Contributions à l'apprentissage statistique : estimation de densité, agrégation d'experts et forêts aléatoires

Jaouad Mourtada

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAX014

Thèse de doctorat

Contributions à l'apprentissage statistique: estimation de densité, agrégation d'experts et forêts aléatoires

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École Doctorale de Mathématique Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue par visio-conférence, le 8 juin 2020, par

JAOUAD MOURTADA

Composition du Jury :

Mme Cristina Butucea Professeur, ENSAE	Président
M. Aurélien Garivier Professeur, Ecole Normale Supérieure de Lyon	Rapporteur
M. Peter Grünwald Professeur, Leiden University & CWI	Rapporteur
M. Francis Bach Directeur de recherche, Inria Paris & ENS	Examinateur
M. Gérard Biau Professeur, Université Paris-VI	Examinateur
M. Gábor Lugosi Professeur, Universitat Pompeu Fabra	Examinateur
M. Stéphane Gaïffas Professeur, Université Paris-VII	Directeur de thèse
M. Erwan Scornet Maître de conférences, École polytechnique	Co-directeur de thèse

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Stéphane Gaïffas et Erwan Scornet. Je souhaite les remercier sincèrement pour leur gentillesse, leurs conseils et leur disponibilité, sans lesquels cette thèse n'aurait jamais vu le jour. Travailler avec vous et interagir avec vos styles complémentaires fut à la fois très formateur scientifiquement et très gratifiant. Je garde un souvenir agréable de nos discussions stimulantes à Polytechnique puis à Paris Diderot, suivies du rituel déjeuner au vietnamien d'à-côté. Cette thèse sous votre encadrement a été une expérience enrichissante, et j'espère que nous garderons le contact dans le futur.

Je remercie chaleureusement Aurélien Garivier et Peter Grünwald d'avoir accepté de rapporter cette thèse. Merci aussi à Francis Bach, Gérard Biau, Cristina Butucea et Gábor Lugosi de m'avoir fait l'immense honneur de faire partie de mon jury.

Je souhaiterais exprimer toute ma reconnaissance envers Gérard Biau, qui m'a orienté vers les statistiques alors que je cherchais ma voie, et à Odalric-ambrym Maillard, qui m'a énormément appris alors que je débutais en apprentissage. Plus largement, je mesure ma dette envers tous mes professeurs de mathématiques au cours de mon cursus, à commencer par Yvonne Kanso et Pascale Richard, qui m'ont transmis leur intérêt pour la discipline.

Effectuer ma thèse au sein du CMAP à Polytechnique fut un plaisir, grâce au compagnonnage des autres doctorants et post-doctorants. Je salue Florian, Mateusz, Mathilde, Nicolas, Raphaël et Tristan avec qui j'ai eu le plaisir de partager un bureau (et pour m'avoir permis tant de fois d'entrer au labo sans mon badge), mais aussi Belhal, Cédric, Cheikh, Frédéric, Geneviève, Hadrien, Imke, Joon, et tous les autres pour les moments de convivialité passés ensemble, au déjeuner, en pause café, à Paris ou en conférence. Je m'excuse au passage auprès de ceux à qui j'avais promis de singer Steve Jobs lors de ma soutenance, j'espère qu'ils me pardonneront ce manquement déplorable à ma parole.

Certaines procédures administratives auraient été pesantes sans l'aide précieuse du personnel de secrétariat du CMAP, je remercie donc Nasséra, Wilfried et Alexandra. Merci aussi à Corinne pour sa gentillesse, sa patience et son soutien pour l'organisation des dernières missions. Enfin, je remercie les professeurs et chargés de TD avec lesquels il m'a été donné de travailler dans le cadre de mes enseignements, pour avoir su montrer l'exemple par leur rigueur et leur enthousiasme communicatif.

Ces derniers mois en postdoc à Gênes furent très agréables et enrichissants. Merci à Lorenzo Rosasco et Silvia Villa pour leur accueil dans l'équipe, mais aussi à tous les membres de l'équipe avec qui j'ai partagé des moments mémorables à Gênes ou à Nice, je pense bien sûr à Andrea, Angel, Cris, Daniele, Gaurvi, Giacomo, Mathurin, Nicolò, Paolo, Vassilis, sans oublier l'inénarrable Cecio. Garder le contact pendant le confinement fut aussi très important pour moi.

Enfin, je souhaite adresser une pensée toute particulière à mes amis. Merci à Jacko, Rémi, Aymeric, Salomé et Amine pour ces moments passés ensemble de Saint-Malo à Tokyo, à Pablo

et Jim pour notre amitié de toujours, à Bertrand pour toutes ces fois où nous avons refait le monde, et à ceux de Rennes, de l'ENS, d'APE, à Maxime et Pierre, à Loucas de P. de V., que j'ai toujours plaisir à retrouver.

Merci enfin à toute ma famille: à Maman, pour son indéfectible soutien, à Papa, pour qui mon affection et mon estime sont toujours aussi vives, à Amélie et Clément, pour tous ces moments de complicité et les liens forts qui nous unissent, mais aussi à Charles pour sa gentillesse.

À la mémoire de mon père.

Contents

Summary of contributions and outline

This manuscript is divided into an introductory chapter (Chapter 1) and three parts, which can be read independently from each other.

Part I. Mondrian Random forests: theory and methodology

This part is devoted to the study of *Mondrian forests* (Lakshminarayanan et al., 2014), a particular *Random forest* procedure. We refer to Section 1.5 for a general introduction to Random forests methods, and in particular to Section 1.5.3 for a more detailed summary of our contributions in Chapters 2 and 3.

Chapter 2. We study a stylized variant of Random forests, namely the *Mondrian forest* estimator introduced by (Lakshminarayanan et al., 2014). By exploiting local and global properties of Mondrian partitions, we establish rates of convergence for this estimator, which are minimax optimal over some nonparametric classes. This contrasts with other randomized tree ensembles (called *purely random forests*) considered in the literature, which achieve suboptimal rates in dimension $d \geqslant 2$ (Breiman, 2004; Biau et al., 2008; Arlot and Genuer, 2014; Klusowski, 2018). Our analysis also exhibits an advantage of forests over individual trees through bias reduction, extending results from Arlot and Genuer (2014).

This chapter is adapted from the article Mourtada et al. (2018), in collaboration with S. Gaïffas and E. Scornet, which has been accepted for publication in *Annals of Statistics*. A preliminary version of this work (Mourtada et al., 2017) appeared in the proceedings of the *Advances in Neural Information Processing Systems (NeurIPS 2017)* conference.

Chapter 3. This chapter complements the previous one from a methodological standpoint. We propose a supervised learning (in particular conditional density estimation) algorithm based on Mondrian forests. Like the original Mondrian forests algorithm (Lakshminarayanan et al., 2014), this procedure can be updated in an online fashion. However, unlike the former this procedure does not require any computing approximation, and satisfies guarantees in the form of oracle inequalities and adaptive convergence rates. The algorithm exploits the updating properties of Mondrian partitions, as well as an efficient aggregation procedure over tree structures introduced by Willems et al. (1995); Helmbold and Schapire (1997). In practice, this method uses more adaptive partitions, as well as a spatially more adaptive regularization level than the estimator studied in Chapter 2. Experiments are performed on several datasets.

This chapter is taken from the paper Mourtada et al. (2019), in collaboration with S. Gaïffas and E. Scornet, which has been submitted for publication.

Part II. Prediction with expert advice

This part deals with the problem of sequential prediction presented in Section 1.2 of the introduction. Chapters 4 and 5, whose results are respectively summarized in Sections 1.2.7 and 1.2.8 of the introduction, can be read independently from each other.

Chapter 4. We study the behavior of the exponential weights (Hedge) algorithm for prediction with expert advice in the online stochastic setting. We prove that anytime Hedge with decreasing learning rate, one of the simplest algorithm for this problem, is both worst-case optimal and adaptive to the easier stochastic and adversarial with a gap problems. Thus, in spite of its small, non-adaptive learning rate, Hedge possesses similar optimal regret guarantees in the stochastic case as more sophisticated adaptive algorithms. Our analysis also exhibits differences with other Hedge variants, such as the fixed-horizon, constant learning rate one, and the one based on the doubling trick, both of which fail to adapt to the easier stochastic setting. Finally, we determine the intrinsic limitations of anytime Hedge in the stochastic case, and discuss the improvements provided by more adaptive algorithms.

This chapter is taken from the article Mourtada and Gaïffas (2019b), in collaboration with S. Gaïffas, which has been published in *Journal of Machine Learning Research*.

Chapter 5. We investigate the problem of sequential prediction, in a context where the class of experts can grow over time. This setting is relevant in situations where one wishes to safely incorporate new learning algorithms as base predictors, or when new variables become available. In this context, we consider several notions of regret, which differ through the considered benchmark one competes against: (1) any expert, from the time it was introduced; (2) arbitrary sequences of experts and (3) sparse sequences of experts, which only switch between a small (but possibly slowly growing) pool of good experts. In each case, building on existing work for fixed classes of experts Vovk (1998); Herbster and Warmuth (1998); Vovk (1999); Bousquet and Warmuth (2002); Koolen et al. (2012), we design efficient algorithms (with complexity linear in the number of rounds and experts) with optimal regret guarantees.

This chapter is adapted from the paper Mourtada and Maillard (2017), in collaboration with O. Maillard, which appeared in the *Proceedings of the international conference on Algorithmic Learning Theory (ALT 2017)*.

Part III. Density estimation, least squares and logistic regression

This part is devoted to regression and density estimation problems, with an emphasis on linear methods. Section 1.3 of the introduction provides background on least squares regression and random covariance matrices, as well as a more detailed exposition of the results from Chapter 6 in Sections 1.3.2 and 1.3.3. In addition, Section 1.4 of the introduction contains a general introduction to the problem of (conditional) density estimation under logarithmic loss, and in particular of logistic regression; our main contributions in Chapter 7 are presented in Sections 1.4.4, 1.4.5 and 1.4.7. While Chapter 7 uses a result from Chapter 6, for the most part they can be read independently.

Chapter 6. We study the standard problem of random-design linear prediction with square loss from a decision-theoretic perspective. It is known from Tsybakov (2003) that, under boundedness constraints on the response (and thus on the regression parameter), the minimax

excess risk scales as $C\sigma^2 d/n$ up to a constant factor, where d is the model dimension, n the sample size, and σ^2 the noise parameter. Here, we do not restrict the optimal regression parameter, and study the excess risk with respect to the full linear class, as a function of the distribution of covariates. We characterize the exact minimax risk in the well-specified case (with linear regression function), relate it to the distribution of *statistical leverage scores* of features and deduce a precise minimax lower bound of $\sigma^2 d/(n-d+1)$ (valid for any distribution of covariates) which nearly matches the risk for centered Gaussian covariates. We then obtain nonasymptotic upper bounds on the minimax risk for covariates that satisfy a "small ball"-type regularity condition, which scale as $(1+o(1))\sigma^2 d/n$ for $d = o(n)$, both in the well-specified and misspecified cases.

Our main technical contribution is the study of the lower tail of the smallest singular value of empirical covariance matrices around 0, which is the object of the second part. First, we establish a lower bound on this lower tail, valid for any distribution with identity covariance in dimension $d \geqslant 2$. We then provide a matching upper bound under a necessary regularity condition on the distribution. Our proof relies on the PAC-Bayesian technique for controlling empirical processes, and extends a prior analysis of Oliveira (2016) devoted to a different part of the lower tail. Equivalently, our results can be stated in terms of controls of moments of the inverse sample covariance matrices. Finally, we show that the aforementioned "small ball" condition on the design holds in the case of independent coordinates with regular distributions.

This chapter is adapted from the manuscript Mourtada (2019).

Chapter 7. We introduce a procedure for predictive (conditional) density estimation under logarithmic loss, which we call SMP (Sample Minmax Predictor). This predictor minimizes a new general excess risk bound, which critically remains valid under model misspecification. On standard examples, this bound scales as d/n where d is the dimension of the model and n the sample size, regardless of the true distribution. The SMP, which is an improper (out-of-model) procedure, improves over proper (within-model) estimators (such as the maximum likelihood estimator), whose excess risk can degrade arbitrarily in the misspecified case. Our bounds also improve over approaches based on online-to-batch conversion, by removing suboptimal $\log n$ factors, addressing an open problem from Grünwald and Kotłowski (2011) for the considered models. For the Gaussian linear model, the SMP admits an explicit expression, and its expected excess risk in the general misspecified case is at most twice the minimax excess risk in the *well-specified case*, but without any condition on the noise variance or approximation error of the linear model. For logistic regression, a penalized SMP can be computed by training two logistic regressions, and achieves a non-asymptotic excess risk of $O((d + B^2R^2)/n)$, where R is a bound on the norm of the features and B the norm of the optimal linear predictor. This improves the rates of proper estimators, which can achieve no better rate than $\min(BR/\sqrt{n}, de^{BR}/n)$ in the worst case (Hazan et al., 2014). This also provides a computationally less demanding alternative to approaches based on online-to-batch conversion of Bayesian mixture procedures (which require approximate posterior sampling), thereby partly answering a question by Foster et al. (2018).

This chapter corresponds to the manuscript Mourtada and Gaïffas (2019a), in collaboration with S. Gaïffas.

Chapter 8. We provide complements to the results from the previous two chapters. We first consider conditional density estimation in the well-specified Gaussian linear model, and

show that even in this case improper estimators can improve over proper ones when the model dimension is high. We then derive a lower bound on Stieltjes transforms of sample covariance matrices, from which we deduce a lower bound for least-squares regression that refines the minimax lower bound of Chapter 6, through a dependence on the signal-to-noise ratio.

Chapter 1

Introduction

Les travaux de cette thèse portent sur le thème général de l'*apprentissage statistique*. Ce domaine formalise dans un cadre général le problème de la *prédiction* à partir de données, et constitue le support théorique et mathématique du champ d'étude dit de l'*apprentissage automatique* (*machine learning* en anglais).

L'apprentissage automatique est un sujet de recherche actif et interdisciplinaire, à la rencontre des statistiques, de l'informatique, des mathématiques, de l'ingénierie, ainsi que des disciplines connexes faisant usage de ces méthodologies. De manière générale, le terme d'apprentissage fait référence au développement de méthodes algorithmiques d'extraction d'information à partir de données, en vue de prédire certaines quantités d'intérêt ou d'effectuer certaines tâches. Parmi les applications contemporaines de l'apprentissage automatique figurent la reconnaissance d'objets dans des images, la traduction automatique, la reconnaissance de parole ainsi que la publicité ciblée. Ce sujet est intimement lié au problème statistique de l'estimation, dont la théorie élémentaire a été largement élaborée au cours du XXe siècle. Il a néanmoins émergé comme un sujet d'étude à part entière, sous l'impulsion notamment des travaux fondateurs de Vapnik et Chervonenkis (Vapnik and Chervonenkis, 1974; Vapnik, 1998). L'apprentissage automatique a fait l'objet d'une activité intense au cours des deux dernières décennies, en raison notamment d'une série de succès significatifs dans ses applications. Ces progrès sont principalement dûs à deux facteurs : d'une part la disponibilité de jeux de données massifs, rendus possibles par les méthodes de collecte automatisée de données (moteurs de recherche, réseaux sociaux) ; d'autre part, les capacités de calcul grandissantes des processeurs modernes, permettant d'utiliser des modèles et algorithmes plus complexes.

Dans cette thèse, nous étudierons plusieurs variantes de ce problème, ainsi que différents algorithmes pour le traiter. Un thème récurrent sera l'obtention de garanties théoriques pour les procédures considérées, que l'on rapportera aux garanties optimales atteignables sous certaines hypothèses. L'intérêt de telles garanties est d'une part de confirmer le bien-fondé des méthodes en question, et d'autre part de comparer différentes procédures entre elles (en termes de leurs comportements respectifs sous diverses hypothèses), ainsi que de suggérer des choix raisonnables de paramètres pour celles-ci.

Cette introduction se décline de la façon suivante. Dans la Section 1.1, nous présentons le problème de l'apprentissage statistique, et décrivons brièvement des résultats généraux et des approches classiques de ce problème. Les sections suivantes traitent de sujets plus spécifiques, étudiés dans les différentes parties de cette thèse ; dans chacun des cas, nous commençons par une introduction au sujet traité et une présentation des résultats existants, avant de décrire

nos contributions. La Section 1.2 traite de la *prédiction séquentielle* (ou *apprentissage en ligne*), une variante du problème de l'apprentissage étudiée dans la Partie II. La Section 1.3 (correspondant au Chapitre 6 de la Partie III), porte sur le problème de la régression linéaire (aussi appelée *agrégation linéaire* ou problème des *moindres carrés*) ainsi que sur l'étude de matrices de covariance aléatoires. Dans la Section 1.4, nous abordons le problème de l'estimation de densité, étudié dans le Chapitre 7 de la Partie III. Enfin, la Section 1.5 est consacrée aux méthodes de *forêts aléatoires*, sur lesquelles porte la Partie I. La Section 1.6 est une annexe technique, qui réunit quelques résultats utilisés dans cette introduction.

1.1 Le problème de l'apprentissage statistique

La plupart des problèmes étudiés dans cette thèse sont des variantes du problème général de l'apprentissage statistique. L'apprentissage statistique (Vapnik, 1998; Friedman et al., 2001; Bousquet et al., 2004; Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014) fournit un cadre commun aux problèmes de prédiction, où le but est de prédire certaines quantités inconnues à partir d'un jeu de données constitué d'observations de même nature. La formalisation de ce problème repose sur une modélisation aléatoire du phénomène étudié, qui permet de relier les observations disponibles aux réalisations non observées des quantités à prédire.

1.1.1 Formulation générale

Le problème de l'apprentissage statistique peut être formulé de façon générale comme suit. Soit $\mathcal{Z}$ un espace mesurable appelé *espace d'observations*, et $\mathcal{G}$ un espace mesurable dont les éléments sont appelés *prédicteurs*. L'adéquation entre prédicteurs et observations est quantifiée par une *fonction de perte* $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ mesurable, où $\ell(g, z)$ s'interprète comme "l'erreur" du prédicteur $g \in \mathcal{G}$ en l'observation $z \in \mathcal{Z}$. Soit P une loi de probabilité sur $\mathcal{Z}$, et Z une variable aléatoire de loi P. La qualité d'un prédicteur $g \in \mathcal{G}$ est mesurée par son *risque* $R(g)$, défini par[1] :

$$R(g) = R_P(g) = \mathbb{E}_{Z\sim P}[\ell(g, Z)]\,. \tag{1.1}$$

Un prédicteur $g \in \mathcal{G}$ est de bonne qualité lorsque son risque $R(g)$ est faible ; le prédicteur optimal $g^* = \arg\min_{g\in\mathcal{G}} R(g)$ est appelé *prédicteur de Bayes*. De manière cruciale, la loi P des observations est inconnue, de sorte que la fonction de risque $R : \mathcal{G} \to \mathbf{R}$ et son minimiseur g^* le sont également. L'objectif du problème consiste à construire, étant donné un *n-échantillon* $Z_1, \dots, Z_n$ indépendant et identiquement distribué (i.i.d.) de loi P, un prédicteur $\widehat{g}_n \in \mathcal{G}$ dépendant de $Z_1, \dots, Z_n$ de faible risque. Plus précisément, introduisons une classe $\mathcal{F} \subset \mathcal{G}$ de prédicteurs, appelée *classe de comparaison* ; on définit l'*excès de risque* relatif à la classe $\mathcal{F}$ d'un prédicteur $g \in \mathcal{G}$ par

$$\mathcal{E}(g) = \mathcal{E}_P(g; \mathcal{F}) := R(g) - \inf_{f\in\mathcal{F}} R(f)\,. \tag{1.2}$$

L'objectif est alors de produire un prédicteur $\widehat{g}_n$ (dans cette introduction, nous utiliserons de manière interchangeable les termes de *prédicteur*, d'*estimateur* et de *procédure* ou *algorithme d'apprentissage*) dont l'excès de risque $\mathcal{E}(\widehat{g}_n)$ est aussi faible que possible. L'excès de risque $\mathcal{E}(\widehat{g}_n)$ étant une variable aléatoire (en tant que fonction de l'échantillon $Z_1, \dots, Z_n$), il est possible de mesurer cette quantité par son espérance

$$\mathcal{E}_n(\widehat{g}_n; P, \mathcal{F}) = \mathbb{E}\big[\mathcal{E}_P(\widehat{g}_n; \mathcal{F})\big]\,, \tag{1.3}$$

ou par ses quantiles

$$\mathcal{E}_{n,\delta}(\widehat{g}_n; P, \mathcal{F}) = \inf\Big\{t \in \mathbf{R} : \mathbb{P}\big(\mathcal{E}_P(\widehat{g}_n; \mathcal{F}) \leqslant t\big) \geqslant 1 - \delta\Big\}\,, \tag{1.4}$$

où $1 - \delta \in (0, 1)$ est un *niveau de confiance*.

Afin d'illustrer les définitions générales précédentes, considérons les exemples suivants :

[1] Dans cette section, nous supposerons implicitement que les espérances considérées sont bien définies dans $\mathbf{R} \cup \{+\infty\}$; c'est notamment le cas lorsque la fonction de perte ℓ est à valeurs positives.

Exemple 1.1 (Estimation de l'espérance)**.** Soit $\mathcal{Z} = \mathbf{R}^d$ pour $d \geqslant 1$, $\mathcal{F} = \mathcal{G} = \mathbf{R}^d$, et $\ell(g, z) = \|g - z\|^2$ (où $\|\cdot\| = \|\cdot\|_2$ désigne la norme euclidienne sur $\mathbf{R}^d$). Supposons que P admette un moment d'ordre 2, c'est-à-dire que $\mathbb{E}[\|Z\|^2] < +\infty$. On a alors, pour tout $g \in \mathbf{R}^d$, $R(g) = \mathbb{E}[\|Z - g\|^2] = \|g - \mathbb{E}[Z]\|^2 + \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2]$, de sorte que $\arg\min_{f \in \mathcal{F}} R(f) = \mathbb{E}[Z]$ et $\mathcal{E}(g) = \|g - \mathbb{E}[Z]\|^2$. Ainsi, le problème d'apprentissage est ici équivalent à celui de l'estimation de la moyenne.

Exemple 1.2 (Estimation de densité)**.** Soit μ une mesure de référence sur l'espace mesurable $\mathcal{Z}$, et $\mathcal{G}$ l'ensemble des densités de probabilité sur $\mathcal{Z}$ par rapport à μ, c'est-à-dire des fonctions mesurables $g : \mathcal{Z} \to \mathbf{R}^+$ telles que $\int_{\mathcal{Z}} g \, \mathrm{d}\mu = 1$. Soit également $\mathcal{F} \subset \mathcal{G}$ une famille de densités sur $\mathcal{Z}$, appelée *modèle statistique*. La *perte logarithmique* (aussi appelée *perte de log-vraisemblance*) est définie par $\ell(g, z) := -\log g(z)$.

Lorsque la loi P admet une densité $p \in \mathcal{G}$, le risque $R(g)$ est minimisé par $g = p$, et coïncide (à une constante près) avec la *divergence de Kullback-Leibler* (ou *entropie relative*) entre p et g :

$$R(g) - R(p) = \mathbb{E}_{Z \sim P}\left[\log\left(\frac{p(Z)}{g(Z)}\right)\right] = \int_{\mathcal{Z}} p \log\left(\frac{p}{g}\right) \mathrm{d}\mu =: \mathrm{KL}(p, g) \geqslant 0\,. \tag{1.5}$$

Ainsi, le problème de l'apprentissage statistique avec perte logarithmique équivaut à celui de l'estimation de densité avec risque de Kullback-Leibler. Dans le cas particulier où $\mathcal{Z} = \mathbf{R}^d$, $\mathcal{F} = \mathcal{G} = \{\mathcal{N}(\theta, I_d) : \theta \in \mathbf{R}^d\}$ est le modèle (de translation) Gaussien et $\mu = (2\pi)^{-d/2}\mathrm{d}z$, ce problème équivaut à l'estimation de la moyenne (Exemple 1.1).

L'excès de risque $\mathcal{E}_n(\widehat{g}_n; P, \mathcal{F})$ d'une procédure $\widehat{g}_n$ par rapport à la classe $\mathcal{F}$ dépend de la loi P, qui est elle-même inconnue. Il est donc souhaitable d'obtenir des garanties valables sur un ensemble $\mathcal{P}$ de lois P sur $\mathcal{Z}$ aussi riche que possible, afin qu'il contienne la loi P du phénomène étudié. En particulier, il est possible de chercher des garanties *uniformes* sur la loi $P \in \mathcal{P}$. Cela conduit naturellement à considérer l'*excès de risque minimax*

$$\mathcal{E}_n^*(\ell, \mathcal{P}, \mathcal{F}) = \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \mathcal{E}(\widehat{g}_n; P, \mathcal{F}) = \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \left(\mathbb{E}[R(\widehat{g}_n)] - \inf_{f \in \mathcal{F}} R(f)\right), \tag{1.6}$$

où l'infimum porte sur tous les estimateurs $\widehat{g}_n$, comme mesure de la difficulté du problème d'apprentissage défini par $(\ell, \mathcal{P}, \mathcal{F})$ (Wald, 1949).

Concluons cette section par une remarque sur la définition du problème. Un estimateur $\widehat{g}_n$ est dit *propre* s'il prend ses valeurs dans la classe de comparaison $\mathcal{F}$, et *impropre* dans le cas contraire. Lorsque l'on se restreint aux estimateurs propres, c'est-à-dire lorsque $\mathcal{G} = \mathcal{F}$, on parle d'*apprentissage propre*. Nous verrons dans la Section 1.4 que la flexibilité additionnelle offerte par l'apprentissage impropre (non restreint) permet pour certains problèmes d'obtenir des garanties inaccessibles au moyen d'estimateurs propres (voir le Chapitre 7).

1.1.2 Apprentissage supervisé

Dans cette thèse, nous étudierons principalement des problèmes d'apprentissage dits *supervisés*. Il s'agit dans ce cas de chercher à prédire une variable de sortie Y, à partir d'une variable d'entrée X. Par exemple, dans un problème de reconnaissance d'objets, X peut encoder une image, et Y l'étiquette "l'image contient une voiture".

Formellement, en apprentissage supervisé, l'espace d'observation $\mathcal{Z}$ est un espace mesurable produit $\mathcal{X} \times \mathcal{Y}$. Les éléments de $\mathcal{X}$ sont appelées *entrées*, *caractéristiques*, *covariables* ou *variables prédictives*, tandis que les éléments de $\mathcal{Y}$ sont appelées *sorties*, *réponses* ou *étiquettes*. On se donne également un espace de prédictions $\widehat{\mathcal{Y}}$, auquel appartiennent les prédictions de la valeur y à partir de x, ainsi qu'une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$; $\ell(\widehat{y}, y)$ correspond à l'erreur de la prédiction $\widehat{y}$ étant donnée la valeur de y. Enfin, l'espace $\mathcal{G}$ des prédicteurs est dans ce cas l'ensemble des fonctions mesurables $g : \mathcal{X} \to \widehat{\mathcal{Y}}$, tandis que $\mathcal{F}$ est une sous-classe de prédicteurs. La fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ permet de définir naturellement une autre fonction de perte $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ (également notée ℓ par léger abus de notation) par $\ell(g, z) := \ell(g(x), y)$ pour $z = (x, y) \in \mathcal{Z}$ et $g \in \mathcal{G}$. Étant donnée une loi jointe P sur le couple $Z = (X, Y)$, le risque d'un prédicteur g s'écrit alors:

$$R(g) = R_P(g) = \mathbb{E}_{(X,Y)\sim P}\big[\ell(g(X), Y)\big]\,. \tag{1.7}$$

Dans ce cas, le prédicteur de Bayes $g^* : \mathcal{X} \to \mathcal{Y}$ s'écrit

$$g^*(x) = \operatorname*{arg\,min}_{\widehat{y}\in\widehat{\mathcal{Y}}} \mathbb{E}[\ell(\widehat{y}, Y)|X = x]\,, \tag{1.8}$$

ce qui justifie son appellation (au vu de l'espérance conditionnelle sur Y sachant X). Le risque $R(g^*)$ correspond à l'erreur "incompressible", due au caractère aléatoire de la réponse Y (y compris sachant X).

Considérons maintenant les exemples suivants, qui constituent des exemples classiques du problème de l'apprentissage supervisé. Ces problèmes sont définis par les espaces de sorties $\mathcal{Y}$ et de prédictions $\widehat{\mathcal{Y}}$, ainsi que la fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$. En pratique, le choix de la fonction de perte dépend de la nature des données et de l'objectif recherché.

Exemple 1.3 (Classification)**.** Lorsque l'on cherche à prédire une réponse (dite *discrète* ou *catégorielle*) appartenant à un ensemble fini $\mathcal{Y}$, il est naturel de considérer l'espace $\widehat{\mathcal{Y}} = \mathcal{Y}$ et la fonction de perte $\ell(\widehat{y}, y) = \mathbf{1}(\widehat{y} \neq y)$ pour $\widehat{y} \in \widehat{\mathcal{Y}}$ et $y \in \mathcal{Y}$, appelée *erreur de classification* (ou *perte* 0-1). Un prédicteur $g : \mathcal{X} \to \mathcal{Y}$ est alors appelé *classifieur*, et son risque $R(g) = \mathbb{P}(g(X) \neq Y)$ est simplement sa probabilité d'erreur. Le classifieur de Bayes (1.8) est alors donné par $g^*(x) = \arg\max_{\widehat{y}\in\mathcal{Y}} \mathbb{P}(Y = \widehat{y}|X = x)$, et le problème de la classification revient à déterminer la sortie la plus probable étant donnée l'entrée.

Exemple 1.4 (Régression)**.** Pour des données *quantitatives*, c'est-à-dire lorsque $\mathcal{Y} = \mathbf{R}$, la fonction de perte la plus courante est la *perte quadratique* $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ pour $\widehat{y}, y \in \mathbf{R}$. Le problème est alors parfois appelé *régression* ou problème des *moindres carrés*. Dans ce cas, un prédicteur $g : \mathcal{X} \to \mathbf{R}$ est une *fonction de régression*.

Le prédicteur de Bayes (1.8) est alors donné par l'*espérance conditionnelle* $g^*(x) := \mathbb{E}[Y|X = x]$ (dès lors que $\mathbb{E}[Y^2] < +\infty$), et est parfois appelé la *fonction de régression* de Y sachant X. De plus, pour tout $g : \mathcal{X} \to \mathbf{R}$, $R(g) - R(g^*) = \mathbb{E}[(g(X) - g^*(X))^2] = \|g - g^*\|^2_{L^2(P_X)}$, où P_X désigne la loi de X. Ainsi, le problème de la régression équivaut à celui de l'estimation de l'espérance conditionnelle de Y sachant X, sous la norme de $L^2(P_X)$ (qui est elle-même inconnue en pratique).

Exemple 1.5 (Estimation de densité conditionnelle)**.** Les deux problèmes précédents sont des exemples de prédiction *ponctuelle* où $\widehat{\mathcal{Y}} = \mathcal{Y}$, et où la prédiction $\widehat{y}$ est une valeur possible de la sortie y, qui doit s'en approcher selon une certaine métrique. Dans certaines situations, il

est souhaitable d'avoir de l'information sur l'incertitude attachée à la réalisation de la sortie. Dans ce cas, il est naturel de former une prédiction *probabiliste* de la réponse y, qui assigne des probabilités aux différentes valeurs possibles de cette variable.

Une fonction de perte possible est alors la perte *logarithmique*, décrite dans l'Exemple 1.2 dans le cas non conditionnel. Étant donné un espace mesurable $\mathcal{Y}$ muni d'une mesure de référence μ, $\widehat{\mathcal{Y}}$ est l'ensemble des densités de probabilité sur $\mathcal{Y}$ par rapport à μ, et la perte s'écrit $\ell(\widehat{y}, y) = -\log \widehat{y}(y)$. Un prédicteur $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ correspond alors à une *densité conditionnelle* de y sachant x, notée $g(y|x) := g(x)(y)$.

Le prédicteur de Bayes (1.8) est donné par la *densité conditionnelle* $g^*(\cdot|x) := \mathrm{d}P_{Y|X=x}/\mathrm{d}\mu$ de Y sachant X (lorsqu'elle existe), tandis que $R(g) - R(g^*)$ vaut $\mathbb{E}[\mathrm{KL}(g^*(\cdot|X), g(\cdot|X))]$ (qui coïncide avec la divergence de Kullback-Leibler entre les lois jointes sur $\mathcal{X} \times \mathcal{Y}$ induites par P_X et g^*, g respectivement). Ainsi, le problème de l'apprentissage supervisé équivaut ici à celui de l'estimation de densité conditionnelle.

Pour plus de détails sur le problème de la classification, outre les références générales indiquées précédemment, nous renvoyons à l'ouvrage de référence Devroye et al. (1996) ; pour une présentation plus succincte du sujet, nous renvoyons à Bousquet et al. (2004). Pour ce qui est de la régression (principalement non paramétrique), nous renvoyons à Györfi et al. (2002); Wasserman (2006); Tsybakov (2009) pour la régression.

1.1.3 Approches générative et discriminative

Comme nous l'avons vu dans la section précédente, le problème de l'apprentissage supervisé se formule de la façon suivante : étant donné un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. de loi P sur $\mathcal{X} \times \mathcal{Y}$, produire un prédicteur $\widehat{g}_n : \mathcal{X} \to \widehat{\mathcal{Y}}$ dont le risque $R(\widehat{g}_n)$ est faible. De plus, le risque minimal atteignable par un prédicteur $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ est le *risque de Bayes* $R(g^*)$, où g^* est le prédicteur de Bayes (1.8). Il est donc naturel de considérer comme objectif de contrôler l'excès de risque par rapport à g^*, c'est-à-dire la différence $R(\widehat{g}_n) - R(g^*)$ entre le prédicteur utilisé et le meilleur possible.

Résultats d'impossibilité. Cependant, comme le montre le résultat suivant (établi par Devroye, 1982), il n'est pas possible de s'approcher de la performance du prédicteur de Bayes avec un échantillon fini sans aucune hypothèse sur la loi P.

Théorème 1.1 (Devroye et al., 1996, Théorème 7.1)**.** *Considérons le problème de la classification binaire (Exemple 1.3), avec $\mathcal{X} = [0,1]$ et $\mathcal{Y} = \{0,1\}$. Pour tous $n \geqslant 1$, $\varepsilon > 0$ et tout classifieur $\widehat{g}_n$, il existe une loi jointe P sur $\mathcal{X} \times \mathcal{Y}$ telle que $R(g^*) = 0$ et $\mathbb{E}[R(\widehat{g}_n)] \geqslant 1/2 - \varepsilon$.*

Pour se convaincre du Théorème 1.1, on peut considérer l'exemple suivant. Supposons X uniforme sur $[0,1]$, et considérons $Y = g^*(X)$, où la fonction $g^* : [0,1] \to \{0,1\}$ prend une valeur constante $a_k \in \{0,1\}$ sur chacun des intervalles $[(k-1)/N, k/N]$, $k = 1, \ldots, N$ (avec $N \geqslant 1$). De plus, les $(a_k)_{1 \leqslant k \leqslant N}$ sont arbitraires (par exemples tirés selon une loi a priori uniforme sur $\{0,1\}^N$). Dans ce cas, un échantillon de taille n ne renseigne que sur (au plus) n des N valeurs de la suite (a_k), donc sur une fraction d'au plus n/N des valeurs de X ; pour les valeurs non observées, il n'est pas possible de faire mieux qu'une prédiction aléatoire. Par conséquent, si $N \gg n$, un échantillon de taille n n'apporte que très peu d'information sur la loi P de (X, Y) : on peut montrer qu'un classifieur $\widehat{g}_n$ ne peut atteindre un risque inférieur à

$(1 - n/N)/2$ dans le pire des cas (en considérant la moyenne de ce risque lorsque la suite (a_k) est elle-même tirée uniformément sur $\{0,1\}^N$).

Le Théorème 1.1 est un résultat de type *no free lunch*, qui affirme qu'il n'est pas possible d'obtenir de garantie non triviale sans hypothèse sur P. Notons cependant que dans le résultat précédent, la loi P mettant en défaut le classifieur $\widehat{g}_n$ dépend de la taille n de l'échantillon. Il existe une contrepartie positive au résultat négatif du Théorème 1.1, obtenue en fixant la loi P et en se plaçant dans un cadre asymptotique où la taille de l'échantillon n tend vers l'infini. Dans ce cas, la notion de garantie de risque à taille d'échantillon fixe est remplacée par la notion asymptotique de *consistance*.

Définition 1.1 (Consistance)**.** Une suite $(\widehat{g}_n)_{n\geqslant 1}$ de prédicteurs (où $\widehat{g}_n$ est fonction d'un échantillon de taille n) est dite *universellement consistante* (relativement à un ensemble $\mathcal{P}$ de lois sur $\mathcal{X} \times \mathcal{Y}$) si, pour toute loi $P \in \mathcal{P}$, on a $\mathbb{E}[R(\widehat{g}_n)] \to R(g^*)$ lorsque $n \to \infty$.

Signalons qu'il existe d'autres variantes de la consistance, obtenues en remplaçant la convergence en espérance dans la Définition 1.1 par la convergence en probabilité (qui est plus faible en général, mais équivalente pour des pertes bornées comme en classification), ou presque sûre (on parle alors de consistance *forte*, qui est plus forte pour des pertes bornées).

Le théorème suivant affirme qu'il existe une suite consistante de classifieurs sur $\mathbf{R}^d$.

Théorème 1.2 (Stone, 1977)**.** *Considérons le problème de la classification binaire avec $\mathcal{X} = \mathbf{R}^d$ et $\mathcal{Y} = \{0,1\}$. Pour $n \geqslant 1$ et $1 \leqslant k \leqslant n$, le classifieur des k-plus proches voisins $\widehat{g}_{n,k}$ est défini comme suit : pour $x \in \mathbf{R}^d$,*

$$\widehat{g}_{n,k}(x) = \mathbf{1}\left(\sum_{j=1}^{k} Y_{(i)} > k/2\right)$$

où $\big((X_{(i)}, Y_{(i)})\big)_{1\leqslant i\leqslant n}$ correspond à l'échantillon $\big((X_i, Y_i)\big)_{1\leqslant i\leqslant n}$ ordonné de sorte que $\|x - X_{(1)}\| \leqslant \ldots \leqslant \|x - X_{(n)}\|$. Alors, si $k_n \to \infty$ et $k_n/n \to 0$, la suite de classifieurs $\widehat{g}_{n,k_n}$ est universellement consistante sur l'ensemble $\mathcal{P}$ des mesures de probabilités sur $\mathbf{R}^d \times \{0,1\}$.

Notons qu'au vu du Théorème 1.1, il n'est pas possible d'obtenir de convergence *uniforme* sur la classe $\mathcal{P}$. Le Théorème 1.2 affirme néanmoins qu'il existe une suite $\widehat{g}_n$ de classifieurs consistante pour toute loi fixe P. Cependant, le résultat négatif suivant (Cover, 1968; Devroye, 1982) montre que la convergence peut être arbitrairement lente, même lorsque la loi P est fixée et lorsque $n \to \infty$.

Proposition 1.1 (Devroye et al., 1996, Théorème 7.2)**.** *Considérons le problème de la classification binaire avec $\mathcal{X} = \mathbf{R}$. Soit $(\varepsilon_n)_{n\geqslant 1}$ une suite décroissante de réels telle que $\varepsilon_1 \leqslant 1/16$ et $\varepsilon_n \to 0$ lorsque $n \to \infty$. Pour toute suite $\widehat{g}_n$ de classifieurs, il existe une loi P telle que $R(g^*) = 0$ et $\mathbb{E}[R(\widehat{g}_n)] \geqslant \varepsilon_n$ pour tout n.*

Approche générative. Il ressort de la discussion précédente qu'il n'est pas possible d'obtenir de garanties sur l'excès de risque $R(\widehat{g}_n) - R(g^*)$ d'un prédicteur $\widehat{g}_n$ sans hypothèse. Il est donc nécessaire d'introduire un *biais inductif* dans la procédure, c'est-à-dire de favoriser certaines distributions, ou certaines formes de dépendance entre la variable d'entrée X et la sortie Y.

Une première façon de le faire consiste à restreindre l'ensemble des lois P considérées, c'est-à-dire à faire une hypothèse de *modélisation* sur cette loi. Dans sa variante la plus forte,

ce type d'approche conduit à faire une hypothèse *paramétrique* sur P, c'est-à-dire à supposer que P appartient à un *modèle* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Dans le cas paramétrique, l'espace des paramètres Θ est un sous-ensemble (par exemple ouvert) de $\mathbf{R}^d$, et la mesure P_θ dépend de manière régulière de θ (au sens où $P_\theta = p_\theta \cdot \mu$ pour une mesure de domination μ fixe, et où la densité p_θ dépend de façon lisse de θ). Cette approche dite *générative* (ou de *modélisation*) a traditionnellement été la plus courante en statistiques (Breiman, 2001b).

L'avantage de l'approche générative est qu'en restreignant le problème, elle en rend possible une analyse fine. Le problème de l'apprentissage tombe alors dans le cadre de la théorie de la décision statistique (Wald, 1949; Lehmann and Casella, 1998; Berger, 1985), qui permet de considérer des notions précises d'optimalité. Il est alors en particulier possible de parler de procédures optimales dans le pire des cas, ou en moyenne selon une certaine loi *a priori* sur les valeurs possibles de P_θ. En outre, la théorie de l'estimation est bien comprise dans le cas paramétrique asymptotique, lorsque la taille de l'échantillon n tend vers l'infini (Ibragimov and Has'minskii, 1981; Le Cam, 1986; van der Vaart, 1998). L'approche générative est également pertinente pour traiter des problèmes de statistique inférentielle, allant au-delà de la stricte prédiction : on s'intéresse alors au paramètre θ en lui-même (ou à une propriété spécifique de la loi P_θ), plutôt que comme intermédiaire permettant d'effectuer des prédictions.

La principale limitation de l'approche générative est qu'elle repose sur l'hypothèse très forte que la loi P appartient à une famille spécifiée. Cette hypothèse n'a en pratique aucune raison d'être satisfaite ; en effet, le modèle ne constitue qu'une approximation de la loi P, cette dernière échappant au contrôle du statisticien. Ainsi, les résultats obtenus sous l'hypothèse de l'appartenance à un modèle ne donnent aucune garantie lorsque cette hypothèse n'est pas satisfaite.

Approche discriminative. Pour des problèmes de nature prédictive, il est en fait possible d'introduire un biais inductif tout en évitant des hypothèses fortement restrictives comme l'appartenance à un modèle paramétrique connu.

Pour s'en convaincre, commençons par considérer un modèle possible $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Définissons, pour tout $\theta \in \Theta$, f_θ le prédicteur de Bayes associé à la loi P_θ :

$$f_\theta(x) := \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathbb{E}_{(X,Y)\sim P_\theta}[\ell(\widehat{y}, Y)|X = x]\,,$$

ainsi que la classe $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ formée par ces prédicteurs. Si $P \in \mathcal{P}$, alors le prédicteur de Bayes appartient à la classe $\mathcal{F}$, et donc pour tout prédicteur g, l'excès de risque par rapport à g^* s'écrit

$$R(g) - R(g^*) = R(g) - \inf_{f \in \mathcal{F}} R(f)\,, \tag{1.9}$$

qui est tout simplement l'excès de risque par rapport à la classe $\mathcal{F}$. Notons à présent que ce dernier peut être défini même lorsque $P \notin \mathcal{F}$, sans référence au prédicteur de Bayes g^*. Il est alors possible de chercher à contrôler l'excès de risque par rapport à $\mathcal{F}$ plutôt que g^*, sans imposer d'hypothèse forte sur P ; intuitivement, cela est envisageable car la complexité de $\mathcal{F}$ est contrôlée, tandis que g^* peut être arbitrairement complexe en fonction de P.

Cette observation motive l'approche dite *discriminative*, où le biais inductif est introduit en restreignant la classe de comparaison $\mathcal{F}$ plutôt que l'ensemble $\mathcal{P}$ de lois considérées. L'objectif est alors de contrôler l'excès de risque par rapport à $\mathcal{F}$. Dans le cas général où $P \notin \{P_\theta : \theta \in$

$\Theta\}$, la relation (1.9) devient la *décomposition en erreurs d'approximation et d'estimation* :

$$R(g) - R(g^*) = \Big(R(g) - \inf_{f\in\mathcal{F}} R(f)\Big) + \Big(\inf_{f\in\mathcal{F}} R(f) - R(g^*)\Big)\,. \tag{1.10}$$

Le premier terme de la décomposition (1.10) est simplement l'excès de risque, aussi appelé *erreur d'estimation* ou *variance* ; lorsque g est un prédicteur $\widehat{g}_n$ construit à partir du jeu de données, ce terme est aléatoire (car dépendant des données). Le second terme de cette décomposition, appelé *erreur d'approximation* ou *biais*, est à l'inverse déterministe (indépendant des données) ; il mesure à quel point la classe $\mathcal{F}$ approche le prédicteur de Bayes g^* en termes de risque. L'approche discriminative sépare ainsi l'étude de l'erreur d'estimation de celle de l'erreur d'approximation.

Exemple 1.6 (Régression linéaire)**.** Considérons le problème de l'apprentissage supervisé avec $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \mathbf{R}$ et perte quadratique (Exemple 1.4). L'approche générative postule un modèle sur (X, Y) ; par exemple, que le vecteur aléatoire (X, Y) est Gaussien :

$$\mathcal{P} = \{\mathcal{N}(\mu_{XY}, \Sigma_{XY}) : \mu_{XY} \in \mathbf{R}^{d+1}, \Sigma_{XY} \in \mathbf{R}^{(d+1)\times(d+1)} \text{ positive}\}\,.$$

Dans ce cas, on a $Y = \langle \beta^*, X\rangle + \varepsilon$ pour un certain $\beta^* \in \mathbf{R}^d$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ $(\sigma^2 > 0)$ est indépendant de X. La fonction de régression s'écrit alors $g^*(x) = \langle \beta^*, x\rangle$, la classe des prédicteurs optimaux associée au modèle $\mathcal{P}$ est donc constituée des fonctions linéaires $\mathcal{F} = \{x \mapsto \langle \beta^*, x\rangle : \beta^* \in \mathbf{R}^d\}$.

À l'inverse, l'approche discriminative conduit à chercher à prédire Y à partir de X avec la même précision que la meilleure fonction linéaire de X, sans nécessairement supposer de forme particulière à la fonction de régression $x \mapsto \mathbb{E}[Y|X = x]$ ou à la loi de $\varepsilon = Y - \langle \beta^*, X\rangle$.

Pour conclure, il n'est en général pas possible d'obtenir de garanties d'excès de risque uniformes du type

$$\sup_{P\in\mathcal{P}} \Big(\mathbb{E}[R(\widehat{g}_n)] - \inf_{f\in\mathcal{F}} R(f)\Big) \tag{1.11}$$

pour un prédicteur $\widehat{g}_n$ calculé à partir d'un échantillon fini, lorsque ni l'ensemble $\mathcal{P}$ de lois considérées ni la classe $\mathcal{F}$ de comparaison ne sont restreintes (par exemple, dans le cas de la classification, lorsque $\mathcal{P}$ est l'ensemble des lois jointes sur $\mathbf{R}^d \times \{0, 1\}$, et $\mathcal{F}$ l'ensemble des fonctions mesurables $\mathbf{R}^d \to \{0, 1\}$). En d'autres termes, afin d'obtenir un excès de risque minimax (1.6) faible, il est nécessaire de restreindre soit l'ensemble $\mathcal{P}$ des lois considérées, soit la classe de comparaison $\mathcal{F}$.

L'approche générative consiste à restreindre la famille $\mathcal{P}$, tandis que l'approche discriminative consiste à restreindre la classe $\mathcal{F}$. La première permet parfois d'obtenir des résultats plus précis que la seconde, mais au prix d'hypothèses plus restrictives. Pour cette raison, l'approche discriminative est généralement privilégiée en apprentissage statistique. Il est cependant courant de combiner les deux approches, en imposant certaines restrictions à la loi P pour analyser plus finement le comportement du risque. En outre, l'approche purement générative est utile pour obtenir des *bornes inférieures* sur la difficulté du problème.

1.1.4 Apprentissage et processus empiriques

Dans cette section, nous passons brièvement en revue un paradigme classique permettant de traiter le problème de l'apprentissage (c'est-à-dire de contrôler l'excès de risque), celui de la

convergence uniforme des processus empiriques. Pour davantage de détails, nous renvoyons à Vapnik and Chervonenkis (1974); van der Vaart and Wellner (1996); van de Geer (1999); Bartlett and Mendelson (2002); Bartlett et al. (2005); Koltchinskii (2006); Boucheron et al. (2005); Koltchinskii (2011); Talagrand (2014); Massart (2007); Boucheron et al. (2013) entre autres références. Nous reprenons le formalisme général de l'apprentissage statistique introduit en Section 1.1.1, en nous restreignant au cas de l'apprentissage propre, pour lequel $\mathcal{F} = \mathcal{G}$.

Minimisation du risque empirique. L'approche sans doute la plus naturelle du problème de l'apprentissage statistique est la *minimisation du risque empirique* (Vapnik, 1998). Définissons, pour tout $f \in \mathcal{F}$, le *risque empirique*

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) \,. \tag{1.12}$$

Notons que $\widehat{R}_n(f)$ est aléatoire (en tant que fonction de $Z_1, \dots, Z_n$), et donc que le risque empirique peut être vu comme un *processus stochastique* $\big(\widehat{R}_n(f)\big)_{f \in \mathcal{F}}$ indexé par les prédicteurs $f \in \mathcal{F}$. Le *minimiseur du risque empirique* (en anglais *Empirical Risk Minimizer*, abrégé ERM) est par définition

$$\widehat{f}_n^{\mathrm{ERM}} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \widehat{R}_n(f) \,, \tag{1.13}$$

où l'on suppose l'existence d'un tel minimiseur (et où l'on fixe un choix de celui-ci lorsqu'il n'est pas unique). D'après la loi des grands nombres, pour tout $f \in \mathcal{F}$ le risque empirique $\widehat{R}_n(f)$ converge presque sûrement vers le risque $R(f)$ lorsque $n \to \infty$, dès lors que $\mathbb{E}[|\ell(f, Z)|] < +\infty$. Il est donc naturel d'espérer de $\widehat{f}_n^{\mathrm{ERM}}$, qui minimise l'approximation $\widehat{R}_n$ de R, qu'il minimise aussi approximativement R.

Borne en terme de l'erreur de généralisation. Pour mener à bien cet argument, la seule convergence *ponctuelle* (c'est-à-dire pour tout $f \in \mathcal{F}$) de $\widehat{R}_n$ vers R s'avère insuffisante. Pour s'en convaincre, on peut considérer l'exemple mentionné à la suite du Théorème 1.1, en prenant pour $\mathcal{F}$ l'ensemble des fonctions mesurables $[0,1] \to \{0,1\}$. Dans ce cas, par l'argument indiqué dans cet exemple, il n'est pas possible d'obtenir de garantie non triviale pour $\widehat{f}_n^{\mathrm{ERM}}$ (ou tout autre prédicteur) sans restriction sur f^*. Il est en revanche possible de contrôler l'excès de risque lorsque la convergence $\widehat{R}_n(f) \to R(f)$ a lieu *uniformément* sur $f \in \mathcal{F}$, comme le montre le théorème suivant.

Théorème 1.3 (Vapnik and Chervonenkis, 1974)**.** *L'excès de risque du minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$ satisfait :*

$$R(\widehat{f}_n^{\mathrm{ERM}}) - \inf_{f \in \mathcal{F}} R(f) \leqslant 2 \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \,. \tag{1.14}$$

Proof. Pour tout $f \in \mathcal{F}$, on a

$$R(\widehat{f}_n^{\mathrm{ERM}}) - R(f) = \big(R(\widehat{f}_n^{\mathrm{ERM}}) - \widehat{R}_n(\widehat{f}_n^{\mathrm{ERM}})\big) + \big(\widehat{R}_n(\widehat{f}_n^{\mathrm{ERM}}) - \widehat{R}_n(f)\big) + \big(\widehat{R}_n(f) - R(f)\big).$$

Le second terme du membre de droite est négatif par définition de $\widehat{f}_n^{\mathrm{ERM}}$, tandis que le premier et le troisième termes sont bornés par $\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$. La borne (1.14) en découle en considérant le supremum du membre de gauche sur $f \in \mathcal{F}$. □

Le Théorème 1.3 majore l'excès de risque en terme de la quantité $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$, appelée *erreur de généralisation*. Cette variable aléatoire est le supremum du processus centré $\big(\widehat{R}_n(f) - R(f)\big)_{f\in\mathcal{F}}$, appelé *processus empirique*. Il existe une riche théorie permettant de contrôler le supremum de processus empiriques (Dudley, 1984; Talagrand, 2014), qui joue un rôle important en statistiques (van der Vaart and Wellner, 1996; van de Geer, 1999; Koltchinskii, 2011). De manière générale, le supremum du processus empirique $(\widehat{R}_n(f) - R(f))_{f\in\mathcal{F}}$ peut être borné (en espérance, ou avec forte probabilité) en fonction de celui d'un autre processus appelé *processus de Rademacher*, au moyen d'une technique générale de *symétrisation* (Giné and Zinn, 1984; Bartlett and Mendelson, 2002; Massart, 2007; Koltchinskii, 2011; Boucheron et al., 2013).

Théorème 1.4 (Giné and Zinn, 1984)**.** *Soit $\varepsilon_1, \dots, \varepsilon_n$ des variables aléatoire indépendantes entre elles et de $Z_1, \dots, Z_n$, telles que $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$. Alors, on a*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\Big] \leqslant 2\,\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\Big|\sum_{i=1}^{n} \varepsilon_i\, \ell(f, Z_i)\Big|\Big]\,. \tag{1.15}$$

La borne (1.15) est essentiellement optimale, car une borne dans le sens inverse est valable (au recentrage de la classe près), voir par exemple (Koltchinskii, 2011, Théorème 2.1). Le terme de droite de l'inégalité (1.15) est appelé *complexité de Rademacher* de la classe

$$\ell \circ \mathcal{F} := \ell(\mathcal{F}, \cdot) = \{\ell(f, \cdot) : f \in \mathcal{F}\}\,.$$

Cette quantité mesure la "richesse" de la classe $\mathcal{F}$, en un sens dépendant de la loi P (à travers les $Z_i \sim P$ dans (1.15)). Lorsque les $\ell(f, \cdot)$ sont bornées par une constante R et lorsque la classe $\mathcal{F}$ est finie, alors il existe une constante absolue $C > 0$ telle que $\mathrm{Rad}_n(\mathcal{F}, P) \leqslant CR\sqrt{\log|\mathcal{F}|/n}$ (Boucheron et al., 2013). De plus, pour des classes $\mathcal{F}$ de "dimension" d (en un sens précis, par exemple la dimension de Vapnik-Chervonenkis, Vapnik 1998 dans le cas de classes de fonctions à valeurs dans $\{0, 1\}$), la complexité de Rademacher satisfait $\mathrm{Rad}_n(\ell \circ \mathcal{F}, P) \leqslant C\sqrt{d/n}$. Enfin, il est aussi possible de contrôler la complexité de Rademacher de classes définies par des conditions de normes (indépendamment de la dimension), pour des fonctions de pertes Lipschitz, en utilisant une inégalité de contraction pour les complexités de Rademacher (Ledoux and Talagrand, 2013).

De manière générale, la complexité de Rademacher $\mathrm{Rad}_n(\mathcal{F}, P)$ (1.15) peut être contrôlée à partir de techniques générales pour l'étude du supremum de processus (sous-) Gaussiens. Ces bornes dépendent de la structure métrique de la classe $\mathcal{F}$ sous la distance $L^2(P)$. La principale technique pour majorer de telles quantités est celle du *chaînage*, dont l'usage remonte à Kolmogorov, exploitée par Dudley (1967) et raffinée par Fernique (1975) et Talagrand, qui consiste à décomposer chaque fonction comme une "chaîne" d'approximations à différents niveaux, puis à contrôler la différence entre niveaux successifs par une majoration du supremum de processus finis sous-Gaussiens par une borne d'union (Talagrand, 2014; Dudley, 1999; Ledoux and Talagrand, 2013; Massart, 2007; Vershynin, 2018). L'usage de cette technique est crucial pour obtenir des bornes optimales sur les complexités de Rademacher de classes "riches", non paramétriques.

Des Théorèmes 1.3 et 1.4, il ressort que l'excès de risque du minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$ peut être contrôlé en fonction de l'erreur de généralisation (1.14), elle-même majorée par la complexité de Rademacher $\mathrm{Rad}_n(\mathcal{F}, P)$. Dans le cas de classes $\mathcal{F}$ de faible complexité (comme les classes finies ou de "dimension" finie), cela implique une borne d'excès de risque

d'ordre $O(1/\sqrt{n})$. Cette vitesse de convergence dite *lente* est optimale dans le pire des cas (c'est-à-dire sans hypothèse particulière sur P) pour certaines fonctions de perte, comme par exemple l'erreur de classification (Exemple 1.3) par un résultat de Devroye and Lugosi (1995).

Localisation. La vitesse lente en $O(1/\sqrt{n})$ s'avère en revanche sous-optimale pour d'autres fonctions de pertes, comme la perte quadratique utilisée en régression (Exemple 1.4) et la perte logarithmique utilisée en estimation de densité (Exemple 1.2). De plus, même dans le cas de la classification, des vitesses plus rapides sont possibles sous certaines hypothèses sur la loi P. Dans ce cas, l'approche consistant à borner l'excès de risque en fonction de l'erreur de généralisation (Théorème 1.3) est fondamentalement sous-optimale : en effet, même dans le cas d'une classe $\mathcal{F} = \{f_0\}$ à un élément, l'erreur de généralisation est d'ordre $O(1/\sqrt{n})$, car l'écart-type de $\widehat{R}_n(f_0)$ est $\sqrt{\mathrm{Var}(\ell(f_0, Z))/n}$.

Commençons par illustrer la possibilité de vitesses rapides par un exemple simple.

Exemple 1.7. On considère le problème de l'estimation de la moyenne (Exemple 1.1) dans le cas où $d = 1$, et l'on pose $\sigma^2 := \mathrm{Var}(Z)$ et $f^* = \mathbb{E}[Z]$ où $Z \sim P$. Le minimiseur du risque empirique est dans ce cas $\widehat{f}_n^{\mathrm{ERM}} = n^{-1} \sum_{i=1}^n Z_i$, et son excès de risque vaut (Exemple 1.1):

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}_n^{\mathrm{ERM}})\big] = \mathbb{E}\big[(\widehat{f}_n^{\mathrm{ERM}} - f^*)^2\big] = \frac{\sigma^2}{n}\,.$$

Au-delà de cet exemple, il est possible d'obtenir des vitesses rapides sous des conditions plus générales sur la fonction de perte et la distribution P. Dans le cas de l'Exemple 1.7, une propriété importante est que *les éléments $f \in \mathcal{F}$ dont l'excès de risque est faible ont une perte fortement corrélée à celle de f^**. En effet, pour tout $f \in \mathcal{F}$, $\ell(f, Z) - \ell(f^*, Z) = (f - f^*)(f + f^* - 2Z)$, de sorte que $\mathrm{Var}(\ell(f, Z) - \ell(f^*, Z)) = 4\sigma^2(f - f^*)^2 = 4\sigma^2\mathcal{E}(f)$.

Ce type de propriété est appelé *hypothèse de marge*. Une hypothèse de cette nature a été introduite par Mammen and Tsybakov (1999); Tsybakov (2004) dans le cas de la classification. La condition de marge et ses conséquences ont également été étudiées par Massart and Nédélec (2006). Nous considérons ici l'hypothèse suivante, introduite par Bartlett and Mendelson (2006) sous le nom de *condition de Bernstein*, et valable pour une fonction de perte générale.

Définition 1.2. Soient $\beta \in (0, 1]$ et $B \geqslant 1$. On dit que la classe $\mathcal{F}$ satisfait la (β, B)-*condition de Bernstein* sous la fonction de perte $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbf{R}$ et la loi P si, en notant $f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}[\ell(f, Z)]$, on a pour tout $f \in \mathcal{F}$,

$$\mathbb{E}\big[\big(\ell(f, Z) - \ell(f^*, Z)\big)^2\big] \leqslant B \cdot \mathbb{E}\big[\ell(f, Z) - \ell(f^*, Z)\big]^\beta\,. \tag{1.16}$$

Remarque 1.1. L'hypothèse de Bernstein est de nature *générative*, en ce sens qu'il s'agit d'une hypothèse sur la loi P de Z (Section 1.1.3).

Puisque $\mathbb{E}\big[(\ell(f, Z) - \ell(f^*, Z))^2\big] \geqslant \mathrm{Var}\big(\ell(f, Z) - \ell(f^*, Z)\big)$, la condition de Bernstein (1.16) implique que $\mathrm{Var}\big(\ell(f, Z) - \ell(f^*, Z)\big) \leqslant B\mathcal{E}(f)^\beta$. Cela signifie précisément que la perte $\ell(f, Z)$ d'éléments $f \in \mathcal{F}$ de faible excès de risque est corrélée à celle de f^*.

Sous une hypothèse de marge de type (1.16) — ou, plus généralement, étant donnée une borne sur $\mathrm{Var}(\ell(f, Z) - \ell(f^*, Z))$ en termes de $\mathcal{E}(f)$ — il est possible d'obtenir des bornes améliorées d'excès de risque. En effet, $\widehat{f}_n^{\mathrm{ERM}}$ peut s'écrire :

$$\widehat{f}_n^{\mathrm{ERM}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \big(\ell(f, Z_i) - \ell(f^*, Z_i)\big)\,; \tag{1.17}$$

la borne d'excès de risque en terme de l'erreur de généralisation implique alors que, avec forte probabilité, $\mathcal{E}(\widehat{f}_n^{\mathrm{ERM}}) \leqslant \delta_1$, où δ_1 dépend de la complexité de Rademacher de la classe $\big(\ell(f,\cdot) - \ell(f^*,\cdot)\big)_{f\in\mathcal{F}}$, et donc de la quantité $\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} \mathbb{E}\big[(\ell(f,Z)-\ell(f^*,Z))^2\big]$ (Talagrand, 1996). Dans ce cas, on a $\widehat{f}_n^{\mathrm{ERM}} \in \mathcal{F}(\delta_1) := \{f \in \mathcal{F} : \mathcal{E}(f) \leqslant \delta_1\}$. Il est alors possible de répéter l'argument précédent, mais en obtenant une nouvelle borne car l'hypothèse (1.16) assure que $\sigma^2(\mathcal{F}(\delta_1)) \leqslant B\delta_1^{\beta}$. En répétant ce processus et en tenant compte des termes supplémentaires qui apparaissent, il est possible d'obtenir une borne améliorée tenant compte de la complexité locale de la classe (c'est-à-dire de celle de sous-classes de la forme $\mathcal{F}(\delta)$). Cette approche fondée sur l'étude de *complexités de Rademacher locales* est due à Koltchinskii (2001, 2006); Bartlett et al. (2005).

Exemple 1.8 (Classe finie)**.** Dans le cas d'une classe $\mathcal{F}$ finie, sous l'hypothèse de Bernstein (1.16) (ainsi qu'une hypothèse de pertes bornées), l'excès de risque $\mathbb{E}[\mathcal{E}(\widehat{f}_n^{\mathrm{ERM}})]$ est majoré par

$$C\left[\Big(\frac{B\log|\mathcal{F}|}{n}\Big)^{1/(2-\beta)} + \frac{\log|\mathcal{F}|}{n}\right]$$

(Boucheron et al., 2005), qui constitue une vitesse améliorée en $O(1/n^{2-\beta})$, allant de $O(1/\sqrt{n})$ (on parle alors de *vitesse lente*) à $O(1/n)$ (vitesse rapide).

Ainsi, l'approche fondée sur les complexités de Rademacher localisées combine l'idée de la convergence uniforme de processus empiriques et celle de localisation. Cette approche est très générale, et fournit des bornes précises sur l'excès de risque du minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$, qui sont optimales dans de nombreuses situations (Koltchinskii, 2006). En particulier, la notion de complexité de Rademacher (et ses variantes localisées) fournit un contrôle précis de la complexité de la classe, en un sens dépendant de la loi P. Mentionnons également que, dans le cas de la régression avec perte quadratique (Exemple 1.4), il existe une variante "ajustée" de la complexité de Rademacher, qui permet d'obtenir des vitesses rapides sans avoir à recourir explicitement aux complexités de Rademacher localisées (Liang et al., 2015).

Le principal inconvénient de l'approche en termes de processus empiriques est sa complexité technique : elle requiert en effet une machinerie sophistiquée permettant le contrôle de déviations de processus empiriques. Cela conduit notamment à des bornes faisant apparaître des constantes souvent sous-optimales, qui ne permettent pas de comparer ou de calibrer précisément des procédures en pratique.

1.1.5 Apprentissage et optimisation stochastique

Nous évoquons maintenant un autre point de vue sur le problème d'apprentissage, celui de l'*optimisation stochastique* (Robbins and Monro, 1951; Polyak and Juditsky, 1992; Nemirovski and Yudin, 1983; Benveniste et al., 1990; Kushner and Yin, 2003; Nesterov, 2004; Boyd and Vandenberghe, 2004; Bach and Moulines, 2013; Bubeck, 2015). Considérons le problème d'apprentissage général décrit en Section 1.1.1, en supposant de plus que $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, où Θ est une partie convexe de $\mathbf{R}^d$. Pour tout $z \in \mathcal{Z}$, on définit la fonction $\ell_z : \Theta \to \mathbf{R}$ par

$$\ell_z(\theta) := \ell(f_\theta, z)\,,$$

de sorte que $R(\theta) = \mathbb{E}[\ell_Z(\theta)]$. Pour $i = 1,\dots,n$, on pose également $\ell_i := \ell_{Z_i}$, de sorte que $\ell_1,\dots,\ell_n$ sont des fonctions aléatoires i.i.d. de même loi que ℓ_Z. En posant $\theta^* = \arg\min R$, le but est de produire un élément $\widehat{\theta}_n \in \Theta$ (dépendant de $\ell_1,\dots,\ell_n$) dont l'excès de risque

$\mathcal{E}(\widehat{\theta}_n) = R(\widehat{\theta}_n) - R(\theta^*)$ est faible. Ainsi posé, ce problème (dit de l'*optimisation stochastique*) est équivalent à celui de l'apprentissage. Cependant, dans cette formulation, le prédicteur f_θ (qui correspond à une fonction $\mathcal{X} \to \widehat{\mathcal{Y}}$ en apprentissage supervisé) est maintenant vu comme un *paramètre* $\theta \in \Theta$, tandis que l'observation Z est représentée par la *fonction* ℓ_Z. Cette différence de formulation conduit à un point de vue complémentaire sur ce problème.

D'une part, dans le cas de l'optimisation stochastique, les hypothèses sont formulées en termes de propriétés analytiques des fonctions ℓ_z ou du risque R. En particulier, la convexité (et ses variantes renforcées) joue un rôle clé dans cette théorie (Boyd and Vandenberghe, 2004; Nesterov, 2004; Bubeck, 2015). Cela tient notamment au fait qu'il est possible de réduire l'étude de problèmes convexes à celle de problèmes *linéaires*. En effet, si la fonction ℓ_z est convexe et différentiable, alors pour tous $\theta, \theta^* \in \Theta$, l'excès de perte est contrôlé par le gradient $\nabla \ell_z(\theta)$:

$$\ell_z(\theta) - \ell_z(\theta^*) \leqslant \langle \nabla \ell_z(\theta), \theta - \theta^* \rangle \, ;$$

nous renvoyons à la Section 1.2.1 pour une conséquence précise de cette inégalité dans le cas de l'optimisation séquentielle. Du point de vue de l'*optimisation*, la convexité garantit que tout minimum local est un minimum global, et permet d'établir la convergence globale d'algorithmes d'optimisation. Du point de vue *statistique*, la *forte convexité* (voir la Section 1.6.3 pour une définition) implique une condition de marge (voir l'Exemple 1.9 ci-dessous), et permet donc d'obtenir des vitesses rapides pour l'excès de risque.

Exemple 1.9 (Forte convexité et condition de Bernstein)**.** Soit Θ une partie convexe de $\mathbf{R}^d$. Supposons que, pour tout $z \in \mathcal{Z}$, la fonction $\ell_z : \Theta \to \mathbf{R}$ est L-Lipschitz (par rapport à la norme euclidienne $\|\cdot\|$ sur $\mathbf{R}^d$) avec $L > 0$. Supposons également que la fonction de risque $R : \Theta \to \mathbf{R}$ est λ-*fortement convexe*, au sens de la Définition 1.8. Supposons enfin que le risque R est minimisé par un élément θ^* intérieur à Θ. Alors, pour tout $\theta \in \Theta$,

$$\mathbb{E}\big[\big(\ell(\theta, Z) - \ell(\theta^*, Z)\big)^2\big] \leqslant L^2 \|\theta - \theta^*\|^2 \leqslant \frac{2L^2}{\lambda}\big(R(\theta) - R(\theta^*)\big)$$

où l'on a utilisé l'inégalité (1.125) Section 1.6.3. Ceci signifie que la condition de Bernstein (Définition 1.2) est satisfaite avec $\beta = 1$ et $B = 2L^2/\lambda$.

Une autre particularité de l'approche "optimisation stochastique" est qu'elle porte souvent sur des procédures *explicites*, c'est-à-dire des algorithmes d'optimisation calculables à partir de certaines quantités associées aux fonctions ℓ_i (par exemple, leurs gradients), là où le point de vue "processus empiriques" conduit plus naturellement à des estimateurs définis implicitement (tel le minimiseur du risque empirique $\widehat{f}_n^{\mathrm{ERM}}$). L'algorithme typique pour les problèmes d'optimisation stochastique, en particulier lorsque le nombre d'observations n et la dimension d sont élevés, est la *descente de gradient stochastique* (en anglais *stochastic gradient descent*, abrégé SGD). La Proposition 1.2 suivante (Nemirovski and Yudin, 1983; Nemirovski et al., 2009; Bubeck, 2015) décrit une variante projetée et "en ligne" de cet algorithme, et énonce une borne d'excès de risque pour sa variante *moyennée* (Polyak and Juditsky, 1992; Ruppert, 1988).

Proposition 1.2. *Supposons que Θ est une partie convexe fermée de $\mathbf{R}^d$ et que, pour tout $z \in \mathcal{Z}$, la fonction $\ell_z : \Theta \to \mathbf{R}$ est convexe, différentiable et L-Lipschitz. Considérons l'algorithme de* descente de gradient stochastique projetée :

- $\widehat{\theta}_1 := \theta_1 \in \Theta$ *fixe ;*

- *pour tout* $i = 1, \dots, n$, *étant donné* $\widehat{\theta}_i$, *on définit :* $\widehat{\theta}_{i+1} := \mathrm{proj}_\Theta(\widehat{\theta}_i - \eta \nabla \ell_{Z_i}(\widehat{\theta}_i))$, *où* $\eta > 0$ *et* $\mathrm{proj}_\Theta(x) := \arg\min_{\theta \in \Theta} \|x - \theta\|$ *pour* $x \in \mathbf{R}^d$.

Soit $B > 0$ *; posons* $\eta = B/(L\sqrt{n})$, *et notons* $\Theta_B := \{\theta \in \Theta : \|\theta - \theta_1\| \leqslant B\}$. *Alors, la moyenne des itérés* $\bar{\theta}_n := \frac{1}{n+1} \sum_{i=1}^{n+1} \widehat{\theta}_i$ *satisfait la borne d'excès de risque suivante :*

$$\mathbb{E}[R(\bar{\theta}_n)] - \inf_{\theta \in \Theta_B} R(\theta) \leqslant \frac{BL}{\sqrt{n+1}}\,. \tag{1.18}$$

Proof. Ce résultat est une conséquence de la conversion online-to-batch (Proposition 1.3) décrite dans la Section 1.2.2 ci-dessous, et de la borne de regret de la descente de gradient en ligne (Proposition 1.11, Section 1.6.4). □

Exemple 1.10 (Prédiction linéaire)**.** Soit $\mathcal{Y}$ un espace mesurable, et $\widetilde{\ell} : \mathbf{R} \times \mathcal{Y} \to \mathbf{R}$ telle que $\widetilde{\ell}(\cdot, y)$ est convexe et C-Lipschitz pour tout $y \in \mathcal{Y}$. Cette condition inclut notamment, lorsque $\mathcal{Y} = \{-1, 1\}$, la *perte logistique* $\widetilde{\ell}(\widehat{y}, y) = \log(1 + e^{-y\widehat{y}})$ (explorée plus en détail dans les Sections 1.4.6 et 1.4.7) et la *perte Hinge* $\widetilde{\ell}(\widehat{y}, y) = \max(-y\widehat{y}, 0)$, avec $C = 1$; la perte *quadratique* bornée $\widetilde{\ell} : [-A, A]^2 \to \mathbf{R}$, $\widetilde{\ell}(\widehat{y}, y) = (\widehat{y} - y)^2$, satisfait également cette condition avec $C = 2A$. Soient également $\Theta = \mathbf{R}^d$ et $\mathcal{Z} = \mathcal{X} \times \mathbf{R}$, où $\mathcal{X} = \{x \in \mathbf{R}^d : \|x\| \leqslant R\}$ pour un certain $R > 0$. Alors, pour tout $z = (x, y) \in \mathcal{Z}$, la fonction $\ell_z(\theta) := \widetilde{\ell}(\langle \theta, x \rangle, y)$ est convexe et CR-Lipschitz ; la borne (1.18) d'excès de risque est donc de $CBR/\sqrt{n+1}$.

La Proposition 1.2 permet d'illustrer les points forts de l'approche en termes d'optimisation stochastique : d'une part, la garantie porte sur un estimateur explicite ; d'autre part, la preuve de la borne d'excès de risque est plus simple que celle fondée sur la convergence uniforme des processus empiriques. En effet, dans le cas particulier de la prédiction linéaire (Exemple 1.10), il est possible d'établir une borne similaire pour le minimiseur du risque empirique restreint à la boule de $\mathbf{R}^d$ de rayon B, par un argument de convergence uniforme reposant sur une inégalité (non triviale) de *contraction* pour les complexités de Rademacher, voir Koltchinskii (2011). Il s'avère de plus qu'il existe des exemples de problèmes d'optimisation stochastique convexe Lipschitz (satisfaisant les hypothèses de la Proposition 1.2) pour lesquels la convergence uniforme du risque empirique n'a pas lieu[2], et où l'excès de risque de l'ERM ne converge pas vers 0 (Shalev-Shwartz et al., 2010).

En outre, il est possible avec cette approche d'obtenir des vitesses rapides de manière plus directe qu'en ayant recours aux complexités de Rademacher localisées (Koltchinskii, 2006; Bartlett et al., 2005) ; une des façons de le faire est de se ramener à la variante séquentielle du problème, décrite dans la Section 1.2, voir par exemple Hazan et al. (2007); Audibert (2009); Lacoste-Julien et al. (2012) (la dernière référence utilisant une variante pondérée du problème séquentiel décrit en Section 1.2).

En contrepartie, cette approche conduit généralement à contrôler le risque par des quantités moins fines que celles apparaissant dans l'analyse en termes de processus empiriques. En effet, les bornes génériques d'excès de risque sous l'hypothèse de régularité Lipschitz (telle la Proposition 1.2) et/ou de forte convexité conduisent typiquement à des bornes dépendant du diamètre du paramètre de comparaison ou de constantes de courbure (forte convexité) uniformes. Or, le diamètre d'une classe est une mesure de complexité moins fine (en particulier, moins adaptative aux propriétés de la loi de Z) que la complexité de Rademacher. Par

[2]Plus précisément, ces contre-exemples sont en dimension infinie, où $\mathbf{R}^d$ est remplacé par un espace de Hilbert. En dimension finie, cela correspond au fait qu'ERM ne satisfait pas de borne de risque générale indépendante de la dimension d (par exemple, en $BL/\sqrt{n}$) en optimisation stochastique convexe Lipschitz.

ailleurs, la forte convexité de la perte en le paramètre θ n'est pas satisfaite pour les problèmes d'apprentissage classiques (telle la régression linéaire, voir la Section 1.3). Si la forte convexité du risque (qui apparaît dans l'Exemple 1.9) est nettement moins restrictive et satisfaite dans certains problèmes "bien conditionnés" en dimension finie (voir la Section 1.4.6), la constante de forte convexité admet généralement une dépendance implicite en la dimension du problème, et est très faible dans un contexte "non paramétrique" de grande dimension (Bach and Moulines, 2013). Il est cependant possible d'obtenir des garanties plus fines, avec une meilleure dépendance en la loi de Z (notamment au spectre de la Hessienne du risque) et sans hypothèse de forte convexité, dans le cas des problèmes quadratiques (Bach and Moulines, 2013; Dieuleveut and Bach, 2016), ou sous des hypothèses plus générales de régularité (Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019) ; nous revenons sur cette question dans la Section 1.4.6 sur la régression logistique.

1.1.6 Le point de vue de la stabilité

Il existe une autre technique générale pour établir des bornes d'excès de risque, reposant sur la notion de *stabilité*. De manière informelle, un prédicteur $\widehat{g}_n$ est dit "stable" s'il n'est pas sensible à de petites perturbations de l'échantillon $Z_1, \dots, Z_n$, et en particulier s'il n'est pas trop influencé par des observations individuelles Z_i.

La notion de stabilité et son lien avec l'erreur de généralisation ont été introduits par Bousquet and Elisseeff (2002), bien que des arguments de même nature remontent à Devroye and Wagner (1979a,b) et Kearns and Ron (1999). Le lien entre la notion de stabilité et les bornes de risque et d'erreur de généralisation a également été étudié par Rakhlin et al. (2005); Sridharan et al. (2009); Shalev-Shwartz et al. (2010). Nous nous plaçons ici dans le cadre de l'apprentissage statistique général introduit en Section 1.1.1.

Définition 1.3 (Stabilité par substitution)**.** Soit $\widehat{g}_n = \widehat{g}_n(Z_1, \dots, Z_n)$ un prédicteur formé à partir de $Z_1, \dots, Z_n$. Pour tous $z \in \mathcal{Z}$ et $i = 1, \dots, n$, notons $\widehat{g}_n^{[Z_i, z]}$ le prédicteur obtenu sur l'échantillon $(Z_1, \dots, Z_{i-1}, z, Z_{i+1}, \dots, Z_n)$ où Z_i a été remplacé par z.

Pour tout $\varepsilon > 0$, on dit que $\widehat{g}_n$ est *ε-stable par substitution (en moyenne)* si, en notant Z une réalisation indépendante de même loi P que $Z_1, \dots, Z_n$, on a

$$\mathbb{E}\big[\ell(\widehat{g}_n, Z) - \ell(\widehat{g}_n^{[Z_n, Z]}, Z)\big] \leqslant \varepsilon\,. \tag{1.19}$$

On dit également que $\widehat{g}_n$ est *uniformément ε-stable par substitution* si, pour tous $Z_1, \dots, Z_n$ et $z \in \mathcal{Z}$, $\ell(\widehat{g}_n, z) - \ell(\widehat{g}_n^{[Z_n, z]}, z) \leqslant \varepsilon$.

La stabilité par substitution uniforme implique la stabilité par substitution en moyenne. La stabilité permet de contrôler l'espérance de l'erreur de généralisation, c'est-à-dire de la différence $R(\widehat{g}_n) - \widehat{R}_n(\widehat{g}_n)$. Dans ce qui suit, nous supposons que $\widehat{g}_n$ dépend de façon symétrique de $Z_1, \dots, Z_n$, c'est-à-dire que sa valeur est inchangée par permutation de Z_i et Z_j, $i \neq j$.

Il existe en fait plusieurs notions de stabilité, en fonction de la notion de "distance" ainsi que du type de "perturbation" de l'échantillon considérées. Par exemple, si $\mathcal{G}$ est un espace vectoriel normé (de norme $\|\cdot\|$), il est en principe possible de définir la stabilité en fonction de la quantité $\|\widehat{g}_n^{[Z_n, Z]} - \widehat{g}_n\|$, ou de considérer la stabilité de l'estimateur à l'ajout (ou la suppression) d'un échantillon (plutôt que le remplacement). Dans la Définition 1.3, nous avons considéré la notion de stabilité de la perte par substitution, car elle conduit naturellement à des bornes d'excès de risque, comme le montrent les résultats suivants.

Lemme 1.1. *Si $\widehat{g}_n$ est un prédicteur symétrique en les observations, on a*

$$\mathbb{E}\big[R(\widehat{g}_n) - \widehat{R}_n(\widehat{g}_n)\big] = \mathbb{E}\big[\ell(\widehat{g}_n, Z) - \ell(\widehat{g}_n^{[Z_n,Z]}, Z)\big]\,. \tag{1.20}$$

En particulier, si $\widehat{g}_n$ est ε-stable par substitution, cette quantité est d'au plus ε.

Proof. Tout d'abord, Z étant indépendant de $Z_1, \dots, Z_n$ (et donc de $\widehat{g}_n$), on a $\mathbb{E}[\ell(\widehat{g}_n, Z)] = \mathbb{E}[\mathbb{E}[\ell(\widehat{g}_n, Z)|\widehat{g}_n]] = \mathbb{E}[R(\widehat{g}_n)]$.

En outre, la loi jointe de $(Z_1, \dots, Z_n, Z)$ est invariante par permutation des variables. Or, permuter Z_n et Z change $\ell(\widehat{g}_n^{[Z_n,Z]}, Z)$ en $\ell(\widehat{g}_n, Z_n)$, donc $\mathbb{E}[\ell(\widehat{g}_n^{[Z_n,Z]}, Z)] = \mathbb{E}[\ell(\widehat{g}_n, Z_n)]$; de même, échanger Z_i et Z_n change $\ell(\widehat{g}_n, Z_n)$ en $\ell(\widehat{g}_n, Z_i)$ (par symétrie de $\widehat{g}_n$ en les observations), donc $\mathbb{E}[\ell(\widehat{g}_n, Z_n)] = \mathbb{E}[\ell(\widehat{g}_n, Z_i)]$ pour $i = 1, \dots, n$. Ainsi

$$\mathbb{E}[\ell(\widehat{g}_n^{[Z_n,Z]}, Z)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(\widehat{g}_n, Z_i)] = \mathbb{E}[\widehat{R}_n(\widehat{g}_n)]\,. \qquad \square$$

Du Lemme 1.1 découle en particulier la borne de risque suivante sur des estimateurs régularisés en optimisation stochastique. Ce résultat est une variation mineure de résultats de Shalev-Shwartz et al. (2010); Sridharan et al. (2009), avec la différence que ces derniers considèrent l'estimateur régularisé (1.21) ci-dessous restreint à la boule de $\mathbf{R}^d$ de rayon $B > 0$.

Corollaire 1.1. *Supposons que $\Theta = \mathbf{R}^d$ et que pour tout $z \in \mathcal{Z}$, la fonction $\theta \mapsto \ell(\theta, z) := \ell(f_\theta, z)$ est convexe et L-Lipschitz. Définissons, pour tout $\lambda > 0$,*

$$\widehat{\theta}_{\lambda,n} := \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i) + \frac{\lambda}{2}\|\theta\|^2 \right\}. \tag{1.21}$$

Alors, la borne d'excès de risque suivante est valide :

$$\mathbb{E}[R(\widehat{\theta}_{\lambda,n})] - \inf_{\theta \in \mathbf{R}^d} \left\{ R(\theta) + \frac{\lambda}{2}\|\theta\|^2 \right\} \leqslant \frac{4L^2}{\lambda n}\,. \tag{1.22}$$

En particulier, pour tout $B > 0$, le choix de $\lambda = 2\sqrt{2} \cdot L/(B\sqrt{n})$ conduit à :

$$\mathbb{E}[R(\widehat{\theta}_{\lambda,n})] - \inf_{\|\theta\| \leqslant B} R(\theta) \leqslant \frac{2\sqrt{2}BL}{\sqrt{n}}\,.$$

Proof. Soit $\widehat{R}_{\lambda,n}(\theta) := \widehat{R}_n(\theta) + \lambda\|\theta\|^2/2$ le risque empirique pénalisé. On a

$$\widehat{\theta}_{\lambda,n}^{[Z_n,Z]} = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left\{ \frac{1}{n}\left[\sum_{i=1}^{n-1} \ell(\theta, Z_i) + \ell(\theta, Z)\right] + \frac{\lambda}{2}\|\theta\|^2 \right\} = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \Big\{ \widehat{R}_{\lambda,n}(\theta) + \big(\ell(\theta, Z) - \ell(\theta, Z_n)\big) \Big\}.$$

Or, la fonction $\widehat{R}_{\lambda,n}$ est λ-fortement convexe (en tant que somme de la fonction λ-fortement convexe $\theta \mapsto \lambda\|\theta\|^2/2$ et de la fonction convexe $\widehat{R}_n$, par convexité de $\ell(\cdot, z)$) ; de plus, la fonction $\theta \mapsto \ell(\theta, Z) - \ell(\theta, Z_n)$ est $2L$-Lipschitz. Par le Lemme 1.3 de l'annexe technique (Section 1.6.3), on en déduit que

$$\ell(\widehat{\theta}_{\lambda,n}, Z) - \ell(\widehat{\theta}_{\lambda,n}^{[Z_n,Z]}, Z) \leqslant \frac{4L^2}{\lambda}\,,$$

c'est-à-dire que $\widehat{\theta}_{\lambda,n}$ est $(4L^2)/\lambda$-uniformément stable par substitution. Par le Lemme 1.1, on a donc $\mathbb{E}[R(\widehat{\theta}_n) - \widehat{R}_n(\widehat{\theta}_{\lambda,n})] \leqslant 4L^2/\lambda$. La borne (1.22) s'en déduit en notant que, pour tout $\theta \in \mathbf{R}^d$,

$$\mathbb{E}[\widehat{R}_n(\widehat{\theta}_{\lambda,n})] \leqslant \mathbb{E}[\widehat{R}_{\lambda,n}(\widehat{\theta}_{\lambda,n})] \leqslant \mathbb{E}[\widehat{R}_{\lambda,n}(\theta)] = R(\theta) + \frac{\lambda}{2}\|\theta\|^2 . \qquad \square$$

La stabilité est un paradigme complémentaire à celui de la convergence des processus empiriques (Section 1.1.4). Cette notion sous-tend les bornes obtenues pour le problème de l'apprentissage séquentiel décrit dans la Section 1.2, et par conséquent la preuve de la Proposition 1.2, voir également la Section 1.2.5. De plus, l'existence d'estimateurs stables minimisant approximativement le risque empirique caractérise la possibilité d'estimateurs consistants dans un cadre général, même lorsque la convergence uniforme du risque empirique vers le risque n'a pas lieu, comme par exemple pour le problème d'optimisation stochastique considéré dans la Proposition 1.2 et le Corollaire 1.1 (Shalev-Shwartz et al., 2010). Dans le Chapitre 7 (voir également la Section 1.4.4 de cette introduction), nous introduisons un raffinement des bornes d'excès de risque en termes de stabilité de la perte, qui conduit dans le cas de l'estimation de densité (Exemples 1.2 et 1.5), et en particulier de la régression logistique, à une procédure admettant des garanties d'excès de risque améliorées.

1.1.7 Décomposition biais-variance et approximation quadratique locale

Considérons le cas de la régression avec perte quadratique (Exemple 1.4). Étant donné un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ (où X prend ses valeurs dans l'espace mesurable $\mathcal{X}$ et Y dans $\mathbf{R}$), le but est de produire un estimateur $\widehat{g}_n$ de la fonction $g^*(x) = \mathbb{E}[Y|X = x]$ de régression de Y sachant X (on suppose ici que $\mathbb{E}[Y^2] < +\infty$). Dans ce cas, l'excès de risque de $\widehat{g}_n$ par rapport à g^* satisfait la *décomposition biais-variance* suivante :

$$\mathbb{E}[R(\widehat{g}_n)] - R(g^*) = \mathbb{E}\big[(\widehat{g}_n(X) - g^*(X))^2\big] = \mathbb{E}\big[(\widehat{g}_n(X) - \bar{g}_n(X))^2\big] + \mathbb{E}\big[(\bar{g}_n(X) - g^*(X))^2\big]$$

où $\bar{g}_n(x) := \mathbb{E}[\widehat{g}_n(x)]$ pour tout $x \in \mathcal{X}$; le premier terme du membre de droite de l'équation ci-dessus correspond à la *variance* (il est égal à $\mathbb{E}[\mathrm{Var}(\widehat{g}_n(X)|X)]$), tandis que le second est par définition le *biais* (voir également la Section 1.5.2, qui traite des estimateurs "ensemblistes").

La décomposition biais-variance simplifie considérablement l'étude de la régression avec perte quadratique, en permettant des calculs explicites (Györfi et al., 2002). Elle est cependant spécifique à cette fonction de perte, à l'inverse des approches fondées sur la convergence de processus empiriques (Section 1.1.4), la stabilité (Section 1.1.6) ou la conversion "online-to-batch" (décrite dans la Section 1.2.2 ci-dessous). Il est en revanche possible d'étendre les résultats obtenus pour la perte quadratique à des fonctions de pertes "lisses" au moyen d'approximations quadratiques locales du risque. Cela est possible lorsque l'on dispose d'un contrôle de l'erreur de cette approximation quadratique ; la notion d'*auto-concordance*, c'est-à-dire un contrôle de la dérivée troisième en fonction de la dérivée seconde, permet d'étendre l'analyse fine des problèmes quadratiques à d'autres cas (Bach, 2010; Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019). Nous reviendrons sur cette approche dans la Section 1.4.6 dédiée à la régression logistique.

1.2 Prédiction séquentielle de suites arbitraires

Dans la Section 1.1.3, nous avons distingué les points de vue génératif (reposant sur l'hypothèse que la loi P appartient à un modèle $\mathcal{P}$) et discriminatif (qui ne fait pas cette hypothèse, mais considère l'excès de risque par rapport à une classe restreinte $\mathcal{F}$) du problème de l'apprentissage statistique. Dans cette section, nous considérons un problème voisin de l'apprentissage statistique, l'*apprentissage séquentiel* (aussi appelé *apprentissage en ligne*, ou *online learning*), pour lequel il est possible de se passer de l'hypothèse de stochasticité des observations, c'est-à-dire de l'existence d'une loi P telle que $Z_1, \dots, Z_n$ soient des variables i.i.d. de loi P.

L'ouvrage de référence sur l'apprentissage séquentiel est Cesa-Bianchi and Lugosi (2006) ; ce compte-rendu complet couvre notamment la théorie de la prédiction séquentielle avec perte logarithmique (dont Merhav and Feder, 1998 fournit une présentation spécifique), qui est l'une des sources de cette théorie, ainsi que les liens avec la théorie des jeux (Blackwell, 1956; Von Neumann and Morgenstern, 1947) et l'optimisation. Pour une approche complémentaire, nous renvoyons également aux ouvrages plus récents Shalev-Shwartz (2012); Hazan (2016), qui mettent en avant le lien avec l'optimisation convexe. Enfin, le problème de l'apprentissage séquentiel admet une variante à information partielle (incluant le problème dit des "bandits"), dont nous ne traiterons pas ici, présentée dans les ouvrages Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2019). Des présentations plus succinctes de la prédiction séquentielle et des problèmes de bandits figurent également dans les articles introductifs Stoltz (2010); Faure et al. (2015).

La prédiction séquentielle fait l'objet de la Partie II de ce manuscrit. Après une introduction au problème (Section 1.2.1), une discussion du lien avec l'apprentissage statistique (Section 1.2.1) et quelques résultats classiques sur l'agrégation à poids exponentiels (Sections 1.2.3 à 1.2.6), nous présentons nos contributions dans les Sections 1.2.7 (Chapitre 4) et 1.2.8 (Chapitre 5).

1.2.1 Apprentissage séquentiel

Dans cette section, nous introduisons le formalisme de l'apprentissage séquentiel, et détaillons quelques formulations de ce problème.

Apprentissage séquentiel général. Nous adoptons ici les mêmes notations que dans la Section 1.1.1. Dans le cas de l'apprentissage séquentiel, étant donnée une fonction de perte $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$, l'objectif est de prédire une à une des observations $z_1, \dots, z_n \in \mathcal{Z}$. Plus précisément, le problème se formule comme un jeu entre un Agent (qui cherche à prédire les observations) et un Environnement (qui génère celles-ci). À chaque instant $t = 1, \dots, n$:

- l'Agent choisit un prédicteur $\widehat{g}_t \in \mathcal{G}$, en se fondant sur les observations précédentes $z_1, \dots, z_{t-1}$;
- l'Environnement révèle la valeur z_t de l'observation au temps t (qui peut dépendre de $\widehat{g}_1, \dots, \widehat{g}_t$). L'Agent subit alors une perte de $\ell(\widehat{g}_t, z_t)$.

L'objectif de l'Agent est d'obtenir une perte cumulée faible. Il n'est pas difficile de voir que cela n'est pas possible sans hypothèse sur la suite $z_1, \dots, z_n$: en effet, à chaque instant t, l'Environnement peut choisir z_t en fonction de $\widehat{g}_t$, de sorte que $\ell(\widehat{g}_t, z_t)$ soit élevé. En d'autres termes, quelle que soit la stratégie de l'Agent, il existe une suite d'observations pour laquelle

la stratégie mène à une erreur élevée. Comme dans le cas de l'apprentissage statistique, cette difficulté est levée en adoptant une approche discriminative (Section 1.1.3), consistant à restreindre la classe de comparaison plutôt que la suite $z_1, \dots, z_n$. Ainsi, étant donnée une sous-classe $\mathcal{F} \subset \mathcal{G}$, l'objectif est de déterminer une stratégie de l'Agent telle que le *regret*

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t) \tag{1.23}$$

soit contrôlé indépendamment de (ou sous de faibles hypothèses sur) la suite d'observations $z_1, \dots, z_n$. À l'instar de l'excès de risque, il est possible de définir l'excès de risque minimax. Étant donné un ensemble $\mathcal{S} \subset \mathcal{Z}^n$ de suites $(z_1, \dots, z_n)$ d'observations et une classe $\mathcal{F}$ de prédicteurs de référence, le *regret minimax* est par définition:

$$\mathsf{Reg}_n^*(\ell, \mathcal{S}, \mathcal{F}) := \inf_{g_1, \dots, g_n} \sup_{(z_1, \dots, z_n) \in \mathcal{S}} \sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t) \tag{1.24}$$

où $(g_1, \dots, g_n)$ est une *stratégie de prédiction*, c'est-à-dire que g_t est une fonction de $\mathcal{Z}^{t-1}$ vers $\mathcal{G}$, et où $\widehat{g}_t := g_t(z_1, \dots, z_{t-1})$. Le biais inductif est alors introduit à travers le choix de la classe $\mathcal{F}$, plutôt que la suite $\mathcal{S}$ d'observations ; il est alors typique de choisir $\mathcal{S} = \mathcal{Z}^n$.

La notion de regret constitue l'analogue séquentiel de celle d'excès de risque. Comme indiqué au début de cette section, il n'est pas nécessaire de supposer ici que la suite $z_1, \dots, z_n$ est formée de n variables i.i.d. Au contraire, l'objectif est d'obtenir des garanties de regret (1.23) valables pour toute suite $z_1, \dots, z_n \in \mathcal{Z}$; on parle alors de garanties valables pour des *suites arbitraires* (*individual sequences* en anglais). Notons que dans le cas de l'apprentissage statistique, l'hypothèse d'observations aléatoires suivant une loi P est nécessaire à la formulation du problème, puisqu'elle permet de définir le risque (erreur moyenne sur une population non observée). Dans le cas de l'apprentissage séquentiel, l'erreur est mesurée sur la suite d'observations $z_1, \dots, z_n$ elle-même (chaque prédicteur $\widehat{g}_t$ étant choisi avant d'observer z_t), ce qui permet de se passer de l'hypothèse de stochasticité.

Nous passerons en revue de manière détaillée deux exemples du problème de l'apprentissage séquentiel : l'agrégation d'experts (Sections 1.2.3 à 1.2.8) et l'estimation de densité en ligne (Section 1.4.3). Nous concluons cette présentation en évoquant la variante *supervisée* de l'apprentissage séquentiel, ainsi que l'*optimisation convexe en ligne*.

Apprentissage séquentiel supervisé. Soient $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$ trois ensembles, et $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ une fonction de perte. Le problème de l'apprentissage séquentiel *supervisé* est défini par la classe $\mathcal{G}$ des fonctions $\mathcal{X} \to \widehat{\mathcal{Y}}$, l'espace $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, la fonction de perte $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ définie par $\ell(g, (x, y)) = \ell(g(x), y)$, ainsi qu'une classe de fonctions $\mathcal{F} \subset \mathcal{G}$. Dans ce cas, choisir une fonction $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ (avant de voir (x, y)) et recevoir la perte $\ell(g, (x, y))$ étant donné (x, y) équivaut à voir x, puis choisir $\widehat{y} = g(x)$ étant donné x, puis subir la perte $\ell(\widehat{y}, y)$ étant donné y. Ainsi, le problème admet la formulation équivalente suivante: pour tout $t \geqslant 1$,

- l'Environnement révèle la valeur $x_t \in \mathcal{X}$ de la variable prédictive ;
- l'Agent effectue une prédiction $\widehat{y}_t \in \widehat{\mathcal{Y}}$ (dépendant de x_t et des observations passées) ;
- l'Environnement révèle la valeur y_t de la réponse au temps t (pouvant dépendre de $x_t, \widehat{y}_t$ ainsi que des observations et prédictions passées). L'Agent subit alors une perte $\ell(\widehat{y}_t, z_t)$.

Le regret s'écrit alors :

$$\mathsf{Reg}_n = \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\,.$$

Optimisation en ligne. Tout comme l'apprentissage statistique, l'optimisation stochastique admet une variante séquentielle, appelée *optimisation (convexe) en ligne* (Zinkevich, 2003; Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012; Hazan, 2016). Le problème se formule de la façon suivante : étant donné un domaine de contrainte $\Theta \subset \mathbf{R}^d$ et une classe $\mathscr{C}$ de fonctions $\Theta \to \mathbf{R}$ (par exemple, si Θ est convexe, les fonctions convexes et L-Lipschitz sur Θ), à chaque instant $t = 1, \dots, n$,

- L'Agent choisit un élément $\widehat{\theta}_t \in \Theta$ (dépendant de $\ell_1, \dots, \ell_{t-1}$) ;
- L'Environnement révèle une fonction $\ell_t \in \mathscr{C}$ (qui peut dépendre de $\widehat{\theta}_1, \dots, \widehat{\theta}_t$). L'Agent subit alors la perte $\ell_t(\widehat{\theta}_t)$.

Le but est alors de déterminer une stratégie de choix de $\widehat{\theta}_1, \dots, \widehat{\theta}_n$ pour laquelle le *regret*

$$\mathsf{Reg}_n = \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \ell_t(\theta) \tag{1.25}$$

est contrôlé indépendamment de la suite $\ell_1, \dots, \ell_n$ de fonctions. De la même manière que dans le cas statistique, l'optimisation en ligne est équivalente à l'apprentissage séquentiel *propre*, c'est-à-dire avec la restriction $\mathcal{F} = \mathcal{G}$.

Tout comme en optimisation stochastique, la notion de convexité joue un rôle clé en optimisation en ligne. Cela tient au fait que le regret dans le cas convexe peut être majoré par le regret dans le cas linéaire. En effet, si Θ et $\ell_t : \Theta \to \mathbf{R}$ sont convexes et si ℓ_t est différentiable[3], alors pour tout $\theta \in \Theta$,

$$\ell_t(\widehat{\theta}_t) - \ell_t(\theta) \leqslant \langle \nabla \ell_t(\widehat{\theta}_t), \widehat{\theta}_t - \theta \rangle\,, \tag{1.26}$$

de sorte qu'en posant $h_t := \nabla \ell_t(\widehat{\theta}_t)$, le regret sur la suite de fonctions (ℓ_t) est majoré par celui sur la suite de fonctions linéaires $\langle h_t, \cdot \rangle$:

$$\mathsf{Reg}_n = \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \ell_t(\theta) \leqslant \sum_{t=1}^{n} \langle h_t, \widehat{\theta}_t \rangle - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \langle h_t, \theta \rangle\,. \tag{1.27}$$

Ceci montre que, du point de vue de l'optimisation en ligne, les fonctions linéaires sont les fonctions convexes les plus "difficiles", et qu'il est possible de transférer des résultats pour les fonctions linéaires aux fonctions convexes générales.

Agrégation d'experts. L'agrégation d'experts (ou *prédiction séquentielle à l'aide d'experts*, en anglais *prediction with expert advice*) est une formulation alternative du problème de la prédiction séquentielle, qui sera utilisée dans la Partie II de cette thèse. Dans ce cas, on se donne une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$, où $\widehat{\mathcal{Y}}$ est l'espace des prédictions et $\mathcal{Y}$ l'espace du *signal*, ainsi qu'un ensemble abstrait Θ d'*experts*, c'est-à-dire de sources de prédictions. Dans ce qui suit, on pourra supposer Θ fini. À chaque étape $t = 1, \dots, n$:

[3] Il est possible de se passer de cette hypothèse si $\widehat{\theta}_t$ est intérieur à Θ, en considérant un *sous-gradient* $h_t \in \partial \ell_t(\widehat{\theta}_t)$ satisfaisant par définition l'inégalité voulue (Boyd and Vandenberghe, 2004).

- l'Environnement révèle les prédictions $(\widehat{y}_{\theta,t})_{\theta\in\Theta} \in \widehat{\mathcal{Y}}^{\Theta}$ des experts $\theta \in \Theta$;
- l'Agent détermine sa propre prédiction $\widehat{y}_t \in \widehat{\mathcal{Y}}$;
- l'Environnement choisit la valeur $y_t \in \mathcal{Y}$ du signal.

Le but est alors de déterminer une stratégie de l'Agent garantissant un regret

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} \ell(\widehat{y}_{\theta,t}, y_t)$$

contrôlé. Notons que, dans ce cadre, les prédictions $\widehat{y}_{\theta,t}$ des experts sont arbitraires, et peuvent être considérées comme des "boîtes noires". Cela confère une grande flexibilité ; par exemple, les experts $\theta \in \Theta$ peuvent eux-mêmes correspondre à des stratégies de prédiction séquentielles, dont les prédictions dépendent des observations précédentes.

L'apprentissage séquentiel général (décrit dans le premier paragraphe de cette section) peut être vu comme un cas particulier de ce problème, en posant $\widehat{\mathcal{Y}} := \mathcal{G}$, $\mathcal{Y} := \mathcal{Z}$ (avec identification des fonctions de pertes), et enfin en posant $\widehat{y}_{\theta,t} := f_\theta$ pour tous $\theta \in \Theta$ et $t \geqslant 1$. L'intérêt de la formulation précédente est toutefois de faire le lien avec le problème général de l'apprentissage statistique, à travers des notations similaires et la conversion "online to batch" (voir la Section 1.2.2).

Il est possible d'envisager l'apprentissage séquentiel supervisé comme une variante de l'agrégation d'experts de manière plus directe. En reprenant les notations du cas supervisé (les espaces $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$, la fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ et la classe $\mathcal{F}$ de fonctions $\mathcal{X} \to \mathcal{Y}$), notons $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. Ce problème se réduit à l'agrégation d'experts avec la même fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$, en posant $\widehat{y}_{\theta,t} := f_\theta(x_t)$ pour tous $t = 1, \dots, n$ et $\theta \in \Theta$.

Enfin, dans les Sections 1.2.4 et 1.2.5, nous verrons que l'étude de l'apprentissage séquentiel général peut se ramener, sous certaines hypothèses sur la fonction de perte ℓ, à celle d'une variante de l'apprentissage séquentiel général décrit en début de section.

1.2.2 Conversion "online to batch"

Il existe un lien général entre l'apprentissage séquentiel et l'apprentissage statistique : il est en effet possible de convertir toute garantie de regret pour un algorithme de prédiction séquentielle en une borne d'excès de risque pour un certain prédicteur. Ce procédé appelé *conversion "online to batch"* (Littlestone, 1989; Cesa-Bianchi et al., 2004) a été introduit dans le cas de l'estimation de densité par Barron (1987); Catoni (1997); Yang (2000). Dans ce qui suit, nous reprenons les notations de la Section 1.1.1.

Proposition 1.3 (Conversion "online to batch"). *Soit $\widehat{g}_1, \dots, \widehat{g}_{n+1}$ des prédicteurs, tels que $\widehat{g}_t$ dépende de $Z_1, \dots, Z_{t-1}$. Supposons que $\mathcal{G}$ est un espace convexe*[4]*, et que la fonction $g \mapsto \ell(g, z)$ est convexe pour $z \in \mathcal{Z}$. Soit P une mesure de probabilité sur $\mathcal{Z}$, et $Z_1, \dots, Z_{n+1}$ des variables i.i.d. de loi P. Définissons le prédicteur moyenné $\bar{g}_n$, dépendant de $Z_1, \dots, Z_n$, par*

$$\bar{g}_n := \frac{1}{n+1} \sum_{t=1}^{n+1} \widehat{g}_t \,. \tag{1.28}$$

[4] C'est-à-dire une partie convexe (mesurable) d'un espace vectoriel réel (mesurable).

Alors, en notant Reg_{n+1} *le regret* (1.23) *de* $\widehat{g}_1, \dots, \widehat{g}_{n+1}$ *par rapport à la classe* $\mathcal{F}$ *sur la suite* $Z_1, \dots, Z_{n+1}$, $\bar{g}_n$ *satisfait la borne d'excès de risque suivante:*

$$\mathbb{E}\big[R(\bar{g}_n)\big] - \inf_{f \in \mathcal{F}} R(f) \leqslant \frac{1}{n+1} \mathbb{E}\big[\mathsf{Reg}_{n+1}\big] \,. \tag{1.29}$$

Proof. Pour tout $t = 1, \dots, n+1$, $\widehat{g}_t$ (qui est une fonction de $Z_1, \dots, Z_{t-1}$) est indépendant de Z_t, de sorte que $\mathbb{E}[\ell(\widehat{g}_t, Z_t)] = \mathbb{E}[\mathbb{E}[\ell(\widehat{g}_t, Z_t)|\widehat{g}_t]] = \mathbb{E}[R(\widehat{g}_t)]$. Ainsi, pour tout $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}\big[R(\bar{g}_n)\big] - R(f) &\leqslant \frac{1}{n+1} \sum_{t=1}^{n+1} \big(\mathbb{E}[R(\widehat{g}_t)] - R(f)\big) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{t=1}^{n+1} \big(\ell(\widehat{g}_t, Z_t) - \ell(f, Z_t)\big)\right] \leqslant \frac{\mathbb{E}[\mathsf{Reg}_{n+1}]}{n+1}
\end{aligned}$$

où la première inégalité découle de la convexité de ℓ (et donc de R). L'inégalité (1.29) s'en déduit en considérant le supremum sur $f \in \mathcal{F}$. □

Remarque 1.2 (Randomisation). L'hypothèse de la Proposition 1.3 que $\mathcal{G}$ est convexe et que ℓ est convexe en son premier argument n'est pas restrictive : en effet, il est toujours possible de s'y ramener en considérant des prédicteurs randomisés. Cela revient à considérer la classe $\mathcal{G}'$ des mesures de probabilités sur $\mathcal{G}$ (dont $\mathcal{G}$ s'identifie à un sous-ensemble, en associant à $g \in \mathcal{G}$ la mesure de Dirac δ_g), ainsi que la fonction de perte $\ell'(\rho, z) = \mathbb{E}_{g\sim\rho}[\ell(g, z)]$ pour $\rho \in \mathcal{G}'$ et $z \in \mathcal{Z}$, qui est linéaire (donc convexe) en ρ et satisfait $\ell'(\delta_g, z) = \ell(g, z)$. La conversion online-to-batch (1.28) revient alors à tirer $\bar{g}_n$ uniformément parmi $\widehat{g}_1, \dots, \widehat{g}_{n+1}$.

Ainsi, l'excès de risque minimax $\mathcal{E}_n^*(\ell, \mathcal{P}, \mathcal{F})$ (1.6) par rapport à une classe $\mathcal{F}$ avec $\mathcal{P} = \mathcal{P}(\mathcal{Z})$ est majoré par le regret minimax $\mathsf{Reg}_{n+1}^*(\ell, \mathcal{Z}^{n+1}, \mathcal{F})/(n+1)$. Nous discuterons plus en détail les avantages et les limites de la réduction de l'apprentissage statistique à l'apprentissage séquentiel dans le cas de l'estimation de densité en ligne (Section 1.4.3).

1.2.3 Agrégation à poids exponentiels

Dans cette section, nous décrivons une stratégie fondamentale de prédiction séquentielle, à savoir l'*agrégation à poids exponentiels* (Vovk, 1990, 1998; Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006). Cette procédure (et son analyse de regret) généralise les stratégies séquentielles de prédiction par *mélange Bayésien* utilisées dans le cas de la perte logarithmique (Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006).

Nous nous plaçons dans le cas du problème de l'agrégation d'experts, formulé à la fin de la Section 1.2.1. Pour tout $t \geqslant 1$, on note

$$\widehat{L}_t := \sum_{s=1}^{t} \ell(\widehat{y}_s, y_s), \qquad L_{\theta,t} := \sum_{s=1}^{t} \ell(\widehat{y}_{\theta,s}, y_s) \,,$$

les pertes cumulées respectives de l'algorithme de prédiction et de l'expert $\theta \in \Theta$.

Limite de la minimisation du risque empirique. La stratégie la plus naturelle consiste à prédire, à chaque étape $t \geqslant 1$, comme l'expert $\widehat{\theta}_{t-1} \in \Theta$ dont la perte cumulée sur les observations précédentes est la plus faible :

$$\widehat{y}_t := \widehat{y}_{\widehat{\theta}_{t-1},t}\,, \qquad \widehat{\theta}_{t-1} := \operatorname*{arg\,min}_{\theta\in\Theta} L_{\theta,t-1}\,. \tag{1.30}$$

Cette stratégie correspond à (la variante séquentielle de) la minimisation du risque empirique, considérée dans le cas statistique en Section 1.1.4. Cependant, à l'inverse du cas statistique, cette approche n'admet aucune garantie non triviale de regret valide pour des suites arbitraires, comme le montre l'exemple suivant.

Exemple 1.11 ("Inconsistance" d'ERM pour des suites arbitraires)**.** Considérons la perte quadratique $\ell(\widehat{y}, y) = (y - \widehat{y})^2$ sur $\widehat{\mathcal{Y}} = \mathcal{Y} = [0, 1]$, ainsi que la classe à deux experts $\Theta = \{1, 2\}$, avec $\widehat{y}_{1,t} = 0$ et $\widehat{y}_{2,t} = 1$ pour tout $t \geqslant 1$. Considérons la suite $(y_t)_{t\geqslant 1}$ donnée par $y_1 = 1/3$, $y_{2k} = 1$ et $y_{2k+1} = 0$ ($k \geqslant 1$). Alors, pour tout $k \geqslant 1$, la stratégie (1.30) prédit $\widehat{y}_{2k} = \widehat{y}_{1,2k} = 1$ au temps $2k$, et subit donc une perte $(\widehat{y}_{2k} - y_{2k})^2 = 1$; de même, au temps $2k + 1$, $(\widehat{y}_{2k+1} - y_{2k+1})^2 = (\widehat{y}_{2,2k+1} - y_{2k+1})^2 = 1$. Ainsi (quel que soit le choix de $\widehat{y}_1$), on a $\widehat{L}_n \geqslant n - 1$ pour tout $n \geqslant 1$, tandis que $\max(L_{1,n}, L_{2,n}) \leqslant n/2 + 1$, de sorte que le regret de la stratégie (1.30) est supérieur à $n/2 - 2$. Le regret est linéaire, le regret moyen ne tend donc pas vers 0.

Dans l'Exemple (1.30), la stratégie (1.30) est mise en défaut même pour une classe Θ de faible complexité ($|\Theta| = 2$). À l'inverse, dans le cas statistique, un argument de convergence uniforme (Section 1.1.4) montre que l'excès de risque d'ERM est d'au plus $O(1/\sqrt{n})$.

Agrégation à poids exponentiels. L'Exemple (1.30) met en évidence la fragilité de la minimisation du risque empirique pour des suites arbitraires, due à l'instabilité de ses prédictions. La stratégie d'agrégation à poids exponentiels, décrite ci-dessous, corrige ce défaut en stabilisant les prédictions.

Dans ce qui suit, supposons que l'ensemble Θ est un espace mesurable. Soit π une mesure de probabilité sur Θ, appelée *loi a priori*, et $\eta > 0$ un paramètre appelé *paramètre d'apprentissage* (ou *température inverse*). L'*agrégation à poids exponentiels* (APE), ou *algorithme Hedge* (Vovk, 1998; Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), de paramètre $\eta > 0$ est par définition la stratégie

$$\widehat{y}_t := \int_\Theta \widehat{y}_{\theta,t}\, v_t(\mathrm{d}\theta)\,, \quad \text{où} \qquad \frac{\mathrm{d}v_t}{\mathrm{d}\pi}(\theta) := \frac{e^{-\eta L_{\theta,t-1}}}{\int_\Theta e^{-\eta L_{\theta',t-1}}\pi(\mathrm{d}\theta')}\,. \tag{1.31}$$

Le paramètre d'apprentissage η quantifie l'attache aux données de l'algorithme : pour $\eta \to 0$, l'APE ne dépend pas des données, et effectue la moyenne selon π des prédictions des experts ; pour $\eta \to \infty$, l'APE se réduit à la minimisation du risque empirique[5]. La mesure de probabilité v_t sur Θ (qui donne les "poids" des différents experts $\theta \in \Theta$) est parfois appelée *postérieur* (par analogie avec le postérieur Bayésien, voir la Section 1.4.3).

Remarque 1.3. Dans ce qui suit, nous supposons donnée une notion d'espérance de variables aléatoires à valeurs dans $\widehat{\mathcal{Y}}$ satisfaisant l'inégalité de Jensen pour les fonctions convexes considérées. Cette propriété est par exemple satisfaite lorsque $\widehat{\mathcal{Y}}$ est l'espace des mesures de

[5]Ou, plus précisément, à la moyenne selon π des experts θ de perte cumulée minimale.

probabilité sur $\mathcal{Y}$ (auquel cas l'espérance est le mélange de distributions) et $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est la perte logarithmique, ou lorsque Θ est fini (car les espérances se réduisent à des combinaisons convexes finies).

Afin d'éviter les questions d'intégrabilité et de continuité, on pourra supposer l'ensemble des experts Θ fini. Dans ce cas, la mesure de probabilité $v \in \mathcal{P}(\Theta)$ s'identifie au vecteur $(v_\theta)_{\theta\in\Theta}$ avec $v_\theta := v(\{\theta\})$, et l'intégrale de $h : \Theta \to \mathbf{R}$, $h(\theta) = h_\theta$ vaut $\int_\Theta h\mathrm{d}v = \sum_{\theta\in\Theta} v_\theta h_\theta$. Nous utilisons cependant des notations "fonctionnelles" génériques pour indiquer que les bornes ne dépendent pas explicitement du nombre $|\Theta|$ d'experts et s'étendent au cas de classes infinies, sous réserve d'intégrabilité.

1.2.4 Le cas des pertes exp-concaves

Dans cette section, nous analysons le comportement de l'APE pour une classe de fonctions de perte, dites "exp-concaves".

Concavité exponentielle. La concavité exponentielle est une propriété de courbure de la fonction de perte, analogue à celle de forte convexité.

Définition 1.4 (Concavité exponentielle)**.** Supposons que $\widehat{\mathcal{Y}}$ est un espace convexe, et soit $\eta > 0$. Une fonction $f : \widehat{\mathcal{Y}} \to \mathbf{R}$ est dite *η-exp-concave* si $\exp(-\eta f) : \widehat{\mathcal{Y}} \to \mathbf{R}$ est concave. De même, une fonction $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est dite *η-exp-concave* si $\ell(\cdot, y)$ l'est pour tout $y \in \mathcal{Y}$.

Une fonction exp-concave est convexe (par composition avec la fonction convexe décroissante $-\eta^{-1}\log$). Une fonction η-exp-concave est également η'-exp-concave pour tout $\eta' < \eta$ (par concavité de $x \mapsto x^{\eta'/\eta}$). De plus, une fonction $f : \Omega \to \mathbf{R}$ deux fois différentiable avec Ω un ouvert convexe de $\mathbf{R}^d$ est η-exp-concave si et seulement si

$$0 \succcurlyeq \nabla^2 \exp(-\eta f) = -\eta \exp(-\eta f)\left[\nabla^2 f - \eta(\nabla f)(\nabla f)^\top\right],$$

c'est-à-dire si et seulement si $\nabla^2 f \succcurlyeq \eta(\nabla f)(\nabla f)^\top$. La concavité exponentielle est donc une propriété de courbure, qui stipule que la Hessienne est minorée dans la direction du gradient.

Exemple 1.12 (Apprentissage supervisé)**.** Considérons le cas de l'apprentissage séquentiel supervisé (Section 1.2.1). Si $\widehat{\mathcal{Y}}$ est convexe et si $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est η-exp-concave, alors la fonction $\ell : (g, (x, y)) \mapsto \ell(g(x), y)$ est η-exp-concave.

Exemple 1.13 (Pertes classiques)**.** Considérons les fonctions de perte suivantes :

- la perte logarithmique (Exemple 1.2) est 1-exp-concave, car $\exp(-\ell(f, z)) = f(z)$ est linéaire en f.
- la perte quadratique $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ sur $\widehat{\mathcal{Y}} \times \mathcal{Y} = [-B, B]^2$ est $1/(8B^2)$-exp-concave.
- la perte absolue $\ell(\widehat{y}, y) = |\widehat{y} - y|$ sur $[0, 1]^2$ est convexe mais n'est pas exp-concave.

Remarque 1.4 (Mélangeabilité). Signalons également une généralisation de la convexité exponentielle, la *mélangeabilité* introduite par Vovk (Vovk, 1998; Haussler et al., 1998; Cesa-Bianchi and Lugosi, 2006). Une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ est dite η-mélangeable si, pour tous $M \geqslant 1$, $\widehat{y}_1, \ldots, \widehat{y}_M \in \mathcal{G}$ et $v_1, \ldots, v_M \geqslant 0$ tels que $\sum_{i=1}^M v_i = 1$, il existe un élément $\widehat{y} \in \mathcal{G}$ tel que, pour tout $y \in \mathcal{Y}$,

$$\exp(-\eta\ell(\widehat{y}, y)) \geqslant \sum_{i=1}^M v_i \exp(-\eta\ell(\widehat{y}_i, y)).$$

Une perte η-exp-concave est η-mélangeable, en prenant pour $\widehat{y}$ la combinaison $\sum_{i=1}^{M} v_i \widehat{y}_i$. La notion de mélangeabilité est donc une extension de la convexité exponentielle, qui est indépendante de la paramétrisation et de l'éventuelle structure convexe de $\widehat{\mathcal{Y}}$. La perte logarithmique étant η-mélangeable si et seulement si $\eta \leqslant 1$, la notion de mélangeabilité n'apporte pas de gain pour cette perte. En revanche, la perte quadratique sur $[-B, B]$ est $1/(2B^2)$-mélangeable (Haussler et al., 1998; Vovk, 1998) ; la mélangeabilité permet donc un gain d'un facteur 4 dans le paramètre de "courbure" η, en ayant recours à une combinaison $\widehat{g}$ différente de la combinaison convexe (Vovk, 1990). En revanche, cette combinaison dépend de la borne supposée connue B sur les valeurs de $y, \widehat{y}$, ce qui la rend plus difficilement applicable en pratique.

Dans la plupart des exemples usuels, la notion de mélangeabilité coïncide avec celle de concavité exponentielle dans une paramétrisation bien choisie. Pour cette raison, nous nous restreignons dans ce texte à la seconde notion.

Remarque 1.5 (Extensions stochastiques). Tout comme la forte convexité, la concavité exponentielle permet d'obtenir des vitesses améliorées. Il existe également, dans le cas statistique, des extensions "stochastiques" des notions de concavité exponentielle et de mélangeabilité, dépendant de la loi P et permettant des vitesses rapides même dans le cas de la classification (Juditsky et al., 2008; Audibert, 2009; van Erven et al., 2015). Ces conditions sont des hypothèses de marge intimement liées à la condition de Bernstein (Définition 1.2) ; nous renvoyons à van Erven et al. (2015) pour plus de détails.

Réduction à la perte de mélange. Si la fonction de perte ℓ est η-exp-concave, alors pour toute mesure de probabilité $v \in \mathcal{P}(\Theta)$, toutes prédictions $(\widehat{y}_\theta)_{\theta\in\Theta}$ des experts et tout $y \in \mathcal{Y}$,

$$\ell\Big(\int_\Theta \widehat{y}_\theta \, v(\mathrm{d}\theta), y\Big) \leqslant -\frac{1}{\eta} \log\Big(\int_\Theta e^{-\eta\ell(\widehat{y}_\theta, y)} v(\mathrm{d}\theta)\Big) . \tag{1.32}$$

Cette inégalité montre qu'il est possible, dans le cas de stratégies utilisant une combinaison convexe des prédictions des experts, de réduire l'agrégation d'experts à un problème d'apprentissage séquentiel général (Section 1.2.1), défini par la fonction de perte suivante.

Définition 1.5 (Perte de mélange)**.** Soit $\mathcal{G} = \mathcal{P}(\Theta)$ et $\mathcal{Z}$ l'espace des fonctions mesurables $\Theta \to \mathbf{R}$. La *perte de mélange* est la fonction de perte $\ell_{\mathrm{mix}} : \mathcal{G} \times \mathcal{Z} \to \mathbf{R} \cup \{-\infty\}$ (à valeurs finies si $h \geqslant 0$ ou si Θ est fini) définie par

$$\ell_{\mathrm{mix}}(v, h) := -\log\Big(\int_\Theta e^{-h(\theta)} v(\mathrm{d}\theta)\Big) . \tag{1.33}$$

Pour tout $t \geqslant 1$, soit $h_t : \Theta \to \mathbf{R}$ (*i.e.*, $h_t \in \mathcal{Z}$) la fonction définie par $h_t(\theta) := \eta \cdot \ell(\widehat{y}_{\theta,t}, y_t)$. Il résulte de l'inégalité (1.32) que, si $\widehat{y}_t = \int_\Theta \widehat{y}_{\theta,t} v_t(\mathrm{d}\theta)$, alors

$$\ell(\widehat{y}_t, y_t) \leqslant \frac{1}{\eta} \ell_{\mathrm{mix}}(v_t, h_t) .$$

Puisqu'en outre $\ell(\widehat{y}_{\theta,t}, y_t) = h_t(\theta)/\eta = \ell_{\mathrm{mix}}(\delta_\theta, h_t)/\eta$ pour tout $\theta \in \Theta$, on a

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \sum_{t=1}^{n} \ell(\widehat{y}_{\theta,t}, y_t) \leqslant \frac{1}{\eta}\Bigg(\sum_{t=1}^{n} \ell_{\mathrm{mix}}(v_t, h_t) - \sum_{t=1}^{n} \underbrace{\ell_{\mathrm{mix}}(\delta_\theta, h_t)}_{h_t(\theta)}\Bigg) .$$

Ainsi, le regret pour l'agrégation d'experts est majoré par (η^{-1} fois) le regret pour la perte $\ell_{\mathrm{mix}} : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ par rapport à la classe $\mathcal{F} = \{\delta_\theta : \theta \in \Theta\}$.

Regret de l'APE. Nous considérons maintenant l'APE appliquée à la perte de mélange. Avec la définition de h_t précédente, le postérieur v_t de l'APE (1.31) s'écrit

$$v_t = \frac{\exp\big(-\sum_{s=1}^{t-1} h_s\big)}{\int_\Theta \exp\big(-\sum_{s=1}^{t-1} h_s\big)\mathrm{d}\pi} \cdot \pi\,,$$

soit :

$$v_1 = \pi\,, \qquad v_{t+1} = \frac{\exp(-h_t)}{\int_\Theta \exp(-h_t)\mathrm{d}v_t} \cdot v_t\,. \tag{1.34}$$

Dans ce qui suit, nous notons $\mathrm{KL}(\rho,\pi) := \int_\Theta \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}\mathrm{d}\rho$ la *divergence de Kullback-Leibler* de ρ par rapport à π (voir la Section 1.6).

Proposition 1.4 (Vovk, 1998; Littlestone and Warmuth, 1994)**.** *Considérons la procédure d'APE donnée par (1.34). Alors, pour toute suite $h_1,\dots,h_n \in \mathcal{Z}$ et toute mesure de probabilité $\rho \in \mathcal{P}(\Theta)$, on a*

$$\sum_{t=1}^n \ell_{\mathrm{mix}}(v_t,h_t) - \int_\Theta \Big(\sum_{t=1}^n h_t(\theta)\Big)\rho(\mathrm{d}\theta) \leqslant \mathrm{KL}(\rho,\pi)\,. \tag{1.35}$$

Proof. Cette proposition est une conséquence de la formule variationnelle de Donsker-Varadhan (Théorème 1.18 de l'annexe technique, Section 1.6). En effet, en combinant l'équation (1.34) avec l'identité (1.122) (avec $f = -h_t$, $\rho = v_t$), il vient pour tous $t = 1,\dots,n$ et $\rho \in \mathcal{P}(\Theta)$,

$$\ell_{\mathrm{mix}}(v_t,h_t) - \int_\Theta h_t(\theta)\rho(\mathrm{d}\theta) = \mathrm{KL}(\rho,v_t) - \mathrm{KL}(\rho,v_{t+1})\,.$$

La borne (1.35) s'obtient alors en sommant sur $t = 1,\dots,n$, en simplifiant la somme télescopique et en utilisant le fait que $v_1 = \pi$ et $\mathrm{KL}(\rho,v_{n+1}) \geqslant 0$. □

De la Proposition 1.4 et de la réduction précédente découlent le résultat suivant.

Corollaire 1.2 (Regret de l'APE : cas exp-concave)**.** *Si la fonction ℓ est η-exp-concave, alors l'APE (1.31) de paramètre η satisfait, pour tout $\rho \in \mathcal{P}(\Theta)$:*

$$\sum_{t=1}^n \ell(\widehat{y}_t,y_t) - \int_\Theta \Big(\sum_{t=1}^n \ell(\widehat{y}_{\theta,t},y_t)\Big)\rho(\mathrm{d}\theta) \leqslant \frac{\mathrm{KL}(\rho,\pi)}{\eta} \tag{1.36}$$

quelles que soient les suites $y_1,\dots y_n \in \mathcal{Y}$ et $(\widehat{y}_{\theta,t})_{\theta\in\Theta,1\leqslant t\leqslant n}$. En particulier, si Θ est fini (ou dénombrable), pour tout $\theta \in \Theta$,

$$\sum_{t=1}^n \ell(\widehat{y}_t,y_t) - \sum_{t=1}^n \ell(\widehat{y}_{\theta,t},y_t) \leqslant \frac{\log(\pi_\theta^{-1})}{\eta}\,; \tag{1.37}$$

si $\Theta = \{1,\dots,M\}$ et π est la loi uniforme sur Θ, alors

$$\sum_{t=1}^n \ell(\widehat{y}_t,y_t) - \min_{1\leqslant i\leqslant M} \sum_{t=1}^n \ell(\widehat{y}_{i,t},y_t) \leqslant \frac{\log M}{\eta}\,. \tag{1.38}$$

La borne (1.36) contrôle la différence entre la perte de l'APE et la perte des experts moyennée selon une certaine loi ρ, en fonction de la "complexité" $\mathrm{KL}(\rho, \pi)$. Notons que cette borne est valide simultanément sur toutes les lois ρ, en particulier pour tout $K = \mathrm{KL}(\rho, \pi)$, pour une même procédure. Bien qu'elle implique les bornes (1.37) et (1.38) dans le cas fini, cette borne ne dépend pas explicitement du nombre $|\Theta|$ d'experts, ni même de la "complexité" de la classe Θ des experts. Il est alors possible d'optimiser le borne supérieure (1.36) sur ρ, par un argument de dualité (Théorème 1.18 de la Section 1.6.2) :

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) \leqslant \inf_{\rho \in \mathcal{P}(\Theta)} \left\{ \int_\Theta L_{\theta,n} \, \rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho, \pi)}{\eta} \right\} = -\frac{1}{\eta} \log \left(\int_\Theta e^{-\eta L_{\theta,n}} \, \pi(\mathrm{d}\theta) \right).$$

La quantité du membre de droite dépend de la loi de la perte $L_{\theta,n}$ selon π. En particulier, pour toute partie $E \subset \Theta$ telle que $\pi(E) > 0$, le choix de $\rho = \pi(E)^{-1} \mathbf{1}(\theta \in E) \cdot \pi$ donne

$$\widehat{L}_t - \sup_{\theta \in E} L_{\theta,n} \leqslant \frac{\log \pi(E)^{-1}}{\eta} ;$$

ainsi, pour tout $\varepsilon \in (0, 1)$, le regret de l'APE par rapport à la fraction ε (au sens de la loi π sur Θ) des meilleurs experts est d'au plus $\log(1/\varepsilon)/\eta$. Cette borne est d'autant plus petite que la loi a priori π attribue une probabilité importante à des paramètres $\theta \in \Theta$ de faible perte. Notons enfin qu'un regret constant (1.38) en $O((\log M)/\eta)$ correspond à un regret moyen en $O((\log M)/(\eta n))$, c'est-à-dire à une vitesse rapide.

1.2.5 Pertes (convexes) bornées et problème de Hedge

Dans cette section, nous ne supposons plus que la perte ℓ est η-exp-concave. Nous supposons cependant que $0 \leqslant \ell \leqslant B$ pour un certain $B > 0$.

Réduction à la perte linéaire. Supposons la fonction $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$ convexe en son premier argument. Si $\widehat{y}_t = \int_\Theta \widehat{y}_{\theta,t} v_t(\mathrm{d}\theta)$, alors l'inégalité de Jensen implique que

$$\ell(\widehat{y}_t, y_t) \leqslant \int_\Theta \ell(\widehat{y}_{\theta,t}, y_t) v_t(\mathrm{d}\theta) = \langle v_t, h_t \rangle ,$$

où $h_t : \Theta \to [0, B]$ est la fonction définie par $h_t(\theta) = \ell(\widehat{y}_{\theta,t}, y_t)$ et où $\langle v, h \rangle := \int_\Theta h \mathrm{d}v$. À l'instar du cas exp-concave (Section 1.2.4), l'agrégation d'experts se ramène alors au problème d'apprentissage séquentiel avec $\mathcal{G} = \mathcal{P}(\Theta)$, $\mathcal{Z}$ l'espace des fonctions mesurables $\Theta \to [0, B]$, ℓ la fonction de perte *linéaire* $(v, h) \mapsto \langle v, h \rangle$ et $\mathcal{F} = \{\delta_\theta : \theta \in \Theta\}$. Ce problème d'apprentissage est appelé *problème de Hedge* (Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), et coïncide avec l'optimisation linéaire en ligne sur le simplexe $\mathcal{P}(\Theta)$.

Si $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$ n'est pas convexe, il est possible de se ramener à ce cas en considérant des prédictions randomisées (Remarque 1.2). En effet, si $\widehat{y}_t = \widehat{y}_{\widehat{\theta}_t,t}$, où $\widehat{\theta}_t \sim v_t$ (conditionnellement aux tirages précédents), alors $\mathbb{E}[\ell(\widehat{y}_t, y_t)] = \langle v_t, h_t \rangle$ dès lors que y_t ne dépend pas de $\widehat{\theta}_t$ (mais peut dépendre de v_t ainsi que de $\widehat{\theta}_1, \ldots, \widehat{\theta}_{t-1}$), ce qui est notamment le cas si la suite $y_1, \ldots, y_n$ est déterministe. La réduction au problème de Hedge permet donc de contrôler le regret moyen (sur le tirage de $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$), et des inégalités de concentration sur les martingales impliquent que le regret est proche de son espérance avec forte probabilité (Cesa-Bianchi and Lugosi, 2006).

Regret de l'APE pour le problème de Hedge. Avec les notations précédentes, l'APE de paramètre $\eta > 0$ pour l'agrégation d'experts (1.31) coïncide avec la stratégie suivante pour le problème linéarisé :

$$v_t = \frac{\exp\big(-\eta \sum_{s=1}^{t-1} h_s\big)}{\int_\Theta \exp\big(-\eta \sum_{s=1}^{t-1} h_s\big) \mathrm{d}\pi} \cdot \pi \,. \tag{1.39}$$

Cet algorithme admet la garantie de regret suivante (Freund and Schapire, 1997; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006) :

Proposition 1.5 (Regret de l'algorithme Hedge)**.** *Pour tous $\eta, B > 0$ et $n \geqslant 1$, toute suite de fonctions $h_1, \dots, h_n : \Theta \to [0, B]$ et toute loi $\rho \in \mathcal{P}(\Theta)$, l'algorithme Hedge (1.39) de paramètre $\eta > 0$ satisfait :*

$$\sum_{t=1}^n \langle v_t, h_t \rangle - \sum_{t=1}^n \langle \rho, h_t \rangle \leqslant \frac{\mathrm{KL}(\rho, \pi)}{\eta} + \frac{\eta B^2 n}{8} \,. \tag{1.40}$$

En particulier, pour tout $K > 0$, le choix de $\eta = B^{-1}\sqrt{8K/n}$ conduit à une borne de regret de $B\sqrt{nK/2}$ par rapport à la classe $\mathcal{F}_K := \{\rho \in \mathcal{P}(\Theta) : \mathrm{KL}(\rho, \pi) \leqslant K\}$. Si $\Theta = \{1, \dots, M\}$ est fini et π est la mesure uniforme sur Θ, le choix $\eta = B^{-1}\sqrt{8(\log M)/n}$ donne :

$$\sum_{t=1}^n \langle v_t, h_t \rangle - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^n h_{i,t} \leqslant B\sqrt{\frac{n \log M}{2}} \,. \tag{1.41}$$

Proof. Pour $t = 1, \dots, n$, posons $h'_t := \eta \cdot h_t$. L'inégalité de Hoeffding (Lemme 1.2, Section 1.6) avec $\lambda = -\eta B$, appliquée à la variable $h_t(\widehat{\theta}_t)/B$, $\widehat{\theta}_t \sim v_t$, à valeurs dans $[0, 1]$, donne

$$\langle v_t, h_t \rangle \leqslant -\frac{1}{\eta} \log \int_\Theta e^{-\eta\, h_t(\theta)} v_t(\mathrm{d}\theta) + \frac{\eta B^2}{8} = \frac{1}{\eta} \ell_{\mathrm{mix}}(v_t, h'_t) + \frac{\eta B^2}{8} \,.$$

La borne de regret (1.40) s'obtient alors en sommant sur $t = 1, \dots, n$, en notant que l'algorithme Hedge (1.39) coïncide avec l'APE sur la suite $h'_1, \dots, h'_n$, et en utilisant la Proposition 1.4. Les assertions restantes s'en déduisent en optimisant le choix de η et en prenant $\rho = \delta_i$ dans le cas où $\Theta = \{1, \dots, M\}$. □

La borne (1.41) correspond à un regret moyen d'au plus $O(\sqrt{(\log M)/n})$. On retrouve ainsi la même vitesse (lente) qu'ERM pour l'apprentissage statistique sur des classes finies et avec perte bornée (Section 1.1.4). La borne de regret de $O(\sqrt{n \log M})$ s'avère optimale au sens minimax dans le cas de classes finies : pour tout algorithme pour le problème de Hedge, il existe une suite de fonctions $h_t : \{1, \dots, M\} \to [0, 1]$, $1 \leqslant t \leqslant n$, pour laquelle le regret de l'algorithme est d'au moins $\Theta(\sqrt{n \log M})$ (Cesa-Bianchi and Lugosi, 2006). Cela découle aussi du fait que, par conversion online-to-batch (Proposition 1.3), le regret minimax (moyen) est supérieur à l'excès de risque minimax pour la classification avec classes finies à M classifieurs, qui est de $\Theta(\sqrt{(\log M)/n})$.

1.2.6 Algorithmes adaptatifs pour le problème de Hedge

Dans cette section, nous considérons le problème de Hedge de la section précédente, en supposant de plus que $\Theta = \{1, \dots, M\}$ et $B = 1$. Ainsi, les fonctions h_t ($t = 1, \dots, M$) s'identifient

à des vecteurs de perte dans $[0,1]^M$, tandis que les lois de probabilités $v_t \in \mathcal{P}(\Theta)$ s'identifient aux vecteurs $(v_{i,t})_{1\leqslant i\leqslant M} \in \mathbf{R}_+^M$ tels que $\sum_{i=1}^M v_{i,t} = 1$. Pour $t = 1, \dots, n$, nous notons

$$H_{i,t} := \sum_{s=1}^{t} h_{i,s}\,, \qquad \widehat{H}_t := \sum_{s=1}^{t} \langle v_t, h_t\rangle$$

les pertes cumulées au temps t de l'expert i et de l'algorithme, respectivement. Nous considérons également des bornes de regret uniformes par rapport aux vecteurs $\rho \in \mathcal{P}(\Theta)$, c'est-à-dire par rapport aux *experts* $i = 1, \dots, M$, telles que la borne (1.41).

Borne minimax, horizon de temps. Par la Proposition 1.5, l'algorithme Hedge avec loi à priori π uniforme, soit

$$v_{i,t} = \frac{e^{-\eta H_{i,t-1}}}{\sum_{j=1}^M e^{-\eta H_{j,t-1}}}$$

pour tous $t = 1, \dots, M$ et $1 \leqslant i \leqslant M$, et de paramètre $\eta = c\sqrt{(\log M)/n}$ (où $c, C > 0$ désignent des constantes numériques), admet la borne de regret

$$\mathsf{Reg}_n := \widehat{H}_t - \min_{1\leqslant i\leqslant M} H_{i,t} = \sum_{t=1}^{n} \langle v_t, h_t\rangle - \min_{1\leqslant i\leqslant M} \sum_{t=1}^{n} h_{i,t} \leqslant C\sqrt{n\log M}\,.$$

(Freund and Schapire, 1997; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006), ce qui correspond au regret minimax (Cesa-Bianchi and Lugosi, 2006). Notons que le choix du paramètre $\eta \asymp \sqrt{(\log M)/n}$ dépend de l'horizon de temps n considéré. Il peut être souhaitable d'obtenir la borne de regret minimax de $O(\sqrt{n\log M})$ *simultanément* pour tout $n \geqslant 1$, pour un même algorithme.

Une façon générique d'obtenir une telle garantie est la *technique du doublement* (*doubling trick* en anglais), qui consiste à utiliser l'algorithme Hedge de paramètre $\eta_p = c\sqrt{(\log M)/2^p}$ sur les intervalles de temps "géométriques" $\{2^p, \dots, 2^{p+1} - 1\}$ (pour $p \geqslant 0$), en réinitialisant les poids à la fin de chaque intervalle (Cesa-Bianchi et al., 1997; Cesa-Bianchi and Lugosi, 2006). On obtient alors, pour tout $n \geqslant 1$, en notant p l'entier tel que $2^p \leqslant n < 2^{p+1}$, un regret d'au plus

$$C\sum_{k=0}^{p} \sqrt{2^k \log M} \leqslant \frac{C}{1-1/\sqrt{2}}\sqrt{2^p \log M} = O(\sqrt{n\log M})\,.$$

Une autre façon d'obtenir cette garantie est d'utiliser un paramètre d'apprentissage η_t variable, c'est-à-dire de choisir à l'instant $t \geqslant 1$,

$$v_{i,t} = \frac{e^{-\eta_t H_{i,t-1}}}{\sum_{j=1}^M e^{-\eta_t H_{j,t-1}}}\,, \tag{1.42}$$

où η_t dépend de t et éventuellement de $h_1, \dots, h_{t-1} \in [0,1]^M$. En effet, pour toute suite décroissante $\eta_1, \eta_2, \dots$ de paramètres, le regret de la stratégie (1.42) satisfait :

$$\mathsf{Reg}_n \leqslant \frac{\log M}{\eta_n} + \frac{1}{8}\sum_{t=1}^{n} \eta_t$$

(Chernov and Zhdanov, 2010). En particulier, le choix $\eta_t = c\sqrt{(\log M)/t}$ $(t \geqslant 1)$ conduit à une borne de regret en $C\sqrt{n\log M}$ pour tout $n \geqslant 1$.

Dépendance en la complexité : bornes de "quantiles". Considérons le cas où le nombre d'experts M est élevé. Il est alors souhaitable d'obtenir des bornes avec une faible dépendance en M, quitte à considérer une classe de comparaison moins "complexe" que l'ensemble des experts $\{1,\dots,M\}$. La Proposition 1.5 affirme que l'algorithme Hedge de paramètre $\eta = c\sqrt{K/n}$ admet une borne de regret d'au plus $C\sqrt{K/n}$ par rapport à tout mélange des experts selon une loi ρ sur $\{1,\dots,M\}$ telle que $\mathrm{KL}(\rho,\pi)$. Comme montré dans la discussion suivant le Corollaire 1.2, en choisissant $K = \log(1/\varepsilon)$ pour $\varepsilon \in (0,1)$, ceci implique une borne de regret de $\sqrt{\log(1/\varepsilon)/n}$ par rapport au $\lceil \varepsilon M \rceil$-meilleur expert, indépendamment de M.

Cependant, cette borne est atteinte en choisissant η en fonction de K (ou, de manière équivalente, en fonction de ε), et n'est donc valable que pour cette valeur de ε. À l'inverse, le Corollaire 1.2 montre que l'APE atteint dans le cas exp-concave la borne de regret $C\log(1/\varepsilon)/n$ simultanément pour tout $\varepsilon \in (0,1)$, par un choix de η indépendant pas de ε. Dans le cas de pertes linéaires considéré ici, il découle de la Proposition 1.5 que le choix de $\eta = c/\sqrt{n}$ conduit à une borne de regret de $C \cdot \mathrm{KL}(\rho,\pi)/\sqrt{n}$ pour tout ρ, c'est-à-dire de $C \cdot \log(1/\varepsilon)/\sqrt{n}$. Cette borne admet une dépendance sous-optimale en ε. Un axe de recherche consiste à obtenir des algorithmes adaptatifs à la "complexité" ε, c'est-à-dire un regret de $C\sqrt{\log(1/\varepsilon)/n}$ pour tout $\varepsilon > 0$ (Chaudhuri et al., 2009; Chernov and Vovk, 2010; Luo and Schapire, 2014, 2015; Koolen and van Erven, 2015) ; ces bornes sont appelées "bornes de quantiles", car elles contrôlent le regret par rapport aux quantiles de niveau $\varepsilon \in (0,1)$ des experts (ordonnés par leur perte).

Au-delà du minimax : algorithmes adaptatifs. L'attrait des bornes minimax est leur robustesse : elles sont en effet valables pour toute suite $h_1,\dots,h_n \in [0,1]^M$ de vecteurs de pertes, et en particulier sans hypothèse de stochasticité. Cependant, la vitesse en $O(\sqrt{n\log M})$, bien qu'optimale dans le pire des cas, peut s'avérer trop pessimiste dans certaines situations. Une littérature importante est dédiée à la conception d'algorithmes *adaptatifs*, qui combinent le regret minimax en $O(\sqrt{n\log M})$ avec des garanties améliorées lorsque la suite $h_1,\dots,h_n$ présente certaines régularités (Cesa-Bianchi et al., 1997; Auer et al., 2002; Cesa-Bianchi et al., 2007; de Rooij et al., 2014; Gaillard et al., 2014; Koolen et al., 2014; Sani et al., 2014; Koolen and van Erven, 2015; Luo and Schapire, 2015; Wintenberger, 2017).

Une première situation où il est possible d'obtenir des garanties améliorées de regret correspond au cas où la perte moyenne $H_n^*/n = \min_{1\leqslant i\leqslant n} H_{i,n}/n$ du meilleur expert est faible. Les bornes de regret dites du *premier ordre* (Cesa-Bianchi et al., 1997; Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006), de la forme $\mathsf{Reg}_n \leqslant C(\sqrt{H_n^*\log M} + \log M)$, permettent d'exploiter cette régularité : elles impliquent un regret d'au plus $O(\sqrt{n\log M})$ dans le pire des cas, mais améliorent cette garantie lorsque $H_n^* \ll n$. Cette garantie est satisfaite par l'algorithme Hedge (1.42) avec $\eta_t = \eta = c\sqrt{(\log M)/(1\vee H_n^*)}$ lorsque H_n^* est connu a priori (Cesa-Bianchi and Lugosi, 2006) ; cette quantité étant généralement inconnue, il est possible d'obtenir la garantie précédente pour tout $H_n^* \in [0,n]$ par la technique du doublement (Cesa-Bianchi et al., 1997), ou en prenant $\eta_t = c\sqrt{(\log M)/(1\vee H_{t-1}^*)}$ (Auer et al., 2002).

La borne du premier ordre permet d'obtenir un regret constant en $O(\log M)$ dans le cas où $H_n^* = O(1)$, c'est-à-dire lorsque le meilleur expert admet une perte négligeable. Cependant, lorsque le meilleur expert admet une perte moyenne non négligeable (ce qui est typiquement le cas pour les problèmes d'apprentissage), la borne du premier ordre n'améliore pas la vitesse du regret, mais au mieux la constante du terme dominant. Cette garantie apporte donc une adaptativité limitée aux propriétés de la suite $h_1,\dots,h_n$: elle n'exploite pas d'autres formes de régularité qui devraient permettre d'obtenir un regret amélioré, par exemple le fait que

l'un des experts prédise mieux que les autres (sans pour autant avoir une perte négligeable). Un second type de garantie, qui raffine la borne du premier ordre, est donné par les bornes du *second ordre* (Cesa-Bianchi et al., 2007; de Rooij et al., 2014; Gaillard et al., 2014; Koolen and van Erven, 2015). Celles-ci dépendent de quantités du second ordre, comme la somme au cours du temps des variances des pertes $(h_{i,t})_{1\leqslant i\leqslant M}$ selon la loi de probabilité $v_t = (v_{i,t})_{1\leqslant i\leqslant M}$ (Cesa-Bianchi et al., 2007; de Rooij et al., 2014), ou des quantités liées (Gaillard et al., 2014; Koolen and van Erven, 2015). Ces bornes sont atteintes par des algorithmes plus sophistiqués, qui calibrent η_t en fonction de ces quantités du second ordre (Cesa-Bianchi et al., 2007; de Rooij et al., 2014), ou reposent sur d'autres choix de poids que les poids exponentiels (1.42), dans le cas de Gaillard et al. (2014); Wintenberger (2017); Koolen and van Erven (2015). Le comportement typique de ces algorithmes est le suivant :

- si la suite $h_1, \dots, h_n$ est "difficile", c'est-à-dire choisie de manière adverse comme dans le pire des cas, alors l'algorithme adoptera un comportement conservateur, similaire à celui de l'algorithme Hedge : les poids ne seront pas excessivement concentrés, en raison d'un niveau non négligeable de régularisation. Dans le cas d'algorithmes à poids exponentiels de type (1.42), cela signifie que $\eta_t \lesssim \sqrt{(\log M)/t}$;
- si au contraire la suite $h_1, \dots, h_n$ est "favorable", c'est-à-dire présente certaines régularités exploitées par l'algorithme, alors celui-ci adoptera un comportement plus "agressif", et aura tendance à éliminer plus rapidement les experts sous-optimaux et à se concentrer sur le meilleur expert. Dans le cas d'un algorithme à poids exponentiels (1.42), cela revient à choisir un paramètre η_t plus élevé, par exemple $\eta_t \gtrsim c$, et donc à se rapprocher de la minimisation du risque empirique.

Les bornes du second ordre impliquent celles du premier ordre (qui impliquent elles-mêmes la borne minimax "d'ordre 0" en $O(\sqrt{n \log M})$), mais les améliorent sensiblement dans certains cas. Par exemple, si l'un des experts domine les autres, alors les poids de l'algorithme vont avoir tendance à se concentrer sur le meilleur expert, ce qui réduira la variance des pertes ; ceci conduit en retour l'algorithme à éliminer les experts sous-optimaux de manière plus agressive, ce qui réduit la variance des pertes d'autant plus vite.

Un type naturel de régularité considéré dans la littérature est le cas *stochastique*, où les vecteurs de pertes $h_1, \dots, h_n \in [0,1]^M$ sont des variables aléatoires i.i.d.[6] (van Erven et al., 2011; Gaillard et al., 2014; Luo and Schapire, 2015). Cela couvre notamment le cas où $h_{i,t} = \ell(f_i, Z_t)$ pour $1 \leqslant i \leqslant M$, avec $f_1, \dots, f_M \in \mathcal{F}$ et $Z_1, \dots, Z_n$ sont des variables i.i.d. de même loi P. Supposons dans ce cas que le meilleur expert $i^* = \arg\min_{1\leqslant i\leqslant n} \mathbb{E}[h_{i,1}]$ est unique. On note alors

$$\Delta := \arg\min_{i \neq i^*} \mathbb{E}[h_{i,t} - h_{i^*,t}] > 0\,. \tag{1.43}$$

Le paramètre Δ, qui correspond à l'écart entre le meilleur expert et les autres, est une mesure de la difficulté du problème : si Δ est suffisamment élevé, alors le meilleur expert aura tendance à dominer rapidement, et sera donc clairement distinguable. Dans ce cas, Gaillard et al. (2014) montre qu'une borne de regret du second ordre implique un regret d'au plus $O((\log M)/\Delta)$; cette garantie améliore la borne minimax de $O(\sqrt{n \log M})$ dès lors que $\Delta \gtrsim \sqrt{(\log M)/n}$, et correspond à un regret constant. La même borne du second ordre (combinée à une garantie de type "quantile") est également atteinte par une procédure de Koolen and van Erven (2015) ;

[6] Notons que l'on suppose ici l'indépendance des pertes $h_{i,t}$ entre les différents instants $t \geqslant 1$, mais pas nécessairement entre les différents experts $1 \leqslant i \leqslant M$ à un même instant.

par ailleurs, un algorithme proposé par Luo and Schapire (2015), qui admet une garantie de regret plus faible qu'une garantie de second ordre, combine également le regret minimax en $O(\sqrt{n \log M})$ avec la garantie améliorée de $O((\log M)/\Delta)$.

Plus généralement, Koolen et al. (2016) ont montré que, si les pertes $(h_{i,t})_{1 \leqslant i \leqslant M}$ satisfont la condition de Bernstein de paramètres (β, B), avec $B > 0$ et $\beta \in [0,1]$ (Définition 1.2), c'est-à-dire $\mathbb{E}[(h_{i,t} - h_{i^*,t})^2] \leqslant B \cdot \mathbb{E}[h_{i,t} - h_{i^*,t}]^\beta$ pour tout i, alors le regret par rapport à i^* satisfait

$$\mathbb{E}[\widehat{H}_n - H_{i^*,n}] \leqslant C\big((B \log M)^{\frac{1}{2-\beta}} n^{\frac{1-\beta}{2-\beta}} + \log M\big)\,, \tag{1.44}$$

avec une borne correspondante en forte probabilité (voir aussi la Proposition 4.4 du Chapitre 4 pour une preuve élémentaire de cette borne en espérance, avec dépendance en B).

1.2.7 Optimalité de l'algorithme Hedge dans le cas stochastique (Chapitre 4)

Des résultats de la section précédente, il découle que :

- l'algorithme de Hedge (1.42), avec paramètre d'apprentissage constant $\eta_t = \sqrt{(\log M)/n}$ ou variable $\eta_t = \sqrt{(\log M)/t}$, ou avec la technique du doublement, admet un regret d'au plus $O(\sqrt{n \log M})$, qui correspond au regret minimax ;
- des stratégies plus sophistiquées admettent des garanties adaptatives plus fines (comme les bornes de second ordre), qui impliquent une borne de regret améliorée en $O((\log M)/\Delta)$ dans le cas stochastique (avec Δ défini par (1.43)), en plus de la borne de $O(\sqrt{n \log M})$ dans le pire des cas. Ces algorithmes adaptatifs se comportent typiquement comme Hedge avec un paramètre d'apprentissage conservateur $\eta_t \asymp \sqrt{(\log M)/t}$ pour une suite "adverse", mais se montrent plus agressifs (proches d'ERM) sur des suites favorables (de Rooij et al., 2014).

Les résultats précédents portent sur des *bornes supérieures* sur le regret des différents algorithmes considérés. De telles garanties ne permettent cependant pas à elles seules de conclure quant aux performances et avantages respectifs des différents algorithmes, le regret pouvant être plus faible que ce que ces bornes supérieures suggèrent.

Dans le Chapitre 4, nous étudions le comportement de l'algorithme Hedge dans le cas stochastique. Tout d'abord, nous montrons que Hedge avec un paramètre d'apprentissage décroissant $\eta_t = c\sqrt{(\log M)/t}$, calibré pour le pire des cas, est également adaptatif au cas stochastique des données, où il atteint la même garantie de $O((\log M)/\Delta)$.

Théorème 1.5 (Théorème 4.1, Chapitre 4)**.** *Si les vecteurs de perte $h_1, \dots, h_n$ sont i.i.d., alors en notant Δ le paramètre (1.43), l'algorithme Hedge avec $\eta_t = 2\sqrt{(\log M)/t}$ admet un regret*[7] *d'au plus*

$$\mathbb{E}[\widehat{H}_n - H_{i^*,n}] \leqslant \frac{4 \log M + 25}{\Delta}\,. \tag{1.45}$$

Une borne similaire est également valable sous des hypothèses plus générales, et en forte probabilité (Corollaire 4.1). En outre, la borne de regret (1.45) admet la dépendance optimale en M et Δ : pour tous $M \geqslant 2, \Delta \in (0,1)$ et toute stratégie, il existe une loi de h_1 pour laquelle le regret de cette stratégie est d'au moins $O((\log M)/\Delta)$ (Proposition 4.2). L'algorithme de

[7] La quantité apparaissant dans la borne (1.45) n'est pas tout-à-fait le regret, mais le regret par rapport à i^*. Ces quantités sont toutefois proches, et il est possible d'obtenir des bornes de regret (voir la Remarque 4.3 et le Corollaire 4.1 du Chapitre 4).

Hedge avec le paramètre $\eta_t = c\sqrt{(\log M)/t}$ calibré pour le pire des cas est donc adaptatif à la difficulté du problème, pour tout $\Delta \in (0, 1)$.

À l'inverse, nous établissons que cette adaptativité au cas stochastique n'est pas partagée par les variantes proches de Hedge, également minimax, obtenues avec le paramètre constant $\eta_t = c\sqrt{(\log M)/n}$ ou par la technique du doublement. En effet, la Proposition 4.3 montre que ces algorithmes exhibent un regret en $\Theta(\sqrt{n \log M})$ du pire des cas même pour des problème stochastiques "faciles" avec $\Delta \simeq 1$. En particulier, le choix d'un paramètre d'apprentissage variable conduit à un algorithme plus adaptatif que la technique du doublement, laquelle n'est adéquate que dans le pire des cas. De plus, le paramètre d'apprentissage variable est préférable au paramètre constant, même lorsque l'horizon de temps n est connu.

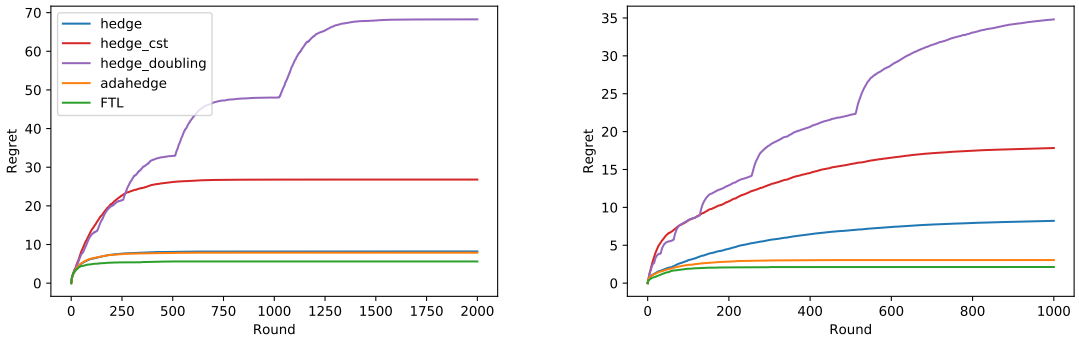
Le Théorème 1.5 montre que l'algorithme de Hedge atteint la même borne de regret dans le cas stochastique que les algorithmes adaptatifs mentionnés dans la Section 1.2.6 (Gaillard et al., 2014; Luo and Schapire, 2015). Il est donc naturel de se demander si ceux-ci apportent un gain dans le cas stochastique. Il s'avère que les algorithmes du second ordre ont bien un avantage sur la variante de Hedge avec paramètre variable. Pour le voir, il est nécessaire de considérer une notion de régularité plus fine que l'écart Δ, à savoir la condition de Bernstein (Définition 1.2). En effet, la borne (1.44) montre qu'une garantie du second ordre implique un regret d'au plus $O(B \log M)$ sur des pertes stochastiques satisfaisant la condition de Bernstein $(1, B)$ avec $B \geqslant 1$. Notons que si $\Delta > 0$, alors cette condition est satisfaite pour $B \leqslant 1/\Delta$. Cependant, cette condition est sensiblement moins restrictive, et peut être satisfaite avec $B = O(1)$ même pour des valeurs arbitrairement faibles de Δ ; nous renvoyons à l'Exemple 4.2 Chapitre 4 pour davantage de détails.

À l'inverse, l'algorithme de Hedge avec paramètre $\eta_t = c\sqrt{(\log M)/t}$ n'est pas adaptatif à la condition de Bernstein. En effet, le Théorème 4.3 implique qu'il existe une loi satisfaisant la condition de Bernstein avec $B = O(1)$, mais pour laquelle cet algorithme admet un regret de $\Theta(\sqrt{n \log M})$. Plus généralement, il s'avère que l'algorithme Hedge avec $\eta_t = c\sqrt{(\log M)/t}$ ne peut pas s'adapter à d'autres régularités que l'écart Δ : quelle que soit la loi de h_1, le regret de cet algorithme pour $n \geqslant C/\Delta^2$ est d'au moins

$$\frac{c}{(\log M)^2 \Delta} .$$

(Théorème 4.4). Ceci caractérise (à un facteur $\log^3 M$ près, en fonction du nombre d'experts presque optimaux) le regret de cet algorithme sur *tout* problème stochastique.

L'avantage d'un paramètre d'apprentissage η_t adaptatif, plus élevé que le paramètre en $c\sqrt{(\log M)/t}$ du pire des cas, sur certains problèmes stochastiques peut se comprendre de la façon suivante. Considérons une loi de h_t avec B peu élevé, mais Δ faible (donc $1/\Delta$ élevé). Le paramètre d'apprentissage en $c\sqrt{(\log M)/t}$ est suffisamment élevé pour éliminer les "mauvais" experts i (tels que $\Delta_i := \mathbb{E}[h_{i,t} - h_{i^*,t}]$ est suffisamment grand) après un nombre d'étapes optimal (en $O((\log M)/\Delta_i^2)$). Cependant, une fois ces experts écartés, les experts presque optimaux (avec $\Delta_i \simeq \Delta$) ne seront éliminés que très tard (après $O((\log M)/\Delta^2)$ étapes). En revanche, la condition de Bernstein implique que les pertes de ces experts sont fortement corrélées à celles de l'expert optimal i^* (voir la Section 1.1.4), de sorte que le niveau de bruit dans leurs pertes relatives $h_{i,t} - h_{i^*,t}$ est faible. Les algorithmes du second ordre utiliseront alors un paramètre d'apprentissage plus élevé, de sorte que ces experts presque optimaux seront éliminés plus tôt et contribueront donc moins au regret.



(a) (b)

Figure 1.1: Regret d'algorithmes d'agrégation d'experts sur deux problèmes stochastiques. Sont évalués les algorithmes Hedge avec paramètre décroissant (`hedge`), constant (`hedge_cst`) et technique du doublement (`hedge_doubling`), l'algorithme adaptatif du second ordre `adahedge` (de Rooij et al., 2014), et la minimisation du risque empirique (`FTL`). (a) Problème stochastique avec un écart ($M = 20$, $\Delta = 0.1$) ; (b) Problème stochastique avec Δ faible, mais satisfaisant la condition de Bernstein ($M = 10$, $\Delta = 0.04$, $B \leqslant 4$).

1.2.8 Prédiction à l'aide d'une classe croissante d'experts (Chapitre 5)

Dans le Chapitre 5, nous étudions une variante du problème de l'agrégation d'experts (Section 1.2.1), dans le cas où la classe d'experts $\Theta_t := \{1, \dots, M_t\}$ s'enrichit au cours du temps : à chaque instant $t \geqslant 1$, $m_t := M_t - M_{t-1} \geqslant 0$ nouveaux experts sont disponibles. Ce problème est motivé par des situations pratiques, où il peut être souhaitable d'incorporer de nouvelles sources de prédictions (comme de nouveaux algorithmes d'apprentissage ou des prédicteurs utilisant de nouvelles variables).

Dans ce contexte, nous considérons plusieurs notions de regret, qui diffèrent par le choix du comparateur : (1) chaque expert, depuis l'instant de son introduction, ce qui revient à se comparer aux suites d'experts ne transitionnant que vers de nouveaux experts ; (2) les suites arbitraires d'experts avec un nombre de changements (soit vers un nouvel expert, soit vers un expert déjà introduit) contrôlé ; (3) les suites d'experts à support "parcimonieux", qui prennent leurs valeurs dans un petit sous-ensemble de "bons" experts.

Dans chaque cas, en étendant les stratégies existantes dans le cas d'une classe fixe d'experts (Vovk, 1998; Herbster and Warmuth, 1998; Vovk, 1999; Bousquet and Warmuth, 2002; Koolen et al., 2012), nous proposons des algorithmes efficaces (avec une complexité linéaire en le temps et le nombre d'experts) admettant des garanties de regret optimales. Ces extensions au cas d'un nombre croissant d'experts utilisent la notion de *spécialistes*, c'est-à-dire d'experts pouvant s'abstenir à certains instants (Freund et al., 1997; Chernov and Vovk, 2009), des (reformulations génériques de) stratégies efficaces équivalentes à l'agrégation de *suites d'experts* avec pour loi a priori une chaîne de Markov (Herbster and Warmuth, 1998; Vovk, 1999; Koolen and de Rooij, 2013; Bousquet and Warmuth, 2002; Koolen et al., 2012), ainsi que la combinaison de ces techniques.

1.3 Régression linéaire et matrices aléatoires

Dans cette section, nous considérons l'un des problèmes d'apprentissage statistique les plus classiques, celui des *moindres carrés*, également appelé *régression linéaire avec design (plan) aléatoire*, ou encore *agrégation linéaire*. Formellement, il s'agit d'un problème de régression (Exemple 1.4), où la classe de comparaison $\mathcal{F}$ est un espace vectoriel de dimension finie $d \geqslant 1$ de fonctions $\mathcal{X} \to \mathbf{R}$. Si $(\varphi_1, \dots, \varphi_d)$ est une base de $\mathcal{F}$, quitte à considérer $X' = \Phi(X) := (\varphi_1(X), \dots, \varphi_d(X))$ on peut supposer que $\mathcal{X} = \mathbf{R}^d$ et que $\mathcal{F}$ est la classe des fonctions linéaires $\mathbf{R}^d \to \mathbf{R}$, données par $f_\beta(x) = \langle \beta, x \rangle$ pour $x \in \mathbf{R}^d$.

Soit (X, Y) un couple de loi P sur $\mathbf{R}^d \times \mathbf{R}$. On suppose que $\mathbb{E}[Y^2] < +\infty$ et que $\mathbb{E}[\|X\|^2] < +\infty$. Le risque associé à la fonction de perte quadratique s'écrit alors pour $\beta \in \mathbf{R}^d$:

$$R(\beta) := R(f_\beta) = \mathbb{E}[(Y - \langle \beta, X \rangle)^2] \,.$$

En supposant que la matrice de covariance[8] $\Sigma := \mathbb{E}[XX^\top]$ est inversible (ce que l'on peut toujours faire, quitte à se restreindre au sous-espace vectoriel de $\mathbf{R}^d$ engendré par le support de X), le risque admet pour unique minimum $\beta^* = \Sigma^{-1}\mathbb{E}[YX]$. De plus, l'excès de risque de tout $\beta \in \mathbf{R}^d$ vaut

$$\mathcal{E}(\beta) = \mathbb{E}[\langle \beta - \beta^*, X \rangle^2] = \|\beta - \beta^*\|_\Sigma^2 \,, \tag{1.46}$$

où l'on note $\|u\|_\Sigma := \langle \Sigma u, u \rangle^{1/2}$ pour $u \in \mathbf{R}^d$. Étant donné un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. de loi P, le problème consiste à produire un prédicteur $\widehat{\beta}_n$ de faible excès de risque. Un estimateur naturel pour ce problème est donné par le minimiseur du risque empirique (1.13), appelé dans ce cas *estimateur des moindres carrés* (en anglais *Ordinary Least Squares estimator*, abrégé OLS), qui s'écrit ici

$$\widehat{\beta}_n^{\mathrm{LS}} = \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \beta, X_i \rangle)^2 \right\} = \widehat{\Sigma}_n^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i \,, \tag{1.47}$$

où $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n X_i X_i^\top$ est la matrice de covariance empirique ; $\widehat{\beta}_n^{\mathrm{LS}}$ est défini de manière unique si et seulement si $\widehat{\Sigma}_n$ est inversible.

1.3.1 Bornes de risque en régression linéaire

La régression linéaire a une longue histoire ; son étude remonte au moins à Gauss (1809) et Legendre (1805), qui ont indépendamment introduit le principe des moindres carrés. Son traitement statistique le plus courant se fait dans le cas d'un *design déterministe*, c'est-à-dire lorsque les X_i sont supposés déterministes, et lorsque l'erreur de prédiction est évaluée sur les mêmes points X_i, avec un nouveau tirage des réponses Y_i' associées. Dans ce cas, si $\mathrm{Var}(Y_i) \leqslant \sigma^2$ pour tout $i = 1, \dots, n$, on vérifie directement que l'excès de risque de l'estimateur $\widehat{\beta}_n^{\mathrm{LS}}$ est d'au plus $\sigma^2 d/n$ en espérance (Lehmann and Casella, 1998). Pour le problème de la régression linéaire avec *design aléatoire* considéré ici, certaines analyses classiques, motivées par une approche non paramétrique, fournissent des garanties de la forme suivante :

[8] À proprement parler, lorsque X n'est pas centré, Σ n'est pas la matrice de covariance de X mais plutôt son second moment, également appelé matrice de Gram. Par un léger abus de terminologie courant sur ce sujet, nous appelons tout de même Σ la "matrice de covariance" de X.

Théorème 1.6 (Györfi et al., 2002, Théorème 11.3). *Supposons qu'il existe $\sigma^2 > 0$ tel que*

$$\mathrm{Var}(Y|X) \leqslant \sigma^2 \tag{1.48}$$

presque sûrement. Supposons de plus qu'il existe une constante $R > 0$ telle que la fonction de régression $g^*(x) := \mathbb{E}[Y|X = x]$ vérifie

$$\sup_{x \in \mathbf{R}^d} |g^*(x)| \leqslant R\,. \tag{1.49}$$

Définissons l'estimateur $\widehat{g}_n^R$ par $\widehat{g}_n^R(x) = \min(-R, \max(R, \langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle))$ pour $x \in \mathbf{R}^d$, obtenu en restreignant à $[-R, R]$ les prédictions de (ou d'un choix de) l'estimateur des moindres carrés $\widehat{\beta}_n^{\mathrm{LS}}$. Alors, on a, pour une certaine constante universelle C,

$$\mathbb{E}[R(\widehat{g}_n^R)] - R(g^*) \leqslant 8\big(R(\beta^*) - R(g^*)\big) + C \max(\sigma^2, R^2) \frac{d(\log n + 1)}{n}\,. \tag{1.50}$$

La borne (1.50) est une *inégalité d'oracle inexacte*, qui majore le risque de l'estimateur (ou plus précisément, l'excès de risque par rapport au prédicteur de Bayes g^*) par un *multiple constant* du risque minimal dans la classe $\mathcal{F}$. Ce type de garantie est utile comme intermédiaire dans le cadre d'une analyse non paramétrique, où la classe $\mathcal{F} = \mathcal{F}_n$ est choisie de telle sorte que l'erreur d'approximation $R(\beta^*) - R(g^*)$ est elle-même d'ordre $O(d/n)$, où $d = d_n = \dim(\mathcal{F}_n)$ (Györfi et al., 2002). Cependant, pour une classe $\mathcal{F}$ donnée, une telle borne ne garantit pas que l'erreur de l'estimateur converge vers l'erreur minimale $R(\beta^*)$ sur $\mathcal{F}$.

Pour ce qui est des garanties sur l'excès de risque (c'est-à-dire la différence $R(\widehat{g}_n) - R(g^*)$), Tsybakov (2003) a montré que l'excès de risque minimax sous les conditions (1.48) et (1.49) est d'ordre $O(\sigma^2 d/n)$, à une constante près. La borne inférieure d'ordre $c\sigma^2 d/n$ (pour une constante universelle $c > 0$) est obtenue par un argument de réduction à un problème de test multiple, en utilisant l'inégalité de Fano (Tsybakov, 2003). Pour ce qui est de la borne supérieure, elle est obtenue en considérant un estimateur $\widehat{\beta}_n$ par projection, supposant la connaissance de la loi P_X de X, qui satisfait

$$\mathbb{E}[R(\widehat{\beta}_n)] - R(\beta^*) \leqslant \frac{(\sigma^2 + R^2)d}{n} \tag{1.51}$$

sous les conditions (1.48), (1.49), et $|X_j| \leqslant R$ presque sûrement. Cela implique que, sous ces conditions de bornes sur P_X et g^*, l'excès de risque minimax est d'ordre $O(\sigma^2 d/n)$.

Outre l'estimateur des moindres carrés $\widehat{\beta}_n^{\mathrm{LS}}$ (c'est-à-dire le minimiseur du risque empirique), une procédure classique est l'estimateur Ridge (Hoerl, 1962), donné par

$$\widehat{\beta}_{\lambda,n} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \beta, X_i\rangle)^2 + \frac{\lambda}{2}\|\beta\|^2 \right\} = (\widehat{\Sigma}_n + \lambda I_d)^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i\,. \tag{1.52}$$

Contrairement à $\widehat{\beta}_n^{\mathrm{LS}}$, cet estimateur est toujours bien défini dès lors que $\lambda > 0$, même lorsque $\widehat{\Sigma}_n$ est non inversible. En particulier, cet estimateur peut être défini dans le cas où $d > n$, ou plus généralement sur un espace de Hilbert à noyau reproduisant (en anglais *Reproducing Kernel Hilbert Space*, RKHS) de dimension infinie (Wahba, 1990; Smola and Schölkopf, 2002; Steinwart and Christmann, 2008). L'analyse du risque des estimateurs Ridge dans le contexte non paramétrique des RKHS a été menée dans une série d'article au cours des années 2000 (Cucker and Smale, 2002a,b; De Vito et al., 2005; Caponnetto and De Vito, 2007; Smale and

Zhou, 2007; Steinwart et al., 2009). Dans ce cas, la dimension d de l'espace ambiant est remplacée par des quantités indépendantes de la dimension, qui ne dépendent que de Σ et de β^* (Wahba, 1990; Caponnetto and De Vito, 2007). Enfin, les estimateurs Ridge et OLS ont été étudiés de manière fine par Audibert and Catoni (2011); Hsu et al. (2014). En particulier, Hsu et al. (2014) effectue une décomposition précise du risque de $\widehat{\beta}_n^{\mathrm{LS}}$ et de $\widehat{\beta}_{\lambda,n}$. Le Théorème 1 de Hsu et al. (2014), affirme que, si $\|\Sigma^{-1/2}X\|/\sqrt{d}$ est borné presque sûrement (par une borne indépendante de la dimension) si l'erreur d'approximation $\langle\beta^*, X\rangle - g^*(X)$ est bornée et le bruit $Y - g^*(X)$ sous-Gaussien, alors pour $n \gtrsim d\log d$, avec forte probabilité :

$$R(\widehat{\beta}_n^{\mathrm{LS}}) - R(\beta^*) = O\Big(\frac{(\sigma^2 + R^2)d}{n}\Big)\,, \tag{1.53}$$

où R contrôle l'erreur d'approximation. Par rapport à la borne (1.51), la borne (1.53) (a) est valide avec forte probabilité, plutôt qu'en espérance ; (b) porte sur l'estimateur classique des moindres carrés $\widehat{\beta}_n^{\mathrm{LS}}$, plutôt que sur l'estimateur par projection $\widehat{\beta}_n$ (qui requiert la connaissance de P_X) ; et (c) requiert une taille d'échantillon $n \gtrsim d\log d$. La dernière condition est nécessaire sans hypothèse additionnelle sur X (Vershynin, 2012).

Enfin, il est possible d'obtenir des garanties d'excès de risque en $O(d/n)$ avec forte probabilité pour $n \gtrsim d$, sous l'hypothèse que le vecteur aléatoire $\Sigma^{-1/2}X$ est *sous-Gaussien*, au sens où les marginales unidimensionnelles $\langle\Sigma^{-1/2}X, \theta\rangle$, $\theta \in S^{d-1}$, sont uniformément sous-Gaussiennes (voir la définition dans l'annexe en Section 1.6 de cette introduction), voir par exemple Vershynin (2012); Lecué and Mendelson (2013). Une telle condition affirme que la variable $\langle\Sigma^{-1/2}X, \theta\rangle^2$ est contrôlée en fonction de son espérance $\mathbb{E}\langle\Sigma^{-1/2}X, \theta\rangle^2 = 1$, dans toutes les directions θ ; cela signifie que la loi de $\Sigma^{-1/2}X$ est suffisamment "isotrope", et n'admet pas de "pics" dans certaines directions particulières. De manière remarquable, l'hypothèse sous-Gaussienne peut être affaiblie de manière significative (Oliveira, 2016; Koltchinskii and Mendelson, 2015; Mendelson, 2014). Il suffit par exemple que les moments d'ordre 4 de la forme $\mathbb{E}[\langle\Sigma^{-1/2}X, \theta\rangle^4]^{1/4}$ soient uniformément bornés pour $\theta \in S^{d-1}$ (Oliveira, 2016) ; cette condition peut elle-même être affaiblie, et remplacée par une condition sur la queue "inférieure" des variables $|\langle\Sigma^{-1/2}X, \theta\rangle|$ (Koltchinskii and Mendelson, 2015; Mendelson, 2014). Ces résultats reposent sur l'observation qu'une variable positive est naturellement "repoussée" par 0, même sans hypothèse forte sur sa queue "supérieure". Nous reviendrons sur ces questions dans la Section 1.3.3.

Les résultats précédents portent sur les estimateurs OLS $\widehat{\beta}_n^{\mathrm{LS}}$ et Ridge $\widehat{\beta}_{\lambda,n}$. Le problème des moindres carrés a également été étudié dans le cadre de l'optimisation stochastique, où l'estimateur considéré est le résultat d'un algorithme d'optimisation, généralement une variante de la descente de gradient ; voir par exemple Ying and Pontil (2008); Bach and Moulines (2013); Rosasco and Villa (2015); Dieuleveut and Bach (2016).

1.3.2 Analyse minimax exacte de la régression linéaire (Chapitre 6)

Dans le Chapitre 6, nous étudions le problème des moindres carrés, du point de vue de l'excès de risque minimax par rapport à la classe $\mathcal{F}$ des fonctions linéaires. Nous étudions en particulier l'effet de la loi P_X de X ainsi que de l'éventuelle erreur d'approximation sur la difficulté du problème. Plus précisément, fixons la loi P_X de X. Étant donnée une famille $\mathcal{P}$ de lois jointes de (X, Y) telles que $X \sim P_X$ (c'est-à-dire, de lois conditionnelles de Y sachant X), on

définit l'excès de risque minimax sur $\mathcal{P}$ par rapport à $\mathcal{F}$ comme étant

$$\mathcal{E}_n^*(\mathcal{P}) := \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}} \left(\mathbb{E}[R(\widehat{\beta}_n)] - \inf_{\beta \in \mathbf{R}^d} R(\beta) \right). \tag{1.54}$$

Cette quantité coïncide avec l'excès de risque minimax $\mathcal{E}_n^*(\ell, \mathcal{P}, \mathcal{F})$ défini par l'équation (1.6), où ℓ est la perte quadratique et $\mathcal{F}$ la classe des fonctions linéaires $\mathbf{R}^d \to \mathbf{R}$.

Nous considérons trois classes $\mathcal{P}$, dépendant chacune de la loi P_X de X ainsi que d'un paramètre $\sigma^2 > 0$ quantifiant le "niveau de bruit", c'est-à-dire la variance du problème. Pour les définir, notons $\varepsilon = Y - \langle \beta^*, X \rangle$ l'erreur du prédicteur linéaire optimal, qui satisfait $\mathbb{E}[\varepsilon^2] \leqslant \mathbb{E}[Y^2] < +\infty$ et $\mathbb{E}[\varepsilon X] = \mathbb{E}[YX] - \Sigma\beta^* = 0$. La loi jointe de (X, Y) est alors caractérisée par la loi P_X de X, le coefficient $\beta^* \in \mathbf{R}^d$ ainsi que la loi conditionnelle de ε sachant X. Pour toute loi P_X de la variable X et pour tout niveau de bruit σ^2, nous définissons les classes suivantes :

$$\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2) := \left\{ P = P_{(X,Y)} : X \sim P_X,\, \varepsilon | X \sim \mathcal{N}(0, \sigma^2) \right\}; \tag{1.55}$$

$$\mathcal{P}_{\text{well}}(P_X, \sigma^2) := \left\{ P = P_{(X,Y)} : X \sim P_X,\, \mathbb{E}[\varepsilon|X] = 0,\, \mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2 \right\}; \tag{1.56}$$

$$\mathcal{P}_{\text{mis}}(P_X, \sigma^2) := \left\{ P = P_{(X,Y)} : X \sim P_X,\, \mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2 \right\}. \tag{1.57}$$

La classe $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ correspond au cas d'un modèle linéaire bien spécifié (au sens où $\mathbb{E}[Y|X] = \langle \beta^*, X \rangle$) avec bruit indépendant Gaussien. Les classes $\mathcal{P}_{\text{mis}}(P_X, \sigma^2)$ contiennent respectivement les lois *bien spécifiées* et *mal spécifiées* (générales), sous une condition sur le second moment de l'erreur ε. En particulier, l'inclusion $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2) \subset \mathcal{P}_{\text{well}}(P_X, \sigma^2) \subset \mathcal{P}_{\text{mis}}(P_X, \sigma^2)$ implique l'inégalité $\mathcal{E}_n^*(\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)) \leqslant \mathcal{E}_n^*(\mathcal{P}_{\text{well}}(P_X, \sigma^2)) \leqslant \mathcal{E}_n^*(\mathcal{P}_{\text{mis}}(P_X, \sigma^2))$ entre les risques minimax.

Risque minimax, levier et bornes inférieures. Dans ce qui suit, nous supposons Σ inversible, ce qui n'est pas restrictif. Nous disons qu'une loi P_X est *non dégénérée* si, pour tout $n \geqslant d$, la matrice de covariance empirique $\widehat{\Sigma}_n$ est inversible presque sûrement. Par ce qui précède, cela équivaut à dire que l'estimateur des moindres carrés $\widehat{\beta}_n^{\text{LS}}$ est bien défini presque sûrement. Par la Définition 6.1 du Chapitre 6, cela équivaut à dire que, pour tout hyperplan H de $\mathbf{R}^d$, $\mathbb{P}(X \in H) = 0$, c'est-à-dire que P_X ne charge aucun hyperplan de $\mathbf{R}^d$.

Théorème 1.7 (Théorème 6.1, Chapitre 6)**.** *Si $n < d$ ou si P_X est dégénérée, alors le risque minimax sur $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ est infini. Sinon, les excès de risque minimax sur les classes $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ et $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ coïncident, et valent*

$$\mathcal{E}_n^*(\mathcal{P}_{\text{well}}(P_X, \sigma^2)) = \mathcal{E}_n^*(\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)) = \frac{\sigma^2}{n} \mathbb{E}\left[\text{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\right]. \tag{1.58}$$

De plus, cet excès de risque est atteint par l'estimateur des moindres carrés $\widehat{\beta}_n^{\text{LS}}$.

En particulier, il découle du Théorème 1.7 que la loi conditionnelle de l'erreur $\varepsilon = Y - \langle \beta^*, X \rangle$ la moins favorable (au sens de l'excès de risque minimax sur la classe de loi $\{(X, \langle \beta^*, X \rangle + \varepsilon) : \beta^* \in \mathbf{R}^d\}$) sous les contraintes $\mathbb{E}[\varepsilon|X] = 0$ et $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$ est la loi $\varepsilon|X \sim \mathcal{N}(0, \sigma^2)$. Il est de plus possible de relier l'excès de risque (1.58) à la notion

de *levier statistique*. Étant donnés $X_1, \dots, X_n, X_{n+1}$, le levier $\widehat{\ell}_{n+1}$ du point X_{n+1} parmi $X_1, \dots, X_n, X_{n+1}$ est par définition

$$\widehat{\ell}_{n+1} := \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle \tag{1.59}$$

(Hoaglin and Welsch, 1978; Chatterjee and Hadi, 1988). Le levier quantifie l'influence d'un point X_{n+1} sur la valeur de la prédiction associée : pour $y_1, \dots, y_{n+1} \in \mathbf{R}$, en notant $\widehat{\beta}_{n+1}$ l'estimateur des moindres carrés obtenu sur l'échantillon $(X_1, y_1), \dots, (X_{n+1}, y_{n+1})$ et $\widehat{y}_i := \langle \widehat{\beta}_{n+1}, X_i \rangle$ (vu comme une fonction de $y_1, \dots, y_{n+1}$), on a $\widehat{\ell}_{n+1} = \partial \widehat{y}_{n+1} / \partial y_{n+1}$. Intuitivement, un levier important est associé à une prédiction "instable" et faiblement déterminée au point X_{n+1} considéré.

Théorème 1.8 (Théorème 6.2, Chapitre 6)**.** *Supposons que $n \geqslant d$ et que la loi P_X est non dégénérée. L'excès de risque minimax sur $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ et $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ s'exprime en fonction de la loi du levier $\widehat{\ell}_{n+1}$ d'un point X_{n+1} parmi $(X_1, \dots, X_{n+1}) \sim P^{n+1}$, de la façon suivante :*

$$\mathcal{E}_n^*(\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)) = \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}} \right]. \tag{1.60}$$

Il en découle que, quelle que soit P_X, l'excès de risque minimax est d'au moins

$$\mathcal{E}_n^*(\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)) \geqslant \frac{\sigma^2 d}{n - d + 1}. \tag{1.61}$$

La première partie du Théorème 1.8 montre qu'une distribution P_X telle que la loi du levier est peu concentrée (de sorte que certains points peuvent avoir un levier élevé) conduit à un risque minimax plus élevé. Cela tient au fait que l'estimateur des moindres carrés est dans ce cas déterminé par un plus petit nombre de points (de fort levier), ce qui induit une variance plus élevée.

La borne inférieure (1.61) découle directement de l'expression (1.60) en termes du levier, et du fait que $\mathbb{E}[\widehat{\ell}_{n+1}] = d/(n+1)$. Cette borne ne dépend pas de la loi P_X ; elle est presque optimale, comme le montre l'expression du risque de $\widehat{\beta}_n^{\mathrm{LS}}$ dans le cas où $X \sim \mathcal{N}(0, \Sigma)$, qui s'obtient à partir de l'espérance de la loi de *Wishart inverse* (Anderson, 2003; Breiman and Freedman, 1983) : si $P \in \mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$,

$$\mathbb{E}[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})] = \frac{\sigma^2 d}{n - d - 1}. \tag{1.62}$$

La borne (1.61) implique en particulier une borne inférieure de $\sigma^2 d/n$; comme nous le verrons plus bas, il s'agit de la vitesse asymptotique exacte lorsque $d/n \to 0$. Cependant, la borne (1.61) est plus précise, notamment en dimension modérément élevée ; elle montre en particulier que la régression avec design aléatoire est plus difficile (du point de vue minimax) que la régression avec design déterministe. En effet, lorsque $d, n \to \infty$, avec $d/n \to \gamma \in (0, 1)$, la borne inférieure (1.61) implique que le risque minimax pour toute distribution P_X est d'au moins $\sigma^2 \gamma / (1 - \gamma)$. De plus, par (1.62), cette borne inférieure est atteinte dans le cas Gaussien où $P_X = \mathcal{N}(0, \Sigma)$; ainsi, la loi Gaussienne est la plus favorable en grande dimension, du point de vue du risque minimax.

Notons qu'il existe des résultats d'universalité, montrant que $\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}}) \to \sigma^2 \gamma / (1 - \gamma)$ presque sûrement si $d, n \to \infty$, $d/n \to \gamma$, et si les coordonnées de $\Sigma^{-1/2} X$ sont indépendantes et

satisfont certaines conditions modestes (voir, par exemple, Tulino and Verdú, 2004) ; cela découle de résultats sur la convergence du spectre de matrices aléatoires (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). Cependant, en dépit de leur universalité apparente, ces résultats reposent sur l'hypothèse très forte d'indépendance des coordonnées, qui induit en grande dimension une géométrie très particulière (El Karoui and Kösters, 2011; El Karoui, 2018). En effet, supposons que $X = (X^j)_{1 \leqslant j \leqslant d}$, où les coordonnées X^j sont indépendantes avec $\mathbb{E}[X^j] = 0$, $\mathbb{E}[(X^j)^2] = 1$ et $\mathbb{E}[(X^j)^4] \leqslant \kappa = O(1)$ pour tout $j = 1, \dots, d$. Alors, $\mathbb{E}[\|X\|^2] = \sum_{j=1}^d \mathbb{E}[(X^j)^2] = d$, tandis que $\mathrm{Var}(\|X\|^2) = \sum_{j=1}^d \mathrm{Var}(X_j^2) \leqslant \kappa d$, ce qui montre qu'avec forte probabilité, $\|X\|^2 = d + O(\sqrt{\kappa d}) = (1 + O(1/\sqrt{d}))d$, et donc $\|X\| = (1 + O(1/\sqrt{d}))\sqrt{d} = \sqrt{d} + O(1)$. Ainsi, X est proche de la sphère de rayon $\sqrt{d}$ avec forte probabilité ; ce résultat peut être précisé en termes de concentration de la mesure (Boucheron et al., 2013; Vershynin, 2018). On montre de même que, si X, X' sont deux réalisations indépendantes de cette loi, alors $\langle X, X' \rangle = O(\sqrt{d})$ en probabilité, de sorte que les vecteurs X, X' sont approximativement orthogonaux : $\langle X, X' \rangle / (\|X\| \cdot \|X'\|) = O(1/\sqrt{d})$.

La borne inférieure (1.61) se généralise au cas du risque de Bayes de l'estimateur Ridge selon une loi a priori Gaussienne sur β^*, de matrice de covariance proportionnelle à Σ^{-1} (c'est-à-dire de densité constante sur les lignes de niveau $\|\beta^*\|_\Sigma = t$, $t > 0$), comme nous le montrons en Section 8.2. Ces résultats montrent qu'au-delà de leur universalité (dans le cas particulier des variables indépendantes), les risques limites obtenus en grande dimension dans le cas de variables prédictives Gaussiennes constituent une *borne inférieure* dans le cas général.

Bornes supérieures : cas bien et mal spécifiés. Venons-en aux bornes *supérieures* sur le risque minimax, qui constituent la principale contribution technique du Chapitre 6. Les résultats suivants reposent sur l'étude quantitative de l'inversibilité de matrices de covariance empiriques $\widehat{\Sigma}_n$, présentée dans la Section 1.3.3 suivante. En particulier, ces bornes sont les premières bornes en espérance pour l'estimateur OLS avec design aléatoire non Gaussien.

Par le Théorème 1.7, pour que l'excès de risque minimax soit fini, il est nécessaire que P_X soit *non dégénérée*, au sens où $\mathbb{P}(X \in H) = 0$ pour tout hyperplan $H \subset \mathbf{R}^d$. Afin d'obtenir une borne explicite sur ce risque, nous introduisons une version quantitative de cette hypothèse :

Hypothèse 1.1 (Régularité, propriété de "petite boule")**.** Il existe des constantes $C \geqslant 1$ et $\alpha \in (0, 1]$ telles que, pour tout hyperplan $H \subset \mathbf{R}^d$ et tout $t > 0$, on a

$$\mathbb{P}(\mathrm{dist}(\Sigma^{-1/2} X, H) \leqslant t) \leqslant (Ct)^\alpha \,. \tag{1.63}$$

De manière équivalente, pour tout $\theta \in \mathbf{R}^d \setminus \{0\}$, en notant $\|\theta\|_\Sigma = \|\Sigma^{1/2}\theta\|$,

$$\mathbb{P}(|\langle \theta, X \rangle| \leqslant t \|\theta\|_\Sigma) \leqslant (Ct)^\alpha \,. \tag{1.64}$$

L'équivalence entre (1.63) et (1.64) vient du fait que, si $\theta' \in S^{d-1}$ est un vecteur unitaire normal à l'hyperplan H, alors $\mathrm{dist}(\Sigma^{-1/2} X, H) = |\langle \theta', \Sigma^{-1/2} X \rangle| = |\langle \theta, X \rangle|$, où $\theta := \Sigma^{-1/2}\theta'$ satisfait $\|\theta\|_\Sigma = \|\theta'\| = 1$ (dont (1.64) découle en normalisant θ). L'Hypothèse 1.1 est une version légèrement renforcée de la propriété de "petite boule" (en anglais *small ball*) considérée par Koltchinskii and Mendelson (2015); Mendelson (2014), qui affirme qu'il existe des constantes $c, p \in (0, 1)$ telles que, pour tout $\theta \neq 0$,

$$\mathbb{P}(|\langle \theta, X \rangle| \leqslant c \|\theta\|_\Sigma) \leqslant p \,. \tag{1.65}$$

Cette condition revient à supposer que l'Hypothèse 1.1 est valide pour *une* valeur de $t \in (0, C^{-1})$. Par une inégalité de Paley-Zygmund, et en notant que $\|\theta\|_\Sigma = \mathbb{E}[\langle\theta, X\rangle^2]^{1/2}$, la condition précédente revient à une équivalence entre les normes L^1 et L^2 des marginales de dimension 1 de X, soit $\mathbb{E}[|\langle\theta, X\rangle|] \geqslant c'\mathbb{E}[\langle\theta, X\rangle^2]^{1/2}$ pour une certaine constante $c' > 0$, uniformément sur $\theta \in \mathbf{R}^d$ (Koltchinskii and Mendelson, 2015). L'Hypothèse 1.1 renforce cette condition à tout $t > 0$ (ce qui implique en particulier que X est non dégénérée), de manière à obtenir des bornes de risque valides à tout niveau de probabilité, ce qui permet d'obtenir des bornes en espérance et donc de contrôler l'excès de risque minimax.

L'Hypothèse 1.1 est d'autant plus faible que α est faible, et C élevé. Cette condition est typiquement vérifiée pour α, C indépendants de la dimension d ; nous verrons dans la section suivante que sous certaines hypothèses, cette condition est satisfaite pour $\alpha = 1$ et $C = O(1)$. Notons que cette hypothèse n'impose aucune condition sur la queue *supérieure* des variables $\langle\theta, X\rangle^2$, au-delà de leur intégrabilité (qui permet de définir Σ et le risque).

L'Hypothèse 1.1 permet à elle seule d'obtenir une borne de $C'\sigma^2 d/n$ pour $d \gtrsim n$ sur l'excès de risque minimax sur la classes $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ et $\mathcal{P}_{\text{mis}}(P_X, \sigma^2)$, où C' est une constante ne dépendant que de C, α (Proposition 6.2 du Chapitre 6). Par la borne inférieure (1.61) de $\sigma^2 d/n$, ces bornes supérieures sont optimales à la constante C' près. Sous une modeste condition supplémentaire sur X, il est possible d'obtenir une borne avec une constante optimale dans le terme du premier ordre pour $d/n \to 0$.

Hypothèse 1.2 (Kurtosis bornée de la norme)**.** On a : $\kappa := \mathbb{E}[\|\Sigma^{-1/2}X\|^4]/d^2 < +\infty$.

Notons que $\mathbb{E}[\|\Sigma^{-1/2}X\|^2] = \mathbb{E}[\mathrm{Tr}(\Sigma^{-1/2}XX^\top\Sigma^{-1/2})] = \mathrm{Tr}(\Sigma^{-1}\Sigma) = d$, de sorte que $\kappa = \mathbb{E}[\|\Sigma^{-1/2}X\|^4]/\mathbb{E}[\|\Sigma^{-1/2}X\|^2]^2$ est la kurtosis de $\|\Sigma^{-1/2}X\|$. De plus, $\kappa \geqslant 1$ par l'inégalité de Cauchy-Schwarz. Dans de nombreuses situations, κ est bornée indépendamment de la dimension ; dans le cas où les coordonnées de X sont indépendantes, centrées, de variance 1 et de kurtosis bornée, on vérifie même que $\kappa = 1 + \mathrm{Var}(\|\Sigma^{-1/2}X\|^2)/d^2 = 1 + O(1/d)$. Sous cette hypothèse supplémentaire, l'excès de risque minimax dans le cas bien spécifié est contrôlé de la façon suivante :

Théorème 1.9 (Théorème 6.3, Chapitre 6)**.** *Sous les Hypothèses 1.1 et 1.2 sur la loi de X, le risque minimax sur $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ satisfait : pour tout $n \geqslant \min(6d/\alpha, 12\log(12/\alpha)/\alpha)$,*

$$\frac{\sigma^2 d}{n} \leqslant \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\text{well}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] = \mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\text{LS}})\big] \leqslant \frac{\sigma^2 d}{n}\Big(1 + 8C'\frac{\kappa d}{n}\Big), \tag{1.66}$$

où C' est une constante ne dépendant que de C, α.

Ainsi, sous les hypothèses précédentes, l'excès de risque minimax est d'ordre $\sigma^2 d/n + O((d/n)^2)$ lorsque $d/n \to 0$; en outre, la borne (1.66) est non asymptotique et valide pour $n \gtrsim d$. De plus, un calcul à l'ordre supérieur (non inclus ici) sous des hypothèses plus fortes (par exemple $\mathbb{E}[\|\Sigma^{-1/2}X\|^8] < +\infty$) montre que la borne supérieure (1.66) est essentiellement fine lorsque d/n, au facteur $8C'$ près[9]. À notre connaissance, le Théorème 1.9 est la première borne sur l'excès de risque de l'estimateur OLS en espérance (et donc sur l'excès de risque minimax) dans le cas d'un design P_X non Gaussien.

Le Théorème 1.9 repose sur le contrôle de la plus petite valeur propre de la matrice de covariance "blanchie" $\widetilde{\Sigma}_n = \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}$, présenté dans la section suivante. Puisque le risque

[9] Ceci suggère que l'Hypothèse 1.2 est essentiellement l'hypothèse minimale sur la queue supérieure de $XX^\top$ permettant d'obtenir un terme du second ordre en $O((d/n)^2)$.

minimax (1.58) est proportionnel à $\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma) = \mathrm{Tr}(\widetilde{\Sigma}_n^{-1})$, une première approche consisterait à majorer $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) \leqslant d \cdot \lambda_{\min}(\widetilde{\Sigma}_n)^{-1}$, puis à contrôler $\mathbb{E}[\lambda_{\min}(\widetilde{\Sigma}_n)^{-1}]$. Cette approche, bien que valide, peut être améliorée dans ce cas, pour deux raisons :

- sous des hypothèses convenables, on a $\mathbb{E}[\lambda_{\min}(\widetilde{\Sigma}_n)^{-1}] = 1 + \Theta(\sqrt{d/n})$, cette approche mène donc à un terme du second ordre en $O((d/n)^{3/2})$ au lieu de $O((d/n)^2)$;
- afin d'obtenir la borne précédente pour $n \gtrsim d$, des hypothèses plus fortes sur X que l'Hypothèse 1.2 sont requises (Oliveira, 2016; Koltchinskii and Mendelson, 2015).

Une approche plus fine consiste à "linéariser" la fonction $A \mapsto \mathrm{Tr}(A^{-1})$ autour de $A = I_d$, de sorte que le terme du premier ordre s'annule en espérance. D'un point de vue technique, l'Hypothèse 1.1 permet de contrôler la queue inférieure de $\lambda_{\min}(\widetilde{\Sigma}_n)$, mais pas la queue supérieure de la plus grande valeur propre $\lambda_{\max}(\widetilde{\Sigma}_n)$. Nous utilisons donc l'inégalité suivante (Lemme 6.6 du Chapitre 6) : pour toute matrice symétrique positive A de dimension d et tout $p \in [1, 2]$,

$$\mathrm{Tr}(A^{-1}) + \mathrm{Tr}(A) - 2d \leqslant \max(1, \lambda_{\min}(A)^{-1}) \cdot \mathrm{Tr}\big(|A - I_d|^{2/p}\big)\,. \tag{1.67}$$

Cette inégalité est alors appliquée à $A = \widetilde{\Sigma}_n$, prise en espérance (en utilisant $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n)] = d$) et combinée avec l'inégalité de Hölder. On montre alors (par les résultats de la section suivante) que, sous l'Hypothèse 1.1, $\mathbb{E}[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-q})]^{1/q} = O(1)$ même pour des valeurs de $q = p/(p-1)$ élevées (jusqu'à $q \asymp n$). Par ailleurs, Hypothèse 1.2 implique que $(1/d)\mathbb{E}[\mathrm{Tr}((\widetilde{\Sigma}_n - I_d)^2)] = O(d/n)$; la borne (1.66) s'obtient alors en optimisant le choix de p.

Le Théorème 1.9 porte sur le cas d'un modèle linéaire *bien spécifié*, où $\mathbb{E}[Y|X]$ est linéaire en X. Considérons maintenant le cas général mal spécifié. Nous introduisons une variante renforcée de l'Hypothèse 1.2 :

Hypothèse 1.3 (Équivalence L^4-L^2 des marginales unidimensionnelles)**.** Il existe une constante $\kappa > 0$ telle que, pour tout $\theta \in \mathbf{R}^d$, $\mathbb{E}[\langle\theta, X\rangle^4] \leqslant \kappa\mathbb{E}[\langle\theta, X\rangle^2]^2 = \kappa\|\theta\|_\Sigma^4$.

L'Hypothèse 1.3 implique l'Hypothèse 1.2, avec $\mathbb{E}[\|\Sigma^{-1/2}X\|^4]/d^2 \leqslant \kappa$. L'Hypothèse 1.3 est cependant plus forte, car elle contrôle $\Sigma^{-1/2}X$ uniformément dans toutes les directions. Cette condition (avec $\kappa = O(1)$) est cependant courante en régression, et peu restrictive dès lors que la variable normalisée $\Sigma^{-1/2}X$ est suffisamment isotrope (à l'instar de l'Hypothèse 1.1, mais pour la queue supérieure au lieu de la queue inférieure). Cette condition est en particulier nettement moins restrictive que l'hypothèse que $\Sigma^{-1/2}X$ est sous-Gaussien, elle-même courante en analyse de la régression linéaire. Elle n'est cependant pas universellement satisfaite : en effet, pour certaines bases de fonctions "localisées" tels des histogrammes, la constante κ peut dépendre de la dimension (Saumard, 2018).

Sous les Hypothèses 1.1 et 1.3, il est possible de contrôler le risque minimax dans le cas mal spécifié.

Proposition 1.6 (Proposition 6.3, Chapitre 6)**.** *Supposons que P_X satisfait les Hypothèses 1.1 et 1.3. Supposons également que $\mathbb{E}[(Y - \langle\beta^*, X\rangle)^4\|\Sigma^{-1/2}X\|^4] \leqslant \chi d^2$. Si $n \geqslant \max(6d, 96)/\alpha$, on a*

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{1}{n}\mathbb{E}\big[(Y - \langle\beta^*, X\rangle)^2\|\Sigma^{-1/2}X\|^2\big] + 276C'^2\sqrt{\kappa\chi}\Big(\frac{d}{n}\Big)^{3/2}. \tag{1.68}$$

En outre,

$$\frac{\sigma^2 d}{n} \leqslant \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n)\big] \leqslant \sup_{P \in \mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{\sigma^2 d}{n}\left(1 + 276C'^2\kappa\sqrt{\frac{d}{n}}\right). \tag{1.69}$$

La Proposition 1.6 combine la borne sur $\lambda_{\min}(\widetilde{\Sigma}_n)$ que nous obtenons sous l'Hypothèse 1.1 avec une borne d'Oliveira (2016) sous l'Hypothèse 1.3 (voir la Section 1.3.3 suivante).

Notons que les inégalités (1.69) impliquent que l'estimateur $\widehat{\beta}_n^{\mathrm{LS}}$ est asymptotiquement minimax sur la classe $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$ lorsque $d/n \to 0$ (en plus d'être exactement minimax sur les classes $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ et $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$). De plus, le risque minimax sur la classe $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$ est équivalent à $\sigma^2 d/n$ lorsque $d/n \to 0$, ce qui montre que le bruit indépendant Gaussien (de variance σ^2) est asymptotiquement la loi conditionnelle de l'erreur $\varepsilon = Y - \langle \beta^*, X \rangle$ la moins favorable sous la contrainte $\mathbb{E}[\varepsilon^2 | X] \leqslant \sigma^2$.

1.3.3 Contrôle de la plus petite valeur propre de matrices de covariance empiriques (Chapitre 6)

Les bornes supérieures de la section précédente reposent sur un contrôle de la queue inférieure de matrices de covariance aléatoires. En effet, par le Théorème 1.9, l'excès de risque minimax dans le cas bien spécifié est $\sigma^2 \mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)]/n$; afin de montrer que cette quantité est finie (et d'ordre $\sigma^2 d/n$ lorsque $d/n \to 0$), il est nécessaire (puisque $\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma) \geqslant \lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2})^{-1}$) de contrôler des moments du type $\mathbb{E}[\lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2})^{-q}]^{1/q}$ pour $q \geqslant 1$ (dans l'esquisse de la preuve du Théorème 1.9, nous avons utilisé le fait que

$$\mathbb{E}[\lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2})^{-q}]^{1/q} = O(1)$$

pour $q \lesssim n$). Nous sommes donc naturellement conduits à majorer la queue inférieure de la variable aléatoire $\lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2})$, c'est-à-dire des probabilités de la forme

$$\mathbb{P}(\lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}) \leqslant t)$$

pour $t \in (0, 1)$. Dans cette section, quitte à remplacer X par $\Sigma^{-1/2}X$, nous supposons dorénavant que $\Sigma = \mathbb{E}[XX^\top] = I_d$.

Bornes existantes sur $\lambda_{\min}(\widehat{\Sigma}_n)$. La littérature sur l'étude de $\lambda_{\min}(\widehat{\Sigma}_n)$, ainsi que de la plus grande valeur propre $\lambda_{\max}(\widehat{\Sigma}_n)$ et d'autres propriétés spectrales de la matrice de covariance aléatoire $\widehat{\Sigma}_n$, est très riche ; nous renvoyons aux ouvrages Bai and Silverstein (2010); Anderson et al. (2010) pour un traitement de la théorie asymptotique des grandes matrices aléatoires, et aux références Vershynin (2012); Tao (2012); Tropp (2015) sur leur étude non asymptotique. Dans ce qui suit, nous nous restreignons aux propriétés de $\lambda_{\min}(\widehat{\Sigma}_n)$; pour davantage de détails (et de références) sur ce sujet, nous renvoyons au compte-rendu de Rudelson and Vershynin (2010) sur les résultats antérieurs à 2010.

Tout d'abord, pour d fixe et $n \to \infty$, la loi (forte) des grands nombres implique que $\widehat{\Sigma}_n \to I_d$ et donc $\lambda_{\min}(\widehat{\Sigma}_n) \to 1$ presque sûrement. De plus, si $\mathbb{E}[\|X\|^4] < +\infty$, le théorème central limite implique que $\sqrt{n}(\widehat{\Sigma}_n - I_d)$ converge vers une loi Gaussienne centrée sur les matrices symétriques $d \times d$, dont la matrice de covariance est déterminée par les moments d'ordre 4 de X (c'est-à-dire par les moments de la forme $\mathbb{E}[X^{j_1}X^{j_2}X^{j_3}X^{j_4}]$, où $X = (X^j)_{1 \leqslant j \leqslant d}$ et $1 \leqslant j_1, j_2, j_3, j_4 \leqslant d$), ce dont on peut déduire la loi asymptotique de $\lambda_{\min}(\widehat{\Sigma}_n)$ et en particulier que $\lambda_{\min}(\widehat{\Sigma}_n) = 1 + O(1/\sqrt{n})$. Ces résultats ne sont cependant pas directement utiles dans notre cas, en raison de leur caractère asymptotique (d est fixé et $n \to \infty$) ; en effet, ils ne disent rien sur la taille n de l'échantillon à partir de laquelle

$$\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant \delta \tag{1.70}$$

pour $d \geqslant 1$ et $t, \delta \in (0,1)$ donnés.

Les résultats classiques sur les grandes matrices aléatoires se placent typiquement dans un régime asymptotique de grande dimension, où $d, n \to \infty$ avec $d/n \to \gamma \in (0, \infty)$ (Wigner, 1958; Marchenko and Pastur, 1967; Bai and Silverstein, 2010). Supposons que $\gamma = \lim d/n < 1$ (si $n < d$, alors $\lambda_{\min}(\widehat{\Sigma}) = 0$). Dans ce cas, si les coordonnées de X sont indépendantes avec un moment d'ordre 4 borné, on a

$$\lambda_{\min}(\widehat{\Sigma}_n) \to (1 - \sqrt{\gamma})^2$$

presque sûrement (Bai and Yin, 1993). Tout comme le précédent, ce résultat est asymptotique (avec ici $n, d \to \infty$) et ne donne pas de borne de type (1.70). Cependant, ce résultat limite suggère que, sous ces hypothèses, $\lambda_{\min}(\widehat{\Sigma}_n) \geqslant 1 - O(\sqrt{d/n})$ pour $n \gtrsim d$, et en particulier $\lambda_{\min}(\widehat{\Sigma}_n) \geqslant c > 0$ avec forte probabilité.

Pour ce qui est des bornes non asymptotiques, valables pour n, d fixés, une inégalité de déviation exponentielle, de la forme

$$\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t_0) \leqslant c_0^n \tag{1.71}$$

pour tout $n \geqslant C_0 d$, où $t_0, c_0 \in (0,1)$ et $C_0 > 1$ sont des constantes indépendantes de n, d, a été obtenue par Bennett et al. (1977) dans le cas où les coordonnées de X sont i.i.d. de loi $\mathbb{P}(X^j = 1) = 1 - \mathbb{P}(X^j = -1) = 1/2$. Rudelson and Vershynin (2008, 2009) ont généralisé ce résultat au cas où les coordonnées de X sont indépendantes et K-sous-Gaussiennes, et pour $n \geqslant d$ arbitraire : pour tout $t \in (0,1)$,

$$\mathbb{P}\Big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t \cdot \frac{(\sqrt{n} - \sqrt{d-1})^2}{n}\Big) \leqslant (Ct)^{(n-d+1)/2} + c^n\,, \tag{1.72}$$

où $C > 1$, $c \in (0,1)$ sont des constantes dépendant de K. Ce résultat couvre notamment le cas délicat $n = d$ (pour lequel $\lambda_{\min}(\widehat{\Sigma}_n) \asymp 1/n^2$). Dans ce qui suit, nous nous plaçons dans le cas $n \geqslant C_0 d$, qui est le régime d'intérêt d'un point de vue statistique. Dans ce cas, la borne (1.72) s'écrit : pour tout $t \in (0,1)$,

$$\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (C't)^{c'n} + c^n\,,$$

ce qui équivaut essentiellement à (1.71) pour certains $t_0, c_0 \in (0,1)$.

La borne (1.71) assure que $\lambda_{\min}(\widehat{\Sigma}_n) \geqslant t_0$ avec une probabilité exponentielle, dans le cas de variables sous-Gaussiennes. Sous cette hypothèse, il est en fait possible de préciser ce résultat : si X est K-sous-Gaussien, il existe des constantes C, c dépendant de K telles que

$$\mathbb{P}\left(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant 1 - C\sqrt{\frac{d}{n}} - t\right) \leqslant e^{-cnt^2} \tag{1.73}$$

(Vershynin, 2012, Théorème 5.39). La borne (1.73) (dont découle la borne précédente (1.71)) est une inégalité de déviation sous-Gaussienne, qui montre de manière non asymptotique que $\lambda_{\min}(\widehat{\Sigma}_n) \geqslant 1 - O(d/n)$ pour $n \gtrsim d$. Une borne analogue sur la plus grande valeur propre est valide sous la même hypothèse.

Les bornes précédentes reposent sur l'hypothèse forte que X est sous-Gaussien (plus précisément, K-sous-Gaussien avec $K = O(1)$). Sans cette hypothèse, une inégalité de Rudelson (1999) montre que $n = O(d \log d)$ échantillons suffisent pour que $\lambda_{\min}(\widehat{\Sigma}_n) \simeq 1$ si $\|X\| \leqslant R$

presque sûrement, avec $R = O(\sqrt{d})$. Le taille $O(d \log d)$ de l'échantillon est également nécessaire sans hypothèse supplémentaire sur X (Vershynin, 2012). Une série de travaux (voir, par exemple, Adamczak et al., 2010, 2011; Mendelson and Paouris, 2014; Srivastava and Vershynin, 2013 et les références qui s'y trouvent) étudie des conditions plus faibles que le caractère sous-Gaussien permettant d'obtenir des bornes d'ordre $1 + O(\sqrt{d/n})$ pour la plus petite et la plus grande valeurs propres $\lambda_{\min}(\widehat{\Sigma}_n)$ et $\lambda_{\max}(\widehat{\Sigma}_n)$.

Une observation cruciale (Oliveira, 2016; Koltchinskii and Mendelson, 2015) est qu'il est possible d'obtenir une inégalité de déviation exponentielle de la forme (1.73) sur $\lambda_{\min}(\widehat{\Sigma}_n)$ sous des hypothèses bien plus faibles sur X que celles requises pour $\lambda_{\max}(\widehat{\Sigma}_n)$. Cela tient essentiellement au fait qu'une somme de variables positives indépendantes ne peut pas être trop proche de 0. Ainsi, Oliveira (2016) montre qu'une inégalité sous-Gaussienne de type (1.73) a lieu sous l'Hypothèse 1.3 d'équivalence L^4-L^2 pour les marginales unidimensionnelles, tandis que Koltchinskii and Mendelson (2015) l'établissent en supposant une équivalence L^p-L^2 pour un exposant $p > 4$. Koltchinskii and Mendelson (2015) (voir aussi Yaskov, 2014, 2015) montrent également que des bornes légèrement plus faibles sont possibles sous des hypothèses moindres. En particulier, il est possible d'obtenir la borne exponentielle (1.71) sous la condition (1.65) de petite boule (c'est-à-dire une équivalence L^1-L^2 des marginales unidimensionnelles), qui n'impose pas l'existence de moments de la forme $\mathbb{E}[\langle \theta, X\rangle^p]$ pour $\theta \in \mathbf{R}^d$ et $p > 2$.

Contrôle des moments négatifs de $\lambda_{\min}(\widehat{\Sigma}_n)$ (Chapitre 6). Les bornes (1.71) et (1.73) contrôlent les probabilités de déviations $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$ de manière essentiellement optimale respectivement pour $c < t < 1-c$ et $c < t$, où c est une constante. Cependant, lorsque $t \to 0$ ces bornes "saturent", c'est-à-dire convergent vers une limite positive de la forme c_0^n avec $c_0 \in (0,1)$. En particulier, la borne sous-Gaussienne (1.73) ne garantit pas que $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \to 0$ lorsque $t \to 0$, et ne permet donc pas de contrôler des moments de la forme $\mathbb{E}[\lambda_{\min}(\widehat{\Sigma}_n)^{-q}]$ pour $q > 0$, qui apparaissent naturellement (notamment pour $q = 1$) dans l'analyse du risque minimax de la régression linéaire (voir la Section 1.3.2 précédente).

Notons que cette saturation est inévitable si l'on suppose seulement que X est sous-Gaussien. En effet, si $X = (X^j)_{1 \leqslant j \leqslant d}$ a des coordonnées i.i.d. de loi $\mathbb{P}(X^j = 1) = 1 - \mathbb{P}(X^j = -1) = 1/2$ (cette loi est sous-Gaussienne, voir par exemple Vershynin, 2012), alors les deux premières colonnes de la matrice $\boldsymbol{X} = (X_i^j)_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant d}$ sont égales avec probabilité $1/2^n$, auquel cas $\boldsymbol{X}$ et donc $\widehat{\Sigma}_n = \boldsymbol{X}^\top \boldsymbol{X}/n$ est de rang au plus $d-1$ et donc non inversible ; ainsi, $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) = 0) \geqslant (1/2)^n$. Cela vaut plus généralement pour toutes les lois P_X *dégénérées*, au sens de la Définition 6.1. Une hypothèse sur la loi X est donc nécessaire.

Nous effectuons dans la Section 6.3 du Chapitre 6 une analyse précise des déviations de la forme $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$ pour tout $t \in (0, c)$, et en particulier pour $t \to 0$. Cela revient à considérer des bornes de la forme (1.70), qui restent non triviales (avec $t = t_\delta > 0$) pour tout niveau de confiance $\delta \in (0,1)$ (et en particulier $\delta \to 1$). Le Proposition 1.7 suivante donne une borne inférieure sur la probabilité $\delta = \delta(t)$ de déviation dans l'inégalité (1.70) en dimension $d \geqslant 2$, valide *pour toute* loi P_X telle que $\mathbb{E}[XX^\top] = I_d$.

Proposition 1.7 (Proposition 6.4, Chapitre 6)**.** *Supposons que $d \geqslant 2$ et que $\mathbb{E}[XX^\top] = I_d$. Alors, pour tout $t \in (0,1)$, il existe $\theta \in S^{d-1}$ tel que*

$$\mathbb{P}(|\langle \theta, X\rangle| \leqslant t) \geqslant 0.16 \cdot t\,.$$

Il en découle que, pour tout $t \in (0,1)$,

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\big) \geqslant (0.025 \cdot t)^{n/2}\,. \tag{1.74}$$

Notons que l'hypothèse $d \geqslant 2$ est nécessaire dans la Proposition 1.7, comme le montre l'exemple de la variable $X \equiv 1$ en dimension 1, pour laquelle $\lambda_{\min}(\widehat{\Sigma}_n) = 1$ presque sûrement. La première inégalité de la Proposition 1.7 s'obtient par un argument probabiliste, en considérant un vecteur θ tiré uniformément sur la sphère S^{d-1}. L'inégalité (1.74) indique que la meilleure borne possible sur la probabilité $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$ est d'au moins $(Ct)^{\alpha n/2}$, pour certaines constantes $C > 0$, $\alpha \in (0, 1]$. Ce résultat montre en particulier que la borne exponentielle (1.71) est optimale pour $t_0 \in (c, 1-c)$ pour toute loi P_X (et pas seulement dans le cas de coordonnées de Bernoulli, pour lesquelles $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) = 0) \geqslant (1/2)^n$), en quantifiant de plus c_0 en fonction de t_0.

De plus, une loi P_X telle que $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (C^2 t)^{\alpha n/2}$ pour certaines constantes C, α et $t \in (0, 1)$ satisfait nécessairement

$$\sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X \rangle| \leqslant t) \leqslant (Ct)^{\alpha} . \tag{1.75}$$

La condition (1.75) coïncide précisément avec la condition de régularité (Hypothèse 1.1) évoquée dans la Section 1.3.1. Comme le montre le Théorème 1.10 suivant, cette condition nécessaire est également suffisante pour obtenir une borne optimale de type $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (Ct)^{cn}$ pour $n \geqslant Cd$, pour certaines constantes $C, c > 0$.

Théorème 1.10 (Théorème 6.4, Chapitre 6)**.** *Supposons que $\mathbb{E}[XX^\top] = I_d$ et que X satisfait l'Hypothèse 1.1. Alors, si $n \geqslant 6d/\alpha$, on a pour tout $t \in (0, 1)$:*

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\big) \leqslant (C' t)^{\alpha n/6} \tag{1.76}$$

où $C' = 3C^4 e^{1+9/\alpha}$ ne dépend que de C et α.

La borne (1.76) est une inégalité de déviation exponentielle non asymptotique, de la forme $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant \exp(-n\psi(t))$, où $\psi(t) = -(6/\alpha)\log(C't)$ est positive pour $t < C'^{-1}$ et $\psi(t) \asymp \log(1/t)$ pour $t \to 0$; cet ordre de grandeur est optimal, par la borne inférieure (1.74), pour $t \in (0, c)$. Cette borne complète la borne sous-Gaussienne (1.73), qui donne $\psi(t) \asymp (1-t)^2$ pour $1 - t \gtrsim C\sqrt{d/n}$, ce qui est optimal pour $t \in (c, 1 - C\sqrt{d/n})$.

Idée de la preuve du Théorème 1.10. La preuve du Théorème 1.10 utilise la technique dite "PAC-Bayésienne" pour contrôler les processus empiriques. Cette technique a été introduite par McAllester (1999b,a) afin d'analyser le risque de prédicteurs randomisés à poids exponentiels en apprentissage statistique ; nous renvoyons à Catoni (2007); Seeger (2002); Audibert (2004); Zhang (2006b); Langford and Shawe-Taylor (2003); McAllester (2003); Germain et al. (2009); Audibert and Catoni (2011); Catoni (2012); Grünwald and Mehta (2019) (et aux références qui s'y trouvent) pour davantage d'informations sur l'application de cette technique à l'apprentissage statistique. Cette technique a été utilisée par Oliveira (2016) afin d'établir la borne sous-Gaussienne (1.73) sous une équivalence L^4-L^2 (Hypothèse 1.3) ; notre analyse utilise une démarche similaire, mais avec des raffinements permettant d'obtenir une borne non triviale (strictement positive) pour $t \to 0$.

Le point de départ de la preuve est la représentation suivante de $\lambda_{\min}(\widehat{\Sigma}_n)$ comme minimum d'un processus empirique :

$$\lambda_{\min}(\widehat{\Sigma}_n) = \inf_{\theta \in S^{d-1}} \langle \widehat{\Sigma}_n \theta, \theta \rangle = \inf_{\theta \in S^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \langle \theta, X_i \rangle^2 \right\} . \tag{1.77}$$

Pour $\theta \in S^{d-1}$ fixé, les variables $Z_i(\theta) := \langle \theta, X_i \rangle^2$, $1 \leqslant i \leqslant n$, sont positives et i.i.d., avec $\mathbb{E}[Z_i(\theta)] = 1$. De plus, l'Hypothèse 1.1 implique que, pour tout $t \in (0,1)$, $\mathbb{P}(Z_i(\theta) \leqslant t) \leqslant (C\sqrt{t})^\alpha = (C^2 t)^{\alpha/2}$. Cette propriété peut s'exprimer en fonction de la transformée de Laplace : $\mathbb{E}[\exp(-\lambda Z_i(\theta))] \leqslant (C^2/\lambda)^{\alpha/2}$, ce qui permet de traiter naturellement la moyenne $Z(\theta) := n^{-1}\sum_{i=1}^n Z_i(\theta)$ de variables indépendantes par la méthode de Chernoff (Boucheron et al., 2013), en utilisant que $\mathbb{E}[\exp(-\lambda n Z(\theta))] = \prod_{i=1}^n \mathbb{E}[\exp(-\lambda Z_i(\theta))]$. On obtient de cette manière, pour tout $\theta \in S^{d-1}$ fixe et tout $t > 0$,

$$\mathbb{P}(Z(\theta) \leqslant t) \leqslant (C_1 t)^{\alpha n/2}$$

pour une certaine constante C_1 (dépendant de C, α).

Une idée naturelle (qui permet dans le cas où X est sous-Gaussien d'obtenir une borne sous-Gaussienne pour $\lambda_{\min}(\widehat{\Sigma}_n)$) serait alors de procéder par un argument d'approximation. Cette approche consiste à considérer un ε-recouvrement $\{\theta_1, \dots, \theta_N\}$ de $\Theta := S^{d-1}$ tel que $\min_{1\leqslant k\leqslant N} \|\theta - \theta_k\| \leqslant \varepsilon$, de cardinal minimal $N \lesssim (2 + 1/\varepsilon)^d$. La valeur de $Z(\theta_k)$ est contrôlée uniformément sur $k = 1, \dots, N$ en utilisant une borne d'union par rapport à la borne précédente, et en contrôlant l'erreur d'approximation $|Z(\theta) - Z(\theta_k)| \lesssim \lambda_{\max}(\widehat{\Sigma}_n)\|\theta - \theta_k\| \leqslant \lambda_{\max}(\widehat{\Sigma}_n) \cdot \varepsilon$ (où k est choisi tel que $\|\theta - \theta_k\| \leqslant \varepsilon$). Le problème de cette approche est qu'elle nécessite une borne de déviation exponentielle sur $\lambda_{\max}(\widehat{\Sigma}_n)$, ce qui n'est pas possible sans hypothèse forte sur la queue *supérieure* de X (par exemple, que X est sous-Gaussien, Vershynin, 2012). Notons que cette difficulté est intrinsèque à l'approximation par recouvrement, car quel que soit le choix du recouvrement, il existe des $\theta \in S^{d-1}$ tels que l'inégalité $|Z(\theta) - Z(\theta_k)| \lesssim \lambda_{\max}(\widehat{\Sigma}_n) \cdot \|\theta - \theta_k\|$ est approximativement une égalité.

Afin de contourner cette difficulté, l'idée consiste à effectuer l'approximation du processus empirique $(Z(\theta), \theta \in \Theta)$ non pas par une discrétisation fixe de Θ, mais par une version "lissée" ou "moyennée" de ce processus. Plus précisément, pour tout mesure de probabilité ρ sur $\Theta = S^{d-1}$, on considère la version moyennée de Z selon ρ :

$$\int_\Theta Z(\theta)\rho(\mathrm{d}\theta) = \int_\Theta \left(\frac{1}{n}\sum_{i=1}^n \langle \theta, X_i\rangle^2\right)\rho(\mathrm{d}\theta)\,.$$

Cette quantité peut être contrôlée de manière *uniforme* sur les lois ρ telles que $\mathrm{KL}(\rho, \pi) \leqslant K$ (où π est une *loi a priori* fixe sur Θ, ne dépendant pas des observations $X_1, \dots, X_n$), au moyen de l'inégalité PAC-Bayésienne suivante (voir, par exemple, Audibert and Catoni, 2011) : pour tous $u \geqslant 0$ et $\lambda > 0$,

$$\mathbb{P}\left(\forall \rho, \int_\Theta (-\lambda n Z(\theta))\rho(\mathrm{d}\theta) \leqslant \int_\Theta \log \mathbb{E}\left[e^{-\lambda n Z(\theta)}\right]\rho(\mathrm{d}\theta) + \mathrm{KL}(\rho, \pi) + u\right) \geqslant 1 - e^{-u}\,. \quad (1.78)$$

Ce résultat est une conséquence de la formule variationnelle de Donsker-Varadhan (Théorème 1.18). Il s'agit alors de considérer une famille de postérieurs $(\rho_\theta)_{\theta\in\Theta}$ indexée par les paramètres $\theta \in \Theta$ (ρ_θ correspond à une loi de "lissage" autour de θ). La borne PAC-Bayésienne (1.78) étant uniforme sur ρ, donc en particulier sur ρ_θ, il suffit alors de contrôler deux termes :

- un terme d'*approximation* : $Z(\theta) - \int_\Theta Z(\theta')\rho_\theta(\mathrm{d}\theta')$ pour $\theta \in \Theta$;
- un terme d'*entropie* : $\mathrm{KL}(\rho_\theta, \pi)$ pour $\theta \in \Theta$.

Oliveira (2016), suivant Audibert and Catoni (2011), utilise un lissage Gaussien : $\Theta = \mathbf{R}^d$, $\pi = \mathcal{N}(0, (\varepsilon^2/d)I_d)$ et $\rho_\theta = \mathcal{N}(\theta, (\varepsilon^2/d)I_d)$ pour tout $\theta \in \Theta$. L'intérêt du choix de lois à priori et a

posteriori Gaussiennes est qu'il permet de calculer explicitement les termes d'approximation et d'entropie :

- en utilisant le fait que $Z(\theta) = \langle \widehat{\Sigma}_n \theta, \theta \rangle$ est une forme quadratique en θ, on vérifie que le terme d'approximation est égal à $(\varepsilon^2/d) \cdot \mathrm{Tr}(\widehat{\Sigma}_n)$;
- en outre, on a $\mathrm{KL}(\rho_\theta, \pi) = \|\theta\|^2/(\varepsilon^2/d) = d/\varepsilon^2$ pour $\theta \in S^{d-1}$.

De manière cruciale, le terme d'approximation est contrôlé en fonction de $(1/d)\mathrm{Tr}(\widehat{\Sigma}_n)$ (c'est-à-dire de la moyenne des valeurs propres de $\widehat{\Sigma}_n$) et non de $\lambda_{\max}(\widehat{\Sigma}_n)$. L'avantage est qu'il est possible de s'arranger pour que $\mathrm{Tr}(\widehat{\Sigma}_n)/d = n^{-1}\sum_{i=1}^n \|X_i\|^2/d$ soit contrôlé (comme nous le verrons par la suite), tandis que ce n'est pas possible pour $\langle \theta, X_i \rangle^2$ simultanément sur toutes les directions $\theta \in S^{d-1}$. Ainsi, l'avantage clé de l'approche PAC-Bayésienne par rapport à celle par discrétisation est qu'elle permet d'obtenir une approximation "isotrope" (moyennée sur toutes les directions), par le choix de la loi ρ_θ isotrope et centrée en θ.

Cependant, dans notre cas, le lissage Gaussien ne permet pas d'obtenir la borne fine pour tout $t > 0$ du Théorème 1.10. Cela tient au fait qu'il est ici nécessaire de choisir $\varepsilon \to 0$ arbitrairement petit lorsque $t \to 0$, afin d'obtenir un terme d'approximation suffisamment faible pour que la borne sur $\lambda_{\min}(\widehat{\Sigma}_n)$ demeure positive lorsque le niveau de probabilité tend vers 1 ; de plus, pour $\varepsilon \to 0$, le terme d'entropie en d/ε^2 diverge trop rapidement.

Intuitivement, il est possible de faire mieux : en effet, la forme quadratique $\widehat{\Sigma}_n$ (et en particulier sa plus petite valeur propre (1.77)) est déterminée par ses valeurs sur S^{d-1}, de sorte que le choix de lois Gaussiennes (sur $\Theta = \mathbf{R}^d$) conduit à des redondances. Notre preuve diverge ici[10] de celle d'Oliveira (2016). Nous considérons $\Theta = S^{d-1}$, π la loi uniforme sur la sphère $\Theta = S^{d-1}$, et pour tout $\theta \in \Theta$, ρ_θ est la mesure uniforme sur la boule

$$\Theta(\theta, \varepsilon) := \Big\{ \theta' \in S^{d-1} : \|\theta' - \theta\| \leqslant \varepsilon \Big\}.$$

Dans ce cas, nous ne disposons plus de formules exactes pour les termes d'approximation et d'entropie. Cependant, il est possible d'obtenir des résultats presque équivalents, et suffisamment précis :

- par des arguments de symétrie, on montre que

$$\int_{S^{d-1}} Z(\theta')\rho_\theta(\mathrm{d}\theta') = \big(1 - \phi(\varepsilon)\big) Z(\theta) + \phi(\varepsilon) \cdot \frac{1}{d}\mathrm{Tr}(\widehat{\Sigma}_n)$$

  pour tout $\theta \in S^{d-1}$, où pour tout $\varepsilon \in (0, 1/2]$

$$\phi(\varepsilon) = \frac{d}{d-1} \int_{S^{d-1}} \big(1 - \langle \theta, v \rangle^2\big) \rho_\theta(\mathrm{d}\theta') \in \Big[0, \frac{d}{d-1}\varepsilon^2\Big] \subset [0, 1/2]\,;$$

- par des arguments de recouvrement et de symétrie, on obtient la majoration $\mathrm{KL}(\rho_\theta, \pi) = \log[1/\pi(\Theta(\theta, \varepsilon))] \leqslant d \log(2 + 1/\varepsilon)$.

[10] Une autre différence technique tient aux bornes sur la transformée de Laplace, issues d'hypothèses différentes adaptées à différentes parties de la queue de $\lambda_{\min}(\widehat{\Sigma}_n)$: Oliveira (2016) utilise une borne de type Bernstein, qui exploite les moments d'ordre 4 des marginales.

Grâce au choix de π, ρ_θ, nous obtenons un terme d'entropie (optimal) en $d\log(1/\varepsilon)$, avec une meilleure dépendance en ε lorsque $\varepsilon \to 0$ que le terme d'entropie Gaussien en d/ε^2. De plus, le terme d'approximation est contrôlé en fonction de $\phi(\varepsilon)\mathrm{Tr}(\widehat{\Sigma}_n)/d \lesssim \varepsilon^2 \cdot \mathrm{Tr}(\widehat{\Sigma}_n)/d$ plutôt que $\lambda_{\max}(\widehat{\Sigma}_n)$. Cette trace peut être contrôlée en remplaçant X par une version "tronquée", ici $X' = \min(\sqrt{d}/\|X\|, 1)X$: en effet, la plus petite valeur propre associée $\lambda_{\min}(\widehat{\Sigma}'_n)$ est inférieure à $\lambda_{\min}(\widehat{\Sigma}_n)$ (donc $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant \mathbb{P}(\lambda_{\min}(\widehat{\Sigma}'_n) \leqslant t)$ pour tout t) ; de plus, $\mathrm{Tr}(\widehat{\Sigma}'_n) \leqslant d$ presque sûrement, et l'on montre que X' satisfait également une condition de "petite boule" similaire à celle de l'Hypothèse 1.1.

La preuve se conclut en combinant les étapes précédentes et en optimisant les paramètres λ, ε en fonction de t. □

La borne du Théorème 1.10 sur la queue inférieure de $\widehat{\Sigma}_n$ peut se formuler de manière équivalente (aux constantes numériques près) en termes des moments de $\lambda_{\min}(\widehat{\Sigma}_n)^{-1}$: si $n \geqslant 12/\alpha$, alors pour tout $q \leqslant \alpha n/12$,

$$\mathbb{E}[\lambda_{\min}(\widehat{\Sigma}_n)^{-q}] \leqslant 2C'^q\,.$$

Enfin, dans la Section 6.3.3, nous discutons l'Hypothèse 1.1 dans le cas de variables indépendantes ; cette condition est alors naturellement satisfaite sous des hypothèses de régularité sur les coordonnées de X. Cela peut être établi soit par un résultat de Rudelson and Vershynin (2014) dans le cas de variables X^j indépendantes de densité bornée, soit par un contrôle uniforme des probabilités de petites boule de fonctions $\langle\theta, X\rangle$ ($\theta \in S^{d-1}$) en fonction de la vitesse de décroissance des fonctions caractéristiques des X^j (Proposition 6.6).

1.4 Estimation de densité et régression logistique

Dans cette section, nous considérons le problème de l'estimation de densité (conditionnelle ou non) avec risque de Kullback-Leibler, c'est-à-dire celui de l'apprentissage statistique avec perte logarithmique (Exemples 1.2 et 1.5). Après quelques définitions et compléments (Section 1.4.1) nous passons en revue deux approches standard du problème d'estimation de densité, à savoir le principe du *maximum de vraisemblance* (Section 1.4.2), ainsi que la réduction à la variante séquentielle du problème (Section 1.4.3).

Dans la Section 1.4.4, nous présentons ensuite notre principale contribution, à savoir une procédure générale pour l'estimation de densité admettant une borne d'excès de risque valide dans le cas mal spécifié. Cet estimateur général est alors appliqué aux modèles linéaire Gaussien (Section 1.4.5) et logistique (Section 1.4.7), pour lesquels nous obtenons de nouvelles bornes d'excès de risque valables uniformément sur toutes les lois P. Les résultats exposés dans les Sections 1.4.4, 1.4.5 et 1.4.7 sont tirés du Chapitre 7.

1.4.1 Préliminaires

Le problème de l'estimation de densité (Exemple 1.2) se formule de la façon suivante : étant donné un espace mesurable $\mathcal{Z}$ et une mesure de référence μ, on définit la perte logarithmique par $\ell(g, z) := -\log g(z)$ pour toute densité de probabilité g par rapport à μ et tout $z \in \mathcal{Z}$. Nous adoptons également un point de vue "apprentissage statistique", et fixons une classe $\mathcal{F}$ de densités de probabilité par rapport à μ. Étant donné un échantillon $Z_1, \dots, Z_n$ i.i.d. de loi P sur $\mathcal{Z}$, le but est de déterminer une densité $\widehat{g}_n$ pour laquelle l'excès de risque

$$\mathcal{E}(\widehat{g}_n) := R(\widehat{g}_n) - \inf_{f \in \mathcal{F}} R(f)\,,$$

où $R(g) := \mathbb{E}[\ell(g, Z)]$ avec $Z \sim P$, soit faible. En particulier, on a $R(g) - R(f) = \mathrm{KL}(P, g \cdot \mu) - \mathrm{KL}(P, f \cdot \mu)$ dès lors que ces quantités sont bien définies, de sorte que l'excès de risque ne dépend pas du choix de la mesure de domination μ (telle que le risque soit bien défini) ; par léger abus de notation, nous identifions donc une densité g à la mesure de probabilité $g \cdot \mu$ sur $\mathcal{Z}$. Étant donnée une famille $\mathcal{P}$ de lois P sur $\mathcal{Z}$ (qui ne sont pas nécessairement absolument continues par rapport à μ), l'excès de risque minimax (1.6) vaut :

$$\mathcal{E}_n^*(\mathcal{P}, \mathcal{F}) := \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \Big(\mathbb{E}[R(\widehat{g}_n)] - \inf_{f \in \mathcal{F}} R(f) \Big)\,. \tag{1.79}$$

La famille $\mathcal{F}$ de densités de probabilité par rapport à μ correspond à un *modèle statistique* sur $\mathcal{Z}$. Lorsque $\mathcal{P} = \mathcal{F}$ (où l'on identifie $\mathcal{F}$ à $\mathcal{F} \cdot \mu = \{f \cdot \mu : f \in \mathcal{F}\}$), on dit que le modèle $\mathcal{F}$ est *bien spécifié*. Conformément à l'approche discriminative exposée en Section 1.1.3, nous nous intéressons au cas général *mal spécifié* où $\mathcal{P} \neq \mathcal{F}$, et où $\mathcal{P}$ est une classe très riche de mesures de probabilités, qui n'impose que peu de restrictions à la loi P des observations.

Nous considérons également le problème de l'estimation de densité *conditionnelle* (Exemple 1.5), c'est-à-dire la variante supervisée de l'apprentissage séquentiel avec perte logarithmique. Dans ce cas, $\mathcal{X}, \mathcal{Y}$ sont des espaces mesurables, et μ est une mesure de référence sur $\mathcal{Y}$. On note alors $\mathcal{G}$ l'ensemble des fonctions mesurables de $\mathcal{X}$ vers l'espace des densités par rapport à μ, c'est-à-dire des densités conditionnelles sur $\mathcal{Y}$ sachant $x \in \mathcal{X}$. Pour $g \in \mathcal{G}$, on note $g(y|x) := g(x)(y)$, et l'on définit pour $g \in \mathcal{G}$ et $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(g, (x, y)) := -\log g(y|x)$. On considère également une classe $\mathcal{F} \subset \mathcal{G}$ de densités conditionnelles (également appelée *modèle*

statistique), ainsi qu'une famille $\mathcal{P}$ de lois jointes sur $\mathcal{X} \times \mathcal{Y}$. Il est alors possible, étant donnée une famille $\mathcal{P}$ de probabilités sur $\mathcal{X} \times \mathcal{Y}$, de définir l'excès de risque minimax (1.79).

En principe, lorsque la loi P_X de X est connue, le problème de l'estimation conditionnelle peut se ramener à celui de l'estimation non conditionnelle, en considérant l'espace $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ muni de la mesure de référence $\mu_{\mathcal{Z}} := P_X \otimes \mu$. Il est cependant utile de considérer le cas conditionnel, afin de distinguer les hypothèses sur X de celles sur la loi conditionnelle de Y sachant X. De plus, pour la variante séquentielle du problème (voir la Section 1.4.3 ci-dessous), il existe une différence non triviale entre ces deux cas. Il est en revanche possible de voir le cas non conditionnel comme un cas particulier du cas conditionnel, en prenant $\mathcal{Y} = \mathcal{Z}$ et $\mathcal{X} = \{x_0\}$ réduit à un point.

Comme pour le problème des moindres carrés (considéré en Section 1.3 et dans le Chapitre 6), il est possible soit d'étudier la dépendance de la difficulté du problème (c'est-à-dire de l'excès de risque minimax) en la loi de X, soit de considérer une loi P_X qui peut elle-même être choisie comme dans le pire des cas, au sein d'une certaine classe. Notons que, dans le cas conditionnel, dire que le modèle est bien spécifié signifie que la loi conditionnelle de Y sachant X appartient à $\mathcal{F}$.

1.4.2 Estimateur du maximum de vraisemblance

Considérons ici le problème de l'estimation de densité non conditionnelle, et notons $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. Une procédure naturelle est donnée par la minimisation du risque empirique (1.13) : $\widehat{f}_n^{\mathrm{ERM}} = f_{\widehat{\theta}_n^{\mathrm{EMV}}}$, où

$$\widehat{\theta}_n^{\mathrm{EMV}} := \arg\min_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta, Z_i) \right\} = \arg\max_{\theta \in \Theta} \left\{ \prod_{i=1}^{n} f_\theta(Z_i) \right\}. \tag{1.80}$$

L'estimateur (1.80) est appelé *estimateur du maximum de vraisemblance* (EMV). L'EMV est sans doute l'estimateur le plus classique, depuis son introduction par Fisher (Fisher, 1925). Les propriétés asymptotiques de l'EMV sont bien comprises dans le cas de modèle dits *réguliers*, que nous discutons ci-dessous.

Supposons que l'espace de paramètre Θ est un ouvert de $\mathbf{R}^d$, que la fonction $\theta \mapsto \ell(\theta, Z) := \ell(f_\theta, z)$ est suffisamment régulière et que la variable aléatoire $\ell(\theta, Z)$ (avec $Z \sim P$) admet suffisamment de moments contrôlés de manière uniforme sur Θ (au moins localement), et que le minimiseur du risque $\theta^* = \arg\min_\theta R(\theta)$ est unique ; pour davantage de détails sur les conditions de régularités requises, nous renvoyons à van der Vaart (1998); Ibragimov and Has'minskii (1981). Il est alors possible de montrer que, lorsque $n \to \infty$, $\widehat{\theta}_n^{\mathrm{EMV}} \to \theta^*$ presque sûrement. De plus, en notant

$$G := \mathbb{E}[\nabla \ell(\theta^*, Z) \nabla \ell(\theta^*, Z)^\top], \quad H := \mathbb{E}[\nabla^2 \ell(\theta^*, Z)]$$

(où la Hessienne et les gradients sont pris par rapport à θ, et où l'on suppose H positive), la variable aléatoire $\sqrt{n}(\widehat{\theta}_n^{\mathrm{EMV}} - \theta^*)$ converge en loi vers $\mathcal{N}(0, H^{-1} G H^{-1})$ (van der Vaart, 1998). Puisque $H = \nabla^2 R(\theta^*)$, pour $\theta \approx \theta^*$ on a $R(\theta) - R(\theta^*) \sim \|\theta - \theta^*\|_H^2/2 = \|H^{1/2}(\theta - \theta^*)\|^2/2$, de sorte que $n\big(R(\widehat{\theta}_n^{\mathrm{EMV}}) - R(\theta^*)\big) = (1 + o(1))\|H^{1/2}\sqrt{n}(\widehat{\theta}_n^{\mathrm{EMV}} - \theta^*)\|^2/2$ converge en loi vers $\|Z\|^2/2$, où $Z \sim \mathcal{N}(0, H^{-1/2} G H^{-1/2})$. Sous réserve que la convergence en loi ait également

lieu[11] dans L^1, on en déduit que

$$\mathbb{E}[\mathcal{E}(\widehat{\theta}_n^{\mathrm{EMV}})] = \big(1+o(1)\big)\frac{\mathbb{E}[\|Z\|^2]}{2n} = \frac{\mathrm{Tr}(\mathbb{E}[ZZ^\top]) + o(1)}{2n} = \frac{\mathrm{Tr}(H^{-1/2}GH^{-1/2})}{2n} + o\Big(\frac{1}{n}\Big) \quad (1.81)$$

Dans le cas *bien spécifié* où $P = P_{\theta^*}$, on a $G = H = I(\theta^*)$, qui est la matrice d'information de Fisher du modèle $\mathcal{F}$ en θ^*. On en déduit que $H^{-1/2}GH^{-1/2} = I_d$, de sorte que $2n \cdot \mathcal{E}(\widehat{\theta}_n^{\mathrm{EMV}})$ convergence en loi vers la loi χ^2_d, c'est-à-dire celle de $\|Z\|^2$ avec $Z \sim \mathcal{N}(0, I_d)$; de plus, l'équation (1.81) devient $\mathbb{E}[\mathcal{E}(\widehat{\theta}_n^{\mathrm{EMV}})] = d/(2n) + o(n^{-1})$. Dans ce cas, l'EMV jouit de propriétés d'optimalité asymptotique fortes : par exemple, sa variance asymptotique $I(\theta^*)/n$, ainsi que son excès de risque en $d/(2n) + o(n^{-1})$, sont asymptotiquement optimaux en un sens précis[12] (Hájek, 1972; Le Cam, 1986; Keener, 2010).

En revanche, lorsque le modèle est mal spécifié, la vitesse de convergence asymptotique (1.81) de l'EMV est de $d_{\mathrm{eff}}/(2n)$, où la constante $d_{\mathrm{eff}} := \mathrm{Tr}(H^{-1/2}GH^{-1/2})$ dépend de la vraie loi P. En particulier, cette constante peut généralement être arbitrairement grande en fonction de P, de sorte qu'une garantie d'excès de risque uniforme (1.79) avec une classe $\mathcal{P}$ riche n'est pas possible.

Exemple 1.14. Considérons $\mathcal{Z} = \mathbf{R}$, muni de la mesure $\mu(\mathrm{d}z) = (2\pi)^{-1/2}\mathrm{d}z$, ainsi que le modèle Gaussien $\mathcal{F} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbf{R}\}$. On a alors, pour tout $\theta, z \in \mathbf{R}$, $\ell(\theta, z) = (\theta - z)^2$, de sorte que $\theta^* = \mathbb{E}[Z]$ et que $G = \mathbb{E}[(\theta^* - Z)^2] = \mathrm{Var}(Z) =: \sigma^2$, tandis que $H = 1$, et donc $d_{\mathrm{eff}} = \sigma^2$. Cela correspond au fait que le risque de l'EMV $\widehat{\theta}_n^{\mathrm{EMV}} := n^{-1}\sum_{i=1}^n Z_i$, qui vaut σ^2/n, se dégrade lorsque la variance de Z est élevée.

Ainsi, la performance de l'EMV se dégrade dans le cas mal spécifié. Nous verrons par la suite, dans plusieurs exemples, que les limitations de l'EMV dans le cas mal spécifié sont communes à tous les estimateurs *propres* (ou de type *plug-in*), c'est-à-dire à tous les estimateur de la forme $\widehat{f}_n = f_{\widehat{\theta}_n}$ appartenant au modèle $\mathcal{F}$.

Enfin, signalons qu'un autre inconvénient des résultats mentionnés ci-dessus est leur caractère asymptotique : ces résultats sont valides lorsque la dimension d du modèle est fixée, tandis que la taille n de l'échantillon tend vers l'infini. L'inconvénient de telles garanties est qu'elles ne disent pas à partir de quelle valeur de n l'erreur devient faible : par exemple, il se pourrait que l'approximation asymptotique ne soit bonne (et le risque faible) que pour $n \geqslant e^d$. Les garanties asymptotiques ne donnent donc pas d'information sur un échantillon fini. De plus, dans une perspective non paramétrique (Tsybakov, 2009; Wasserman, 2006), il peut être désirable de faire croître la dimension $d = d_n$ du modèle avec n, ce qui sort du cadre asymptotique classique où d est fixé (Massart, 2007). Pour ces raisons, il est souhaitable d'obtenir des garanties explicites et non-asymptotiques. Les résultats asymptotiques ne servent ici qu'à illustrer les limitations de l'EMV dans le cas mal spécifié.

[11] Cette étape est non triviale, et parfois même fausse, sans domination suffisante sur $n\big(R(\widehat{\theta}_n^{\mathrm{EMV}}) - R(\theta^*)\big)$. Par exemple, dans le cas de la régression logistique, l'espérance du risque de l'EMV est typiquement infinie (Chapitre 7). En revanche, le lemme de Fatou (Durrett, 2010) fournit une *borne inférieure* sur l'excès de risque en espérance de l'EMV.

[12] Par exemple, la variance asymptotique $I(\theta^*)/n$ est optimale parmi les M-estimateurs réguliers, et la vitesse en $d/(2n)$ est optimale en un sens asymptotiquement localement minimax, c'est-à-dire minimax sur les voisinages de θ^* de taille $K/\sqrt{n}$, avec $K \gg 1$ (Hájek, 1972; Le Cam, 1986). Ces définitions précises sont motivées par l'existence de pathologies telles que les estimateurs super-efficients, exhibée par Hodges (voir par exemple Lehmann and Casella, 1998, Exemple 2.5 et Keener, 2010, Exemple 1.6).

1.4.3 Prédiction séquentielle avec perte logarithmique

Dans cette section, nous décrivons brièvement une variante du problème considéré, à savoir la *prédiction séquentielle avec perte logarithmique*, ou *estimation de densité en ligne* (Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006) ; cette théorie est intimement liée à celle de la compression sans perte de données, étudiée en théorie de l'information (Cover and Thomas, 2006), ainsi qu'à celle de la longueur minimale de description (en anglais *Minimum Description Length*, abrégé MDL ; voir Rissanen, 1985; Barron et al., 1998; Grünwald, 2007). Ce problème est un cas particulier de celui de la prédiction séquentielle, considérée dans la Section 1.2 de cette introduction, ainsi que dans les Chapitres 4 et 5.

Stratégies de prédiction, regret. Tout d'abord, rappelons que la perte logarithmique $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ est 1-exp-concave (Exemple 1.13) ; en particulier, les bornes de regret de la Section 1.2.4 lui sont applicables. Cependant, dans le cas de la perte logarithmique, le regret admet une expression simple, qui permet de réinterpréter l'agrégation à poids exponentiels (ainsi que la preuve de sa borne de regret) et d'obtenir une expression exacte pour le regret minimax. Les résultats mentionnés ci-dessous sont classiques ; nous renvoyons à Cesa-Bianchi and Lugosi (2006, Chapitre 9) ou à Merhav and Feder (1998) pour plus de détails.

Une *stratégie de prédiction* (c'est-à-dire, un algorithme de prédiction séquentielle) correspond à une suite $(g_t)_{1 \leqslant t \leqslant n}$, où $g_t : \mathcal{Z}^{t-1} \to \mathcal{G}$ associe aux observations passées $z_1, \dots, z_{t-1}$ une densité $\widehat{g}_t = g_t(z_1, \dots, z_{t-1}) \in \mathcal{Z}$. Notons que g_t peut être vu comme une densité conditionnelle sur $\mathcal{Z}$ sachant $(z_1, \dots, z_{t-1}) \in \mathcal{Z}^{t-1}$, en notant $g_t(z|z_1, \dots, z_{t-1}) := g_t(z_1, \dots, z_{t-1})(z)$. Il en découle que la donnée d'une stratégie de prédiction $(g_1, \dots, g_n)$ équivaut à celle d'une densité jointe sur $\mathcal{Z}^n$ (par rapport à μ^n). En effet, on peut associer à la stratégie de prédiction $(g_1, \dots, g_n)$ la densité sur $\mathcal{Z}^n$ donnée par

$$\bar{g}_n(z_1, \dots, z_n) := \prod_{t=1}^{n} g_t(z_t|z_1, \dots, z_{t-1})$$

pour tout $(z_1, \dots, z_n) \in \mathcal{Z}^n$. Réciproquement, à la densité $\bar{g}_n$ sur $\mathcal{Z}^n$ est associée[13] la stratégie de prédiction $(g_1, \dots, g_n)$, où pour $1 \leqslant t \leqslant n$,

$$g_t(z_t|z_1, \dots, z_{t-1}) := \frac{\int_{\mathcal{Z}^{n-t}} \bar{g}_n(z_1, \dots, z_{t-1}, z_t, z'_{t+1}, \dots, z'_n)\mu(\mathrm{d}z'_{t+1}) \cdots \mu(\mathrm{d}z'_n)}{\int_{\mathcal{Z}^{n-t+1}} \bar{g}_n(z_1, \dots, z_{t-1}, z'_t, \dots, z'_n)\mu(\mathrm{d}z'_t) \cdots \mu(\mathrm{d}z'_n)} ;$$

les deux transformations ci-dessous sont inverses l'une de l'autre. Dans ce qui suit, étant donnée une densité p sur $\mathcal{Z}^n$, nous notons pour $1 \leqslant t \leqslant n$:

$$p(z_1, \dots, z_t) = \int_{\mathcal{Z}^{n-t}} p(z_1, \dots, z_t, z'_{t+1}, \dots, z'_n)\mu(\mathrm{d}z'_{t+1}) \cdots \mu(\mathrm{d}z'_n) ,$$
$$p(z_t|z_1, \dots, z_{t-1}) = \frac{p(z_1, \dots, z_t)}{p(z_1, \dots, z_{t-1})} .$$

L'erreur cumulée d'une stratégie de prédiction $\bar{g}_n$ sur une suite $z_1, \dots, z_n$ est simplement donnée par l'opposé du logarithme de la probabilité jointe attribuée à cette suite :

$$\sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) = \sum_{t=1}^{n} -\log g_t(z_t|z_1, \dots, z_{t-1}) = -\log \prod_{t=1}^{n} g_t(z_t|z_1, \dots, z_{t-1}) = -\log \bar{g}_n(z_1, \dots, z_n).$$

[13] Nous supposons ici les densités strictement positives sur $\mathcal{Z}$; cette restriction sur les stratégies considérées est naturelle, car une densité nulle en un point conduit à une perte infinie.

Considérons également une famille $(\bar{f}_\theta)_{\theta\in\Theta}$ de stratégies de prédiction, indexée par le paramètre $\theta \in \Theta$ (nous verrons par la suite à quoi correspond $\bar{f}_\theta$ dans les cas non conditionnel et conditionnel). En appliquant également le résultat ci-dessus à $\bar{f}_\theta$, on obtient l'expression suivante pour le regret :

$$\sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \sum_{t=1}^{n} \ell(f_{\theta,t}, z_t) = -\log\Big(\frac{\bar{g}_n(z_1,\dots,z_n)}{\bar{f}_\theta(z_1,\dots,z_n)}\Big)\,, \tag{1.82}$$

où $f_{\theta,t} := \bar{f}_\theta(\cdot|z_1,\dots,z_{t-1})$.

Poids exponentiels et mélange Bayésien. Commençons par exprimer l'agrégation à poids exponentiels dans le formalisme précédent. Pour tout $t = 1,\dots,n$, en notant $L_{t-1}(\theta)$ la perte cumulée de θ jusqu'à l'instant $t-1$ et en prenant $\eta = 1$, on a

$$\exp(-\eta L_{t-1}(\theta)) = \prod_{s=1}^{t-1} f_{\theta,s}(z_s) = \bar{f}_\theta(z_1,\dots,z_{t-1})\,,$$

de sorte que le postérieur $\widehat{\pi}_t := \pi_{\exp(-\eta L_{t-1})}$ a pour densité $\theta \mapsto \bar{f}_\theta(z_1,\dots,z_{t-1})$ par rapport à π ; il s'agit donc du *postérieur Bayésien* sur Θ (Berger, 1985; Robert, 2007; Gelman et al., 2013). Si l'on considère une variable aléatoire $(\Theta, Z_1,\dots,Z_n)$ sur $\Theta\times\mathcal{Z}^n$ telle que $\Theta\sim\pi$ et $(Z_1,\dots,Z_n)|\Theta \sim \bar{f}_\Theta\cdot\mu^n$, alors $\widehat{\pi}_t$ est la loi conditionnelle de Θ sachant $Z_1,\dots,Z_{t-1}$. De plus, l'APE (1.31) avec pour loi a priori π sur Θ s'écrit

$$\widehat{g}_t(z_t) = \frac{\int_\Theta \bar{f}_\theta(z_t|z_1,\dots,z_{t-1})\bar{f}_\theta(z_1,\dots,z_{t-1})\pi(\mathrm{d}\theta)}{\int_\Theta \bar{f}_\theta(z_1,\dots,z_{t-1})\pi(\mathrm{d}\theta)} = \frac{\int_\Theta \bar{f}_\theta(z_1,\dots,z_{t-1},z_t)\pi(\mathrm{d}\theta)}{\int_\Theta \bar{f}_\theta(z_1,\dots,z_{t-1})\pi(\mathrm{d}\theta)}\,,$$

ce qui correspond au *postérieur prédictif Bayésien* ; si $\Theta\sim\pi$, et $(Z_1,\dots,Z_n)|\Theta\sim\bar{f}_\Theta$, alors $\widehat{g}_t$ est la densité conditionnelle de Z_t sachant $Z_1,\dots,Z_{t-1}$. Enfin, la stratégie de prédiction $\bar{g}_n$ de l'APE correspond à la loi jointe

$$\bar{g}_n(z_1,\dots,z_n) = \prod_{t=1}^{n} \frac{\int_\Theta \bar{f}_\theta(z_1,\dots,z_t)\pi(\mathrm{d}\theta)}{\int_\Theta \bar{f}_\theta(z_1,\dots,z_{t-1})\pi(\mathrm{d}\theta)} = \int_\Theta \bar{f}_\theta(z_1,\dots,z_n)\pi(\mathrm{d}\theta)\,, \tag{1.83}$$

qui est simplement le *mélange* des stratégies $(\bar{f}_\theta)_{\theta\in\Theta}$ selon la loi a priori π. Pour cette raison, l'APE pour la perte logarithmique est également appelée la stratégie de *mélange Bayésien*. Dans le cas où Θ est discret, on retrouve directement la borne de regret (1.38) (Corollaire 1.2) en combinant (1.82), (1.83) et le fait que pour $\theta\in\Theta$ et $z_1,\dots,z_n\in\mathcal{Z}$:

$$\bar{g}_n(z_1,\dots,z_n) = \sum_{\theta'\in\Theta} \pi(\theta')\bar{f}_{\theta'}(z_1,\dots,z_n) \geqslant \pi(\theta)\bar{f}_\theta(z_1,\dots,z_n)\,.$$

L'un des avantages des stratégies de mélange Bayésien est qu'elles sont calculables *en ligne*, et ne nécessitent pas de connaître à l'avance les prédictions $f_{\theta,s}$ pour $t < s \leqslant n$. De plus, dans le cas paramétrique et sous des hypothèses assez générales, ces stratégies admettent un regret asymptotiquement minimax en $d\log(n)/2 + O(1)$ (Xie and Barron, 2000; Merhav and Feder, 1998). Cette vitesse de regret peut être vue comme une variante cumulée de la vitesse optimale d'excès de risque en $d/(2n) + o(1/n)$ dans le cas statistique bien spécifié.

La principale limitation pratique de l'approche bayésienne est son coût computationnel : en effet, son implémentation requiert de calculer l'intégrale $\int_\Theta f_\theta(z_1)\cdots f_\theta(z_t)\pi(\mathrm{d}\theta)$. Bien que

dans certains cas, cette quantité admette une expression exacte (on parle alors de *familles conjuguées*, voir par exemple Berger, 1985; Robert, 2007; Gelman et al., 2013), son calcul se fait en général de manière approchée, avec un coût non négligeable ; nous reviendrons sur ce point dans le cas de la régression logistique (Section 1.4.6).

Regret et stratégie minimax. Considérons à présent le regret minimax (1.24) (sur toutes les suites $(z_1, \dots, z_n) \in \mathcal{Z}^n$) par rapport à la famille $(\bar{f}_\theta)_{\theta\in\Theta}$. Il découle de (1.82) que le regret de $\bar{g}_n$ s'écrit, pour tous $z_1, \dots, z_n \in \mathcal{Z}$,

$$\sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} \ell(f_{\theta,t}, z_t) = -\log\left(\frac{\bar{g}_n(z_1, \dots, z_n)}{\sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)}\right);$$

le regret minimax vaut donc

$$\mathcal{R}_n := \sup_{z_1,\dots,z_n\in\mathcal{Z}} \left(\sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} \ell(f_{\theta,t}, z_t) \right) = \sup_{z_1,\dots,z_n\in\mathcal{Z}} -\log\left(\frac{\bar{g}_n(z_1, \dots, z_n)}{\sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)}\right).$$

Si $\mathcal{R}_n < +\infty$, on a pour tous $z_1, \dots, z_n \in \mathcal{Z}$, $\bar{g}_n(z_1, \dots, z_n) \geqslant e^{-\mathcal{R}_n} \sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)$, de sorte que

$$1 = \int_{\mathcal{Z}^n} \bar{g}_n(z_1, \dots, z_n)\mu(\mathrm{d}z_1)\cdots\mu(\mathrm{d}z_n) \geqslant e^{-\mathcal{R}_n} \int_{\mathcal{Z}^n} \sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)\mu(\mathrm{d}z_1)\cdots\mu(\mathrm{d}z_n),$$

soit

$$\mathcal{R}_n \geqslant \mathcal{R}_n^* := \log\left(\int_{\mathcal{Z}^n} \sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)\mu(\mathrm{d}z_1)\cdots\mu(\mathrm{d}z_n) \right).$$

Réciproquement, si $\mathcal{R}_n^* < +\infty$, la stratégie

$$\bar{g}_n^*(z_1, \dots, z_n) = \frac{\sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)}{\int_\Theta \sup_{\theta\in\Theta} \bar{f}_\theta(z'_1, \dots, z'_n)\mu(\mathrm{d}z'_1)\cdots\mu(\mathrm{d}z'_n)} \tag{1.84}$$

admet un regret constant de $\mathcal{R}_n^*$, de sorte que $\mathcal{R}_n \leqslant \mathcal{R}_n^*$. Ainsi, le regret minimax vaut

$$\log\left(\int_{\mathcal{Z}^n} \sup_{\theta\in\Theta} \bar{f}_\theta(z_1, \dots, z_n)\mu(\mathrm{d}z_1)\cdots\mu(\mathrm{d}z_n) \right), \tag{1.85}$$

et est atteint par la stratégie minimax $\bar{g}_n^*$ (1.84), dite du *maximum de vraisemblance normalisé* (en anglais *Normalized Maximum Likelihood*, NML). Le regret minimax (1.85) est parfois appelé *complexité NML* ou *intégrale de Shtarkov*. Ce résultat est dû à Shtarkov (1987).

Le cas non conditionnel. Considérons le cas de l'estimation de densité en ligne, par rapport à une classe de densité $(f_\theta)_{\theta\in\Theta}$ sur $\mathcal{Z}$. On cherche alors à contrôler le regret

$$\sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} \ell(f_\theta, z_t), \tag{1.86}$$

ce qui revient à se comparer à la classe des stratégies de prédictions *statiques* (ou *constantes*) $\bar{f}_\theta(z_1, \dots, z_n) := f_\theta(z_1)\cdots f_\theta(z_n)$. Dans ce cas, le regret minimax (1.85) s'écrit

$$\sup_{z_1,\dots,z_n\in\mathcal{Z}} \left(\sum_{t=1}^{n} \ell(\widehat{g}_t, z_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} \ell(f_\theta, z_t) \right) = \log\left(\int_{\mathcal{Z}^n} \sup_{\theta\in\Theta} f_\theta(z_1)\cdots f_\theta(z_n)\mu(\mathrm{d}z_1)\cdots\mu(\mathrm{d}z_n) \right),$$

et est atteint (lorsqu'il est fini) par la stratégie

$$\bar{g}_n^*(z_1,\dots,z_n) = \frac{\sup_{\theta\in\Theta}\left[f_\theta(z_1)\cdots f_\theta(z_n)\right]}{\int_\Theta \sup_{\theta\in\Theta}\left[f_\theta(z_1)\cdots f_\theta(z_n)\right]\mu(\mathrm{d}z_1')\cdots\mu(\mathrm{d}z_n')}\,.$$

Dans le cas d'une famille $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ paramétrique régulière de dimension d, avec un espace de paramètre "borné", le regret minimax asymptotique lorsque $n \to \infty$ est de

$$\frac{d}{2}\log\left(\frac{n}{2\pi}\right) + C(\mathcal{F}) + o(1) \tag{1.87}$$

avec

$$C(\mathcal{F}) = \log\int_\Theta \sqrt{\det I(\theta)}\mathrm{d}\theta\,,$$

où $I(\theta)$ désigne la matrice d'information de Fisher définie plus haut[14].

Limitations de la stratégie NML. La stratégie NML est la stratégie minimax exacte, ce qui est un résultat remarquable. En revanche, cette procédure pose certaines difficultés pratiques : (1) elle dépend généralement de l'horizon de temps n, de sorte qu'il n'est pas possible de l'utiliser pour un nombre d'étapes indéterminé; (2) son calcul nécessite de considérer toutes les suites possibles $(z_1,\dots,z_n) \in \mathcal{Z}^n$, puisque sa première prédiction est donnée par

$$\bar{g}_n^*(z_1) = \frac{\int_{\mathcal{Z}^{n-1}} \sup_{\theta\in\Theta}\left[f_\theta(z_1)f_\theta(z_2')\cdots f_\theta(z_n')\right]\mu(\mathrm{d}z_2')\dots\mu(\mathrm{d}z_n')}{\int_{\mathcal{Z}^n} \sup_{\theta\in\Theta}\left[f_\theta(z_1')\cdots f_\theta(z_n')\right]\mu(\mathrm{d}z_1')\dots\mu(\mathrm{d}z_n')}\,.$$

Ainsi, même lorsque $|\mathcal{Z}| = 2$, la complexité du calcul de la première prédiction de $\bar{g}_n^*$ est exponentielle en l'horizon de temps n. Dans certains cas cependant, la stratégie NML ne dépend pas de l'horizon de temps, et coïncide dans ce cas avec une stratégie de mélange Bayésien (Bartlett et al., 2013; Hedayati and Bartlett, 2017).

Une autre limitation tient au fait que le regret minimax est infini pour certaines familles non bornées, comme le montre l'exemple suivant :

Exemple 1.15. Soit $\mathcal{F} = \{\mathcal{N}(\theta,1) : \theta \in \mathbf{R}\}$ le modèle (de translation) Gaussien sur $\mathbf{R}$. Considérons la mesure de référence $\mu = (2\pi)^{-1/2}\mathrm{d}z$. Alors, pour tous $z_1,\dots,z_n \in \mathbf{R}$, on a $\sup_{\theta\in\mathbf{R}}[f_\theta(z_1)\cdots f_\theta(z_n)] = \exp(-\sum_{i=1}^n (z_i - \bar{z}_n)^2)$ avec $\bar{z}_n := n^{-1}\sum_{i=1}^n z_i$. En particulier, cette quantité est minorée par $e^{-1/2}$ sur le "tube" (de mesure de Lebesgue μ infinie) $\{(z_1 + t,\dots,z_{n-1}+t,t) : \sum_{i=1}^{n-1} z_i^2 \leqslant 1, t \in \mathbf{R}\}$, de sorte que le regret minimax (1.85) est infini. Cela tient au fait que le regret est non borné dès la première observation.

Le cas conditionnel. Considérons à présent le problème de l'estimation de densité conditionnelle en ligne (Section 1.4.1). À tout $\theta \in \Theta$ correspond une densité conditionnelle $x \mapsto f_\theta(\cdot|x)$ par rapport à la mesure de référence μ sur $\mathcal{Y}$. Étant donnée la suite $(x_1,\dots,x_n) \in \mathcal{X}^n$ des covariables, on associe à tout $\theta \in \Theta$ la stratégie de prédiction

$$\bar{f}_\theta(y_1,\dots,y_n|x_1,\dots,x_n) := f_\theta(y_1|x_1)\cdots f_\theta(y_n|x_n)\,, \tag{1.88}$$

dont la perte cumulée est celle de f_θ. Il résulte donc de ce qu'il précède que, lorsque la suite $(x_1,\dots,x_n) \in \mathcal{X}^n$ est fixée et connue à l'avance, le regret minimax (sur la suite $(y_1,\dots,y_n) \in$

[14] Le premier terme dans l'expression de $C(\mathcal{F})$ correspond au logarithme du volume de $\mathcal{F}$, sous la forme volume définie par la métrique d'information de Fisher.

$\mathcal{Y}^n$) par rapport à $\mathcal{F} = (f_\theta)_{\theta\in\Theta}$ vaut

$$\log\left(\int_{\mathcal{Y}^n} \sup_{\theta\in\Theta} \left[f_\theta(y_1|x_1)\cdots f_\theta(y_n|x_n)\right]\mu(\mathrm{d}y_1)\cdots\mu(\mathrm{d}y_n)\right),$$

et est atteint par la stratégie de prédiction

$$\bar{g}_n^*(y_1,\ldots,x_n|x_1,\ldots,x_n) = \frac{\sup_{\theta\in\Theta}\left[f_\theta(y_1|x_1)\cdots f_\theta(y_n|x_n)\right]}{\int_{\mathcal{Y}^n}\sup_{\theta\in\Theta}\left[f_\theta(y_1'|x_1)\cdots f_\theta(y_n'|x_n)\right]\mu(\mathrm{d}y_1')\cdots\mu(\mathrm{d}y_n')}.$$

La principale limitation de cette approche est qu'elle requiert de connaître la suite $(x_1,\ldots,x_n)$ à l'avance, ce qui est rarement le cas en pratique. Dans le cas contraire, déterminer la stratégie minimax revient à résoudre un problème minimax non trivial (Rakhlin and Sridharan, 2015; Foster et al., 2018). De plus, même dans le cas où la suite $(x_1,\ldots,x_n)$ est connue à l'avance, le calcul des prédictions associées requiert de considérer toutes les suites $(y_1,\ldots,y_n)\in\mathcal{Y}^n$ afin de calculer la constante de normalisation. Dans ce cas, la stratégie de mélange bayésien (qui ne requiert pas de connaître $(x_1,\ldots,x_n)$ à l'avance) est préférable à la stratégie minimax.

Conclusion : limites de la réduction au problème en ligne. Les limitations suivantes tiennent au fait que le problème de la prédiction séquentielle est plus difficile que celui de l'apprentissage statistique "hors ligne". Notons au passage que la principale différence tient au caractère cumulé du critère (qui inclut donc des instants $t \ll n$ pour lesquels peu d'observations sont disponibles), davantage qu'au caractère non stochastique du problème. En effet, l'excès de risque cumulé est du même ordre de grandeur que le regret minimax (Grünwald, 2007).

1. **Vitesse sous-optimale :** les vitesses obtenues par conversion online-to-batch sont sous-optimales d'un facteur logarithmique ($O(d\log(n)/n)$ au lieu de $O(d/n)$). L'obtention de vitesses en d/n pour l'apprentissage statistique avec perte logarithmique est un problème ouvert (Grünwald and Kotłowski, 2011). D'un point de vue pratique, le fait de considérer la moyenne des itérés a tendance à dégrader la qualité de l'estimateur, à cause de la mauvaise qualité des premiers itérés (obtenus à partir de peu d'observations).

2. **Dépendance aux conditions initiales/à la complexité globale de $\mathcal{F}$:** Outre le facteur logarithmique sous-optimal, la borne d'excès de risque asymptotique (1.87) comporte un terme en $C(\mathcal{F})/n$; ce terme de complexité "globale" de $\mathcal{F}$ (lié à son volume) reflète le fait que la borne n'est pas suffisamment "localisée" autour de θ^*. Cela tient intuitivement au fait que l'on effectue la moyenne de bornes d'excès de risque à différents instants $1 \leqslant t \leqslant n$, correspondant à différents degrés de localisation.

3. **Familles non bornées :** De même, pour certaines classes (même paramétriques) de "volume" infini, le regret minimax est infini, tandis que l'excès de risque minimax est fini (Exemple 1.15). Cela est dû au caractère non borné de la classe, qui induit un regret non borné aux premiers instants dans le problème de prédiction en ligne.

4. **Cas conditionnel :** Dans le cas conditionnel, le regret (et la stratégie) minimax sont difficiles à déterminer. De plus, le fait de considérer les pires entrées x_t possibles est trop conservateur, car la loi de X peut avoir certaines régularités permettant d'obtenir de plus ou moins bonnes bornes de risque.

À ces limitations générales de la réduction au problème séquentiel s'ajoutent des difficultés spécifiques aux procédures considérées (la stratégie minimax NML et les stratégies de mélange Bayésien), en particulier leur complexité computationnelle.

1.4.4 Un estimateur (presque) optimal dans le cas mal spécifié (Chapitre 7)

Dans le Chapitre 7, nous introduisons une procédure générale pour l'estimation de densité (conditionnelle ou non), qui satisfait une borne générale d'excès de risque, valide dans le cas général mal spécifié. Dans cette section, nous décrivons brièvement cette procédure, ainsi que sa borne d'excès de risque, et des cas particuliers simples dans le cas non conditionnel. Dans les sections suivantes, nous étudierons cette procédure appliquée à deux modèles conditionnels classiques, à savoir le *modèle linéaire Gaussien* ainsi que le *modèle logistique*.

La procédure que nous introduisons, appelée *Sample Minmax Predictor* (SMP), est en fait valide pour l'apprentissage supervisé avec une fonction de perte générale. Elle apparaît naturellement comme la procédure minimisant une nouvelle borne d'excès de risque générale pour l'apprentissage supervisé. Nous reprenons les notations de la Section 1.1.2 ; en particulier, $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ désigne la fonction de perte, et $\mathcal{F}$ une classe de fonctions $\mathcal{X} \to \widehat{\mathcal{Y}}$. Pour $z = (x, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ et $g : \mathcal{X} \to \widehat{\mathcal{Y}}$ un prédicteur, on note $\ell(g, z) := \ell(g(x), y)$.

Théorème 1.11 (Théorème 7.1, Chapitre 7)**.** *Soit $\widehat{g}_n$ un estimateur dépendant de l'échantillon i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$. Notons $Z_i = (X_i, Y_i)$, et pour tout $z \in \mathcal{Z}$*

$$\widehat{f}_n^{(z)} := \arg\min_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \ell(f, Z_i) + \ell(f, z) \right\}.$$

Alors, l'excès de risque $\mathcal{E}(\widehat{g}_n) := R(\widehat{g}_n) - \inf_{f \in \mathcal{F}} R(f)$ de $\widehat{g}_n$ satisfait :

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n)\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{g}_n(X), y) - \ell(\widehat{f}_n^{(X,y)}(X), y) \Big\} \Big] \tag{1.89}$$

où $Z = (X, Y) \sim P$ est indépendant de Z_1^n. De plus, la borne (1.89) est minimisée par le prédicteur

$$\widehat{g}_n^{\mathrm{SMP}}(x) = \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell(\widehat{f}_n^{(x,y)}(x), y) \Big\}, \tag{1.90}$$

que nous appelons SMP lorsqu'il est bien défini. Dans ce cas, la borne générale (1.89) s'écrit

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\inf_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell(\widehat{f}_n^{(X,y)}(X), y) \Big\} \Big]. \tag{1.91}$$

Le SMP admet également une variante régularisée, avec une borne d'excès de risque correspondante (voir l'énoncé exact du Théorème 7.1 du Chapitre 7). Dans le cas de la perte logarithmique, le SMP (1.90) et sa borne d'excès de risque (1.91) admettent une expression explicite.

Théorème 1.12 (Théorème 7.2, Chapitre 7)**.** *Dans le cas de l'estimation de densité conditionnelle, le SMP s'écrit*

$$\widehat{g}_n^{\mathrm{SMP}}(y|x) = \frac{\widehat{f}_n^{(x,y)}(y|x)}{\int_{\mathcal{Y}} \widehat{f}_n^{(x,y')}(y'|x) \mu(\mathrm{d}y')} \tag{1.92}$$

dès lors que le dénominateur de (1.92) est fini. De plus, son excès de risque est borné de la façon suivante :

$$\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})] \leqslant \mathbb{E}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_n^{(X,y)}(y|X) \mu(\mathrm{d}y) \Big) \Big]. \tag{1.93}$$

Notons que le SMP est en général un prédicteur *impropre*, tout comme les prédicteurs obtenus par (conversion online-to-batch de) mélange bayésien ou NML. Nous verrons que cet estimateur contourne les limitations inhérentes aux estimateurs propres (comme les approches fondées sur la conversion online-to-batch), et permet d'obtenir des bornes en $d/n + o(1/n)$ (contrairement à ces dernières).

L'expression (1.93) du SMP fait apparaître une intégrale comme constante de renormalisation. Pour les exemples que nous allons considérer, cette constante se calcule explicitement. Cependant, notons que, contrairement aux approches Bayésiennes où l'intégrale de la constante de renormalisation porte sur le paramètre $\theta \in \Theta$, celle-ci porte sur la réponse $y \in \mathcal{Y}$. Dans de nombreux exemples d'estimation de densité conditionnelle, l'espace des paramètres Θ est bien plus complexe que celui des sorties $\mathcal{Y}$. C'est notamment le cas pour le problème de la régression logistique, dont nous discuterons en Section 1.4.7, où $\mathcal{Y} = \{-1, 1\}$ tandis que $\Theta = \mathbf{R}^d$.

Exemple 1.16 (Modèle Gaussien, Proposition 7.2)**.** Dans le cas du modèle Gaussien $\mathcal{F} = \{\mathcal{N}(\theta, I_d) : \theta \in \mathbf{R}^d\}$, le SMP vaut $\widehat{g}_n^{\mathrm{SMP}} = \mathcal{N}(\bar{Z}_n, (1 + 1/n)^2 I_d)$, et sa borne d'excès de risque (1.93) vaut $d \log(1 + 1/n) \leqslant d/n$, quelle que soit la loi de Z telle que $\mathbb{E}[\|Z\|] < +\infty$.

À l'inverse, pour l'EMV et plus généralement tout estimateur propre $f_{\widehat{\theta}_n}$, une dépendance en la quantité $d_{\mathrm{eff}} = \mathbb{E}[\|Z\|^2]$ est inévitable : pour tout $t > 0$, il est impossible d'obtenir une borne uniformément meilleure que $t/(2n)$ sur la classe des lois de Z telles que $\mathbb{E}[\|Z\|^2] = t$ (Section 7.3.2).

En outre, le regret minimax (1.85) par rapport à la classe $\mathcal{F}$ est infini (Exemple 1.15), il n'est donc pas possible d'obtenir de garantie d'excès de risque uniforme sur $\mathcal{F}$ par conversion online-to-batch.

Exemple 1.17 (Modèle multinomial, Proposition 7.1)**.** Considérons le cas où $\mathcal{Z}$ est fini, de cardinal d, et considérons le modèle multinomial $\mathcal{F} = \mathcal{P}(\mathcal{Z})$ (qui est toujours bien spécifié). Dans ce cas, le SMP correspond à l'*estimateur de Laplace*, donné par $\widehat{g}_n^{\mathrm{SMP}}(z) = (N_n(z) + 1)/(n + d)$, où $d = |\mathcal{Z}|$ et où $N_n(z)$ est le nombre d'occurrences de z parmi $Z_1, \ldots, Z_n$. De plus, sa borne d'excès de risque (1.93) est de $\log[(n + d)/(n + 1)] \leqslant (d - 1)/n$.

L'excès de risque de l'EMV est quant à lui infini avec probabilité positive (Section 7.3.1), donc aussi en espérance. Enfin, le regret minimax étant d'ordre $(d - 1)(\log n)/2 + O(1)$ (le nombre de paramètres est ici $d-1$), la conversion online-to-batch ne peut fournir qu'une borne en $\Theta(d \log(n)/n)$.

1.4.5 Application au modèle linéaire Gaussien (Chapitre 7 et Section 8.1)

Considérons à présent les espaces $\mathcal{X} = \mathbf{R}^d$ et $\mathcal{Y} = \mathbf{R}$, $\mu = (2\pi)^{-1/2}\mathrm{d}y$ et la famille de densités conditionnelles $\mathcal{F} = \{f_\beta : \beta \in \mathbf{R}^d\}$, où $f_\beta(\cdot|x) = \mathcal{N}(\langle \beta, x\rangle, 1)$. Ainsi, pour tous $\beta \in \mathbf{R}^d$ et $(x, y) \in \mathbf{R}^d \times \mathbf{R}$,

$$\ell(\beta, (x, y)) = \frac{1}{2}(\langle \beta, x\rangle - y)^2 .$$

Lorsque l'on se restreint aux prédicteurs propres, de la forme $\widehat{f}_n = f_{\widehat{\beta}_n}$, le problème est donc équivalent à celui des moindres carrés mentionné à la Section 1.3 et étudié dans le Chapitre 6. Cependant, le problème est de nature différente, puisqu'il s'agit d'effectuer une prédiction *probabiliste* de la réponse, c'est-à-dire d'estimer la loi conditionnelle de Y sachant X et non son espérance conditionnelle (Exemple 1.5). La possibilité d'utiliser des estimateurs impropres

permet à nouveau d'obtenir des garanties améliorées pour ce problème, en particulier dans le cas mal spécifié.

Borne d'excès de risque uniforme. Commençons par considérer le SMP (1.92) appliqué à la classe $\mathcal{F}$, ainsi que la borne d'excès de risque correspondante.

Théorème 1.13 (Théorème 7.4, Chapitre 7)**.** *Dans le cas du modèle linéaire Gaussien, le SMP vaut* $\widehat{g}_n^{\mathrm{SMP}}(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle, (1 + \langle (n\widehat{\Sigma}_n)^{-1}x, x\rangle)^2)$*. De plus, si* P_X *est non dégénérée (au sens discuté en Section 1.3.3) et si* $\mathbb{E}[Y^2] < +\infty$*, sa borne d'excès de risque* (1.93) *vaut*

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \mathbb{E}\big[\log(1 + \langle (n\widehat{\Sigma}_n)^{-1}X, X\rangle)\big] \leqslant \mathbb{E}\Big[\log\Big(1 + \frac{1}{n}\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\Big)\Big]\,. \tag{1.94}$$

De plus, dans le cas bien spécifié, on a $\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})] = \mathbb{E}\big[\log(1+\langle (n\widehat{\Sigma}_n)^{-1}X, X\rangle)\big] - d/(2(n+1))$.

La borne (1.94) montre que l'excès de risque $\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})]$ est d'au plus $\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)]/n$, c'est-à-dire au plus deux fois celui de l'EMV dans le cas bien spécifié[15]. Par le Théorème 1.9, il en découle que sous les Hypothèses 1.1 et 1.2 sur la loi P_X, l'excès de risque du SMP vérifie

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \frac{d}{n}\Big(1 + 8C'\frac{\kappa d}{n}\Big) = \frac{d}{n} + O\Big(\Big(\frac{d}{n}\Big)^2\Big)\,,$$

uniformément sur toutes les lois de Y sachant X telles que $\mathbb{E}[Y^2] < +\infty$ (Corollaire 7.1, Chapitre 7). À l'inverse, le risque de l'EMV dans le cas mal spécifié est de

$$\frac{1}{2n}\mathbb{E}\big[(Y - \langle \beta^*, X\rangle)^2 \|\Sigma^{-1/2}X\|^2\big] + O\Big(\Big(\frac{d}{n}\Big)^{3/2}\Big)$$

sous des hypothèses convenables sur P_X (Proposition 1.6). Cette vitesse dépend de l'erreur d'approximation du modèle linéaire, soit $\mathbb{E}[Y|X] - \langle \beta^*, X\rangle$, ainsi que de la variance conditionnelle $\mathrm{Var}(Y|X)$, puisque $\mathbb{E}[(Y - \langle \beta^*, X\rangle)^2|X] = (\mathbb{E}[Y|X] - \langle \beta^*, X\rangle)^2 + \mathrm{Var}(Y|X)$. Plus généralement, sur la classe des lois de Y sachant X telles que $\mathbb{E}[(Y - \langle \beta^*, X\rangle)^2|X] \leqslant \sigma^2$, l'excès de risque minimax parmi les estimateurs *propres* est d'au moins $\sigma^2 d/(2n)$ (par le Théorème 1.8 de la Section 1.3.2 sur la régression linéaire). Cela tient au fait que le SMP quantifie mieux l'incertitude sur la valeur de Y sachant X que tout estimateur propre ; le SMP exploite aussi implicitement la "courbure" (mélangeabilité) globale de la perte logarithmique, qui peut être nettement supérieure à celle de la perte restreinte au modèle.

Remarque 1.6 (Cas bien spécifié). Il est possible de montrer que la première borne de (1.94) vaut précisément le double de l'excès de risque minimax dans le cas *bien spécifié* (voir la Section 8.1 ainsi que la fin de cette section). Ainsi, la performance du SMP dans le cas mal spécifié est proche de la performance optimale atteignable même dans le cas bien spécifié, quelle que soit la loi des variables X. Cela montre notamment que l'excès de risque minimax dans le cas mal spécifié vaut au plus deux fois celui du cas bien spécifié.

Remarque 1.7 (Lien avec le levier). La première borne de (1.94) peut s'exprimer en fonction de la loi du levier statistique $\widehat{\ell}_{n+1}$ (voir la Section 1.3.2) d'un point X_{n+1} parmi $X_1, \ldots, X_{n+1}$:

$$\mathbb{E}[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})] \leqslant -\mathbb{E}[\log(1 - \widehat{\ell}_{n+1})]\,.$$

[15] En raison du facteur 1/2 devant la perte quadratique, le risque de $\widehat{\beta}_n^{\mathrm{LS}}$ dans le cas bien spécifié où $Y|X \sim \mathcal{N}(\langle \beta^*, X\rangle, 1)$ est $\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)]/(2n)$.

Ainsi, le levier caractérise l'excès de risque minimax (et celui du SMP) pour le problème de l'estimation de densité conditionnelle, comme pour celui de la régression (Section 1.3). L'interprétation est la même : un levier $\widehat{\ell}_{n+1}$ "déséquilibré" (non concentré autour de d/n) signifie que le prédicteur optimal β^* est mal estimé dans certaines directions, et que $\widehat{\beta}_n^{\mathrm{LS}}$ est déterminé par un plus petit nombre d'observations, ce qui accroît sa variabilité.

SMP régularisé et bornes non uniformes. Nous nous intéressons à présent à une variante régularisée du SMP (nous renvoyons au Chapitre 7 pour une définition générale du SMP régularisé), obtenue en considérant la pénalité $\beta \mapsto \lambda\|\beta\|^2/2$ pour un certain $\lambda > 0$. Cet estimateur satisfait une borne d'excès de risque *non uniforme* sur la classe $\mathcal{F}$, dépendant de la norme $\|\beta\|$ du paramètre de comparaison. Cette procédure est utile dans le cas où des bornes uniformes satisfaisantes ne sont pas possibles, c'est-à-dire lorsque (1) la loi P_X ne satisfait pas l'Hypothèse 1.1 (Section 1.3.2) de régularité, garantissant un excès de risque en d/n pour le SMP non régularisé; ou (2) la dimension d est élevée et supérieure à n, c'est-à-dire dans un cadre *non paramétrique*.

Pour ce qui est du premier cas, le résultat suivant montre qu'il est possible d'obtenir une borne de risque en dimension finie sans hypothèse de petite boule (Hypothèse 1.1), au prix d'une (modeste) dépendance en la norme $\|\beta\|$.

Proposition 1.8 (Proposition 7.3, Chapitre 7)**.** *Pour tout $\lambda > 0$, le SMP pénalisé s'écrit $\widehat{g}_{\lambda,n}(\cdot|x) = \mathcal{N}(\widetilde{\mu}_\lambda(x), \widetilde{\sigma}^2_\lambda(x))$, où $\widetilde{\mu}_\lambda(x), \widetilde{\sigma}^2_\lambda(x)$ sont indiqués dans la Proposition 7.3. Supposons que $\mathbb{E}[Y^2] < +\infty$ et $\|X\| \leqslant R$ presque sûrement. Alors, pour tout $B > 0$, le choix $\lambda = d/(B^2(n+1))$ conduit à :*

$$\mathbb{E}\big[R(\widehat{g}_{\lambda,n})\big] - \inf_{\|\beta\|\leqslant B} R(\beta) \leqslant \frac{5d\log\big(2 + BR/\sqrt{d}\big)}{n+1} = O\left(\frac{d}{n}\log\left(2 + \frac{BR}{\sqrt{d}}\right)\right). \tag{1.95}$$

En particulier, lorsque $R = O(\sqrt{d})$ et $\|\beta\| = O(1)$ (ce qui correspond au cas où X est approximativement "isotrope", c'est-à-dire où Σ est bien conditionnée[16], et où la "force du signal" $\|\beta\|_\Sigma$ est bornée), cette borne est d'ordre $O(d/n)$. La borne (1.95) est un raffinement de la garantie obtenue par conversion online-to-batch. En effet, Kakade and Ng (2005) montrent que la stratégie de mélange Bayésien $(\widehat{f}_{\nu,t})_{1\leqslant t\leqslant n+1}$ sur $\mathcal{F}$, avec pour loi a priori sur β donnée par $\pi = \mathcal{N}(0, \nu^2)$, satisfait la borne de regret

$$\sum_{t=1}^{n+1} \ell(\widehat{f}_{\nu,t}, (X_t, Y_t)) - \inf_{\|\beta\|\leqslant B} \sum_{t=1}^{n+1} \ell(f_\beta, (X_t, Y_t)) \leqslant \frac{B^2}{2\nu^2} + \frac{d}{2}\log\left(1 + \frac{\nu^2 R^2(n+1)}{d}\right)$$

dès lors que $\|X_t\| \leqslant R$ pour tout t. Par conversion online-to-batch (Proposition 1.3), on en déduit que, sous les hypothèses de la Proposition 1.8, l'estimateur de mélange $\bar{f}_{\nu,n} := (n+1)^{-1}\sum_{t=1}^{n+1}\widehat{f}_{\nu,t}$ avec $\nu = B/\sqrt{d}$ satisfait

$$\mathbb{E}\big[R(\bar{f}_{\nu,n})\big] - \inf_{\|\beta\|\leqslant B} R(\beta) \leqslant \frac{d(1 + \log(1 + B^2R^2(n+1)/d^2))}{2(n+1)} = O\left(\frac{d}{n}\log\left(2 + \frac{B^2R^2}{d}\frac{n}{d}\right)\right).$$

[16] Au sens où $C^{-1}I_d \preccurlyeq \Sigma \preccurlyeq CI_d$, avec $C = O(1)$. En pratique, le bon conditionnement peut s'obtenir en normalisant les X_i par la matrice de covariance empirique (éventuellement calculée sur une fraction séparée du jeu de données), c'est-à-dire par une étape de *pré-conditionnement*.

Cette borne est du même type que celle (1.95) du SMP pénalisé, mais avec un facteur additionnel en $\log(n/d)$. Par exemple, si $R = O(\sqrt{d})$ et $B = O(1)$, la borne précédente est en $O(d\log(n/d)/n)$, correspondant à une vitesse asymptotique en $O(d\log(n)/n)$ pour $n \gg d$.

Par ailleurs, il est également possible d'obtenir des bornes pour le SMP régularisé lorsque $d \geqslant n$. En effet, le Théorème 7.5 du Chapitre 7 montre que, sous les hypothèses de la Proposition 1.8, le SMP pénalisé $\widetilde{f}_{\lambda,n}$ satisfait, pour tout $\beta \in \mathbf{R}^d$,

$$\mathbb{E}[R(\widehat{g}_{\lambda,n})] - R(\beta) \leqslant 1.25 \cdot \frac{\mathrm{Tr}[(\Sigma + \lambda I_d)^{-1}\Sigma]}{n+1} + \frac{\lambda\|\beta\|^2}{2}\,.$$

Dans cette borne, la dimension d est remplacée par la quantité $\mathrm{Tr}((\Sigma + \lambda I_d)^{-1}\Sigma)$, correspondant aux *degrés de liberté* de l'estimateur Ridge (Wahba, 1990; Friedman et al., 2001; Wasserman, 2006). À nouveau, cette borne est valable uniformément sur les lois P telles que $\mathbb{E}[Y^2] < +\infty$.

Cas bien spécifié : sous-optimalité de l'EMV en grande dimension (Section 8.1). Considérons à présent le cas bien spécifié, où $Y|X \sim \mathcal{N}(\langle\beta^*, X\rangle, 1)$ pour un certain $\beta^* \in \mathbf{R}^d$, c'est-à-dire où $P \in \mathcal{P} := \mathcal{P}_{\mathrm{Gauss}}(P_X, 1)$ (cf. (1.55)). Supposons également P_X non dégénérée (Définition 6.1). Dans ce cas, comme souligné plus haut ainsi que dans la Section 1.3, l'excès de risque minimax parmi les estimateurs *propres* vaut

$$\frac{1}{2n}\mathbb{E}\big[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\big]\,,$$

et est atteint par l'EMV $f_{\widehat{\beta}_n^{\mathrm{LS}}}$. En outre, comme mentionné plus haut (voir le Théorème 8.2 de la Section 8.1), l'excès de risque minimax (sans restriction, donc en autorisant les estimateurs impropres) est

$$\frac{1}{2}\mathbb{E}\big[\log\big(1 + \langle(n\widehat{\Sigma}_n)^{-1}X, X\rangle\big)\big]\,,$$

atteint par l'estimateur impropre $\widehat{g}_n(\cdot|x) := \mathcal{N}(\langle\widehat{\beta}_n^{\mathrm{LS}}, \cdot\rangle, 1 + \langle(n\widehat{\Sigma}_n)^{-1}x, x\rangle)$. Le premier risque est supérieur à $d/(2(n-d+1))$ pour toute loi P_X, et vaut $d/(2(n-d-1))$ lorsque $X \sim \mathcal{N}(0,\Sigma)$. Par un argument similaire, on montre que le second risque est supérieur à $-\log(1-d/(n+1))/2$ pour toute loi P_X, et inférieur à $-\log(1-d/(n-1))/2$ si $P_X = \mathcal{N}(0,\Sigma)$.

Considérons maintenant le régime asymptotique de *grande dimension*, où $d, n \to \infty$ avec $d/n \to \gamma \in (0,1)$. Dans ce cadre, la performance optimale atteignable par un estimateur propre est de $\gamma/(2(1-\gamma))$ (avec égalité dans le cas où $P_X \sim \mathcal{N}(0,\Sigma)$). Cependant, le risque optimal atteint par l'estimateur impropre $\widehat{g}_n$ est (dans le cas Gaussien) de

$$-\frac{1}{2}\log(1-\gamma) = \frac{1}{2}\log\Big(1 + \frac{\gamma}{1-\gamma}\Big) < \frac{\gamma}{2(1-\gamma)}\,.$$

Ainsi, même dans le cas bien spécifié où P appartient à la classe $\mathcal{F}$, les estimateurs propres (restreints à $\mathcal{F}$) sont sous-optimaux lorsque la dimension d est relativement élevée (de l'ordre de n). Ceci contraste avec le cadre asymptotique classique (évoqué en Section 1.4.2), où d est fixé et $n \to \infty$, pour lequel l'EMV est asymptotiquement optimal. Intuitivement, le prédicteur $\widehat{g}_n$ admet un meilleur risque que l'EMV car il quantifie mieux l'incertitude sur la loi P (et donc sur ses réalisations futures) que l'EMV.

1.4.6 Régression logistique

Nous considérons à présent un autre problème classique d'estimation de densité conditionnelle, à savoir la *régression logistique* (Berkson, 1944; McCullagh and Nelder, 1989; van der Vaart, 1998). Ce problème correspond au cas où la réponse Y est binaire, ou plus généralement catégorielle.

Définitions et notations. Ici, $\mathcal{Y} = \{-1, 1\}$ est binaire, et $\mu = \delta_1 + \delta_{-1}$ est la mesure de comptage sur $\mathcal{Y}$. De plus, $\mathcal{X} = \mathbf{R}^d$, et $\mathcal{F}$ est le *modèle logistique* $\{f_\beta : \beta \in \mathbf{R}^d\}$, où

$$f_\beta(1|x) = 1 - f_\beta(-1|x) = \sigma(\langle \beta, x \rangle) \tag{1.96}$$

pour tous $\beta, x \in \mathbf{R}^d$, avec $\sigma : \mathbf{R} \to (0, 1)$ la *fonction sigmoïde* $\sigma(u) = e^u/(1 + e^u)$. Puisque $\sigma(-u) = 1 - \sigma(u)$, on a $f_\beta(y|x) = \sigma(y\langle \beta, x \rangle)$ pour tout $y \in \{-1, 1\}$. Ainsi, en notant $\ell(u) = \log(1 + e^u)$ pour $u \in \mathbf{R}$, on a pour tous $\beta, x \in \mathbf{R}^d$ et $y \in \{-1, 1\}$,

$$\ell(f_\beta, (x, y)) = \log(1 + e^{-y\langle \beta, x \rangle}) = \ell(-y\langle \beta, x \rangle) = \ell(\langle \beta, z \rangle)$$

où $z := -yx$. L'EMV est par définition, en notant $Z_i = -Y_i X_i$,

$$\widehat{\beta}_n^{\text{EMV}} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \widehat{R}_n(\beta) := \frac{1}{n} \sum_{i=1}^n \ell(\langle \beta, Z_i \rangle) \right\},$$

dès lors que ce minimiseur existe. Tout d'abord, le risque empirique $\widehat{R}_n(\beta)$ est convexe en β, car $u \mapsto \ell(u)$ est convexe. Considérons alors les cas de figure suivants :

- Le jeu de données est *(linéairement) séparé (au sens strict)*, s'il existe $\theta \in \mathbf{R}^d$ tel que $\langle \theta, Z_i \rangle < 0$ pour $i = 1, \dots, n$; cela revient à dire que $\langle \theta, X_i \rangle < 0$ si $Y_i = 1$, et $\langle \theta, X_i \rangle > 0$ si $Y_i = -1$, c'est-à-dire que l'hyperplan $\{x \in \mathbf{R}^d : \langle \theta, x \rangle = 0\}$ sépare les classes $\{X_i : Y_i = 1\}$ et $\{X_i : Y_i = -1\}$. Dans ce cas, $\widehat{R}_n(t\theta) \to 0$ lorsque $t \to +\infty$; comme par ailleurs $\widehat{R}_n > 0$, $\widehat{R}_n$ n'admet pas de minimum sur $\mathbf{R}^d$. On peut alors étendre la classe $\mathcal{F}$ en y ajoutant les densités conditionnelles f^θ ($\theta \in \mathbf{R}^d$, $\|\theta\| = 1$), où $f^\theta(y|x)$ vaut 1 si $y\langle \theta, x \rangle > 0$, 0 si $y\langle \theta, x \rangle < 0$ et $1/2$ si $y\langle \theta, x \rangle = 0$. Les minimiseurs du risque empirique sont alors donnés par les hyperplans séparateurs ; un tel hyperplan n'est pas unique.

- Le jeu de donné est dit *(strictement) non séparé* si pour tout $\theta \in \mathbf{R}^d \setminus \{0\}$, il existe un i tel que $\langle \theta, Z_i \rangle > 0$. Dans ce cas, la fonction $\beta \mapsto \widehat{R}_n(\beta)$ diverge lorsque $\|\beta\| \to \infty$; par continuité, elle admet donc un minimum dans $\mathbf{R}^d$. Enfin, les Z_i engendrent linéairement $\mathbf{R}^d$ (sinon il existerait un $\theta \neq 0$ tel que $\langle \theta, Z_i \rangle = 0$ pour tout i), ce qui implique que $\widehat{R}_n$ est strictement convexe et donc que son minimiseur $\widehat{\beta}_n^{\text{EMV}}$ est unique.

En revanche, pour tout $\lambda > 0$, l'estimateur pénalisé (de type Ridge)

$$\widehat{\beta}_{\lambda,n} = \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\langle \beta, Z_i \rangle) + \frac{\lambda}{2} \|\beta\|^2 \right\} \tag{1.97}$$

est bien défini de manière unique, par λ-forte convexité du risque pénalisé.

Garanties pour la régression logistique. Il découle de résultats classiques sur l'EMV que, sous des hypothèses modestes sur la loi jointe de (X, Y) (non séparation, moments contrôlés de $\|X\|$), $\widehat{\beta}_n^{\mathrm{EMV}}$ est consistant : $\widehat{\beta}_n^{\mathrm{EMV}} \to \beta^* = \arg\min_{\beta \in \mathbf{R}^d} R(\beta)$ en probabilité, et asymptotiquement normal (van der Vaart, 1998), au sens où $\sqrt{n}(\widehat{\beta}_n^{\mathrm{EMV}} - \beta^*)$ converge en loi vers $\mathcal{N}(0, H^{-1}GH^{-1})$, avec les notations de la Section 1.4.2. La constante $d_{\mathrm{eff}} = \mathrm{Tr}(H^{-1/2}GH^{-1/2})$, qui caractérise l'excès de risque asymptotique de l'EMV (voir la Section 1.4.2) peut être significativement plus élevée que d. En effet, si $\|X\| \leqslant R$ presque sûrement et $\|\beta^*\| \leqslant B$, d_{eff} peut être aussi élevée que de^{BR} (Bach and Moulines, 2013). De plus, l'excès de risque en espérance de l'EMV (avec choix arbitraire de l'hyperplan séparateur dans le cas séparé) est typiquement infini, par exemple lorsque $\mathbb{P}(Y = 1|X) \in (0, 1)$ presque sûrement : en effet, dans ce cas le jeu de données est séparé avec probabilité positive, l'EMV est alors un hyperplan séparateur qui admet une erreur infinie dans la région de $\mathbf{R}^d \times \{-1, 1\}$ à laquelle il attribue une probabilité nulle. Plus généralement, l'EMV tend à produire des prédictions trop confiantes, c'est-à-dire proches de 0 ou 1 (Sur and Candès, 2019).

Supposons dans ce qui suit que $\|X\| \leqslant R$ presque sûrement, et considérons des bornes d'excès de risque par rapport à la classe $\mathcal{F}_B := \{f_\beta : \beta \in \mathbf{R}^d, \|\beta\| \leqslant B\}$. Ce problème peut être envisagé comme un problème d'optimisation stochastique (Section 1.1.5). Tout d'abord, la perte $\ell(\beta, Z)$ est convexe en β et R-Lipschitz car $\|Z\| \leqslant R$. Cela implique qu'un excès de risque en $O(BR/\sqrt{n})$ est atteignable, de plusieurs façons :

- par minimisation du risque empirique sur $\mathcal{F}_B$ (par un argument de convergence uniforme de processus empiriques, voir la Section 1.1.4, et une inégalité de contraction sur les complexités de Rademacher, Ledoux and Talagrand, 2013) ;
- par l'estimateur pénalisé $\widehat{\beta}_{\lambda,n}$ (1.97) avec $\lambda = R/(B\sqrt{n})$ (par le Corollaire 1.1) ;
- par descente de gradient stochastique projetée sur $\mathcal{F}_B$ et moyennée (par la Proposition 1.2) ;
- par agrégation à poids exponentiels sur β (suivi d'une moyenne/conversion online to batch), avec une loi a priori sur β uniforme sur la boule $\{\beta \in \mathbf{R}^d : \|\beta\| \leqslant B\}$, par la Proposition 1.5 (dans ce cas, la borne obtenue est en fait de $BR\sqrt{(\log n)/n}$).

De plus, on vérifie que la fonction de perte $\ell(\beta, Z)$ est également e^{-BR}-exp-concave sur la boule de norme B. Ceci implique qu'une vitesse en $O(de^{BR}\log(n)/n)$ est possible, par conversion online-to-batch à partir des poids exponentiels avec $\eta = e^{-BR}$, ou de l'algorithme *Online Newton Step* (Hazan et al., 2007; Mahdavi et al., 2015). Une vitesse en $O(de^{BR}/n)$ peut également être obtenue par minimisation du risque empirique, pénalisée (Koren and Levy, 2015) ou non (Gonen and Shalev-Shwartz, 2018), ou d'autres procédures (Mehta, 2017).

De ce qui précède, il ressort qu'il est possible d'obtenir une borne d'excès de risque en

$$O\Big(\min\Big(\frac{BR}{\sqrt{n}}, \frac{de^{BR}}{n}\Big)\Big). \tag{1.98}$$

La première vitesse ci-dessus est une vitesse lente ; la seconde est une vitesse rapide (il s'agit de la vitesse asymptotique de l'EMV lorsque d est fixe et $n \to \infty$, pour une loi P choisie telle que $d_{\mathrm{eff}} \asymp de^{BR}$), mais avec une dépendance exponentielle prohibitive en BR (qui est typiquement d'ordre $\sqrt{d}$). Il s'avère en fait, par un résultat de Hazan et al. (2014), qu'il n'est pas possible d'améliorer la borne (1.98) pour un estimateur *propre* sans hypothèse supplémentaire sur P :

pour tout estimateur $\widehat{\beta}_n$, il existe une loi P pour laquelle l'excès de risque $\mathbb{E}[\mathcal{E}(f_{\widehat{\beta}_n})]$ par rapport à $\mathcal{F}_B$ est d'au moins (1.98). Pour des valeurs typiques de BR, d, n, on a $n = O(e^{BR})$ et la vitesse lente dans (1.98) est la meilleure. Ainsi, pour $B = \|\beta^*\| = O(1)$ et $R = O(\sqrt{d})$ (ce qui correspond au cas de "dimension finie" et "bien conditionné", voir la Section 1.4.5), cette vitesse est de $O(\sqrt{d/n})$.

Afin de contourner cette borne inférieure dans le pire des cas, il est naturel de faire des hypothèses supplémentaires, afin d'obtenir des bornes avec une dépendance explicite en certaines quantités dépendant de la loi P. Dans le pire des cas, ces bornes mènent inévitablement à la vitesse lente (1.98), mais pour certaines lois plus favorables (telles que $d_{\text{eff}} \ll de^{BR}$) la borne obtenue peut être bien meilleure. Pour mener à bien ce type d'analyse, la seule convexité (qui conduit à une vitesse lente) ou la courbure *globale* (qui induit une dépendance exponentielle en BR) ne sont pas suffisamment précises. Intuitivement, la difficulté du problème est déterminée, au moins asymptotiquement, par les propriétés *locales* de la fonction de perte $\ell(\beta, z)$ et du risque $R(\beta)$, pour β proche de l'optimum β^*. En particulier, pour $\beta \approx \beta^*$, le risque est approximativement quadratique : $R(\beta) - R(\beta^*) \approx \frac{1}{2}\|\beta - \beta^*\|_H^2$, où $H = \nabla^2 R(\beta^*)$. Afin d'obtenir des garanties non-asymptotiques fines en termes de quantités locales telles que H et d_{eff}, un contrôle global de la qualité de l'approximation quadratique locale de R est nécessaire.

Pour cela, une propriété utile de la fonction de perte logistique $\ell(\beta, z)$ est la *(pseudo-)auto-concordance*, introduite par Bach (2010) et exploitée par Bach (2014); Bach and Moulines (2013); Ostrovskii and Bach (2018); Marteau-Ferey et al. (2019) afin d'analyser la régression logistique (ainsi que d'autres problèmes d'optimisation stochastique) de manière non asymptotique. La notion d'*auto-concordance* (Nesterov and Nemirovskii, 1994), c'est-à-dire une majoration de la dérivée tierce d'une fonction en fonction de la puissance $3/2$ de sa dérivée seconde, est utilisée dans le cadre de l'analyse d'algorithmes d'optimisation du second ordre tels que la méthode de Newton (Nesterov and Nemirovskii, 1994; Boyd and Vandenberghe, 2004). La notion de *(pseudo-)auto-concordance* introduite par Bach (2010) correspond également à un contrôle de la dérivée troisième en fonction de la dérivée seconde, mais sans l'exposant $3/2$. Dans le cas de la perte logistique, cette propriété s'écrit, pour tout $u \in \mathbf{R}$,

$$|\ell'''(u)| = |\sigma(u)\big(1 - \sigma(u)\big)\big(1 - 2\sigma(u)\big)| \leqslant \sigma(u)\big(1 - \sigma(u)\big) = \ell''(u)\,.$$

À un haut niveau, l'auto-concordance est préférable à une borne uniforme sur la dérivée troisième, car elle permet de contrôler les variations *relatives* de la Hessienne

$$H(\beta) := \nabla^2 R(\beta) = \mathbb{E}[\ell''(\langle \beta, Z\rangle) ZZ^\top]\,.$$

En utilisant la propriété d'auto-concordance, Bach (2010) obtient une vitesse rapide pour la régression logistique avec pénalisation Ridge (1.97), dans le cas d'un design déterministe et d'un modèle logistique bien spécifié.

Dans le cas général mal spécifié et avec un design déterministe, supposons l'existence de $\beta^* = \arg\min_{\beta \in \mathbf{R}^d} R(\beta)$, et considérons les matrices suivantes :

$$\begin{aligned}\Sigma &= \mathbb{E}[XX^\top] = \mathbb{E}[ZZ^\top] \\ G(\beta) &= \mathbb{E}[\nabla \ell(\beta, Z) \nabla \ell(\beta, Z)^\top] = \mathbb{E}[\sigma(\langle \beta, Z\rangle)^2 ZZ^\top]\,,\end{aligned}$$

ainsi que $G = G(\beta^*)$, $H = H(\beta^*)$. L'excès de risque asymptotique de l'EMV est caractérisé par la *dimension effective* $d_{\text{eff}} = \operatorname{Tr}(H^{-1/2} G H^{-1/2})$ (voir la Section 1.4.2). Une autre quantité importante (Bach and Moulines, 2013) est la norme d'opérateur $\rho = \lambda_{\max}(H^{-1/2}\Sigma H^{-1/2})$,

c'est-à-dire la plus petite constante telle que $\Sigma \preccurlyeq \rho H$. Puisque $\ell'' = \sigma(1-\sigma) \leqslant 1/4$, on a nécessairement $\rho \geqslant 4$. De plus, si $B = \|\beta^*\|$ et $\|X\| \leqslant R$ presque sûrement, alors $\ell''(\langle \beta^*, X\rangle) \gtrsim e^{-BR}$, et donc $\rho \lesssim e^{BR}$; dans le pire des cas, $\rho \asymp e^{BR}$, mais ρ peut être bien plus faible en pratique (Bach and Moulines, 2013; Ostrovskii and Bach, 2018).

Bach (2014) considère l'algorithme de descente de gradient stochastique, avec un pas de $C/(R^2\sqrt{n})$ et avec moyenne des itérés, et établit une borne d'excès de risque de

$$\mathbb{E}\big[R(\widehat{\beta}_n)\big] - R(\beta^*) \lesssim \frac{R^2}{\mu n}\big(B^4R^4 + 1\big)\,, \tag{1.99}$$

avec $B = \|\beta^*\|$, et où μ désigne la plus petite valeur propre de H, c'est-à-dire la constante de forte convexité *locale* du risque au voisinage de β^*.

La borne (1.98) est sensible au conditionnement des données, c'est-à-dire au caractère isotrope (ou non) de la matrice de covariance Σ. Pour le voir, définissons $\lambda = \lambda_{\min}(\Sigma)$ la plus petite valeur propre de Σ. Alors

$$\lambda = \lambda_{\min}(\Sigma) \leqslant \frac{1}{d}\mathrm{Tr}(\Sigma) = \frac{\mathbb{E}[\|X\|^2]}{d} \leqslant \frac{R^2}{d}\,,$$

de sorte que

$$\frac{R^2}{\lambda} \geqslant d\,.$$

Par ailleurs, l'inégalité $H \preccurlyeq \Sigma/4$ implique que $\mu \leqslant \lambda/4$, et donc $R^2/\mu \geqslant 4R^2/\lambda \geqslant 4d$. De plus, pour les problèmes "mal conditionnés" où X est anisotrope, c'est-à-dire ici $\mathrm{Tr}(\Sigma)/\lambda_{\min}(\Sigma) \gg d$, on a $R^2/\mu \geqslant 4R^2/\lambda \gg d$. Enfin, notons que $\mu \geqslant \lambda/\rho$, et qu'il est possible d'avoir égalité. Dans le cas où $\mu \asymp \lambda/\rho$, on obtient $R^2/\mu \asymp \rho R^2/\lambda \gtrsim \rho d$, avec égalité (à une constante près) dans le cas bien conditionné. Ainsi, du fait de sa dépendance implicite en $\lambda = \lambda_{\min}(\Sigma)$, la borne (1.98) est principalement adaptée au cas paramétrique (avec $d \ll n$) bien conditionné. Dans ce cas, avec les ordres de grandeur $R^2/\mu = O(\rho d)$ et $BR = O(\sqrt{d})$, la borne (1.98) donne la vitesse $O(\rho \cdot d^3/n)$, avec une dépendance sous-optimale en d. Toutefois, en choisissant un pas différent (dépendant de B), il est possible de remplacer le terme $O(B^4R^4 + 1)$ par $O(B^2R^2)$, ce qui donne dans ce cas une vitesse en $O(\rho \cdot d^2/n)$.

À partir d'un algorithme et d'une analyse raffinées, Bach and Moulines (2013) montrent qu'il est possible d'éviter la dépendance en λ dans une borne de risque. Plus précisément, Bach and Moulines (2013) proposent un algorithme d'optimisation stochastique en deux étapes qui admet la garantie suivante : pour $n \gtrsim B^4R^4$,

$$\mathbb{E}[R(\widehat{\beta}_n)] - R(\beta^*) \lesssim \frac{\kappa^{3/2}\rho^3 d}{n}(B^4R^4 + 1)\,, \tag{1.100}$$

où κ est une borne sur la kurtosis des marginales unidimensionnelles des variables $[\ell''(\langle \beta, Z\rangle)]^{1/2}Z$ pour $\beta \in \mathbf{R}^d$. Cette borne correspond essentiellement à la précédente (1.99) dans le cas bien conditionné, mais cette fois-ci sans dépendance en λ lorsque $\lambda \ll R^2/d$. En particulier, en dimension finie avec $BR = O(\sqrt{d})$, cette borne mène à un excès de risque en $O(\rho^3 d^3/n)$ pour $n \gtrsim d^2$. Plus récemment, au moyen d'une analyse fine exploitant également l'auto-concordance, Ostrovskii and Bach (2018) ont obtenu des garanties améliorées pour l'estimateur du maximum de vraisemblance $\widehat{\beta}_n^{\mathrm{EMV}}$. En supposant que les vecteurs aléatoires "décorrélés" $\Sigma^{-1/2}Z$, $G^{-1/2}\ell'(\langle \beta^*, Z\rangle)Z$ et $H^{-1/2}[\ell''(\langle \beta, Z\rangle)]^{1/2}Z$ (pour β proche de β^*) sont

sous-Gaussiens, Ostrovskii and Bach (2018) montrent que, pour $n \gtrsim \max(\rho d_{\text{eff}}, d \log d)$, l'excès de risque est borné par

$$R(\widehat{\beta}_n^{\text{EMV}}) - R(\beta^*) \lesssim \frac{d_{\text{eff}}}{n}$$

avec forte probabilité, ce qui correspond à une variante non asymptotique du risque asymptotique de $\widehat{\beta}_n^{\text{EMV}}$. Cette borne améliore la précédente (1.100), tant du point de vue du risque que de la valeur de n requise. Notons que cette borne dépend des normes sous-Gaussiennes des vecteurs cités plus haut, qui dépendent elles-mêmes implicitement de la loi P, en particulier de la loi P_X, de β^* et de la loi de Y sachant X. Dans le cas où $X \sim \mathcal{N}(0, \Sigma)$ (par invariance de l'EMV par transformation linéaire, il est possible de supposer $\Sigma = I_d$), Ostrovskii and Bach (2018) montrent que ces normes peuvent être bornées en fonction de $B = \|\beta^*\|_\Sigma = \|\beta^*\|$, avec dans le cas de $G^{-1/2}\ell'(\langle \beta^*, Z\rangle)Z$ l'hypothèse supplémentaire que le modèle est bien spécifié. Enfin, Marteau-Ferey et al. (2019) considèrent le cas *non paramétrique* bien spécifié, et obtiennent des bornes non asymptotiques où la dimension d est remplacée par les degrés de libertés $\operatorname{Tr}[(H + \lambda I_d)^{-1} H]$ de H, et qui dépendent de la décroissance des coefficients de β^* dans une base de vecteurs propres de H. Ces bornes correspondent aux vitesses non paramétriques obtenues par l'estimateur Ridge dans le cas des moindres carrés (Caponnetto and De Vito, 2007).

Rappelons que dans le cas général où l'on suppose seulement que $\|X\| \leqslant R$, les bornes précédentes exhibent toutes une dépendance inévitable en e^{BR} dans le pire des cas, en raison de la borne inférieure (1.98) de Hazan et al. (2014) pour les estimateurs propres. Il est cependant possible de contourner cette borne inférieure sans hypothèse supplémentaire, en ayant recours à des prédicteurs *impropres* (Foster et al., 2018), tels que les estimateurs par mélange Bayésien (Section 1.4.3). Plus précisément, Kakade and Ng (2005) et Foster et al. (2018) ont montré que la stratégie de prédiction en ligne par mélange Bayésien, avec pour lois a priori respectives $\mathcal{N}(0, B^2/d)$ et la loi uniforme sur la boule de norme B, admettent une borne de regret en

$$O\Big(d \log\Big(2 + \frac{BRn}{d}\Big)\Big)$$

dès lors que $\|X_t\| \leqslant R$ pour tout t. Par convergence online-to-batch (Proposition 1.3), Foster et al. (2018) en déduit une borne d'excès de risque de

$$\mathbb{E}[R(\bar{f}_n)] - \inf_{\|\beta\| \leqslant B} R(\beta) = O\Big(\frac{d}{n} \log\Big(2 + \frac{BRn}{d}\Big)\Big) \tag{1.101}$$

pour la moyenne $\bar{f}_n$ des itérés. Signalons au passage que Foster et al. (2018) proposent également un autre estimateur $\widehat{f}_{n,\delta}$ pour lequel ils énoncent une borne d'excès de risque avec forte probabilité $1 - \delta$ (Théorème 10). Cette borne est néanmoins erronée : en effet, la preuve applique l'inégalité de Markov sur l'excès de risque d'un estimateur impropre intermédiaire ; cependant, cet estimateur étant impropre (en dehors de la classe $\mathcal{F}_B$), son excès de risque par rapport à la classe $\mathcal{F}_B$ peut prendre des valeurs négatives, de sorte qu'il n'est pas possible de lui appliquer l'inégalité de Markov.

La borne (1.101) a ceci de remarquable qu'elle ne dépend pas des constantes ρ, d_{eff} et n'admet donc pas de dépendance exponentielle en la norme BR. L'estimateur $\bar{f}_n$ contourne donc la borne inférieure (1.98) pour les estimateurs propres dans le pire des cas. Lorsque $BR = O(\sqrt{d})$, cette borne fournit une vitesse en $O(d \log(n)/n)$, optimale au facteur $\log n$ près.

Le principal inconvénient de cette approche est sa complexité algorithmique, soulignée par Foster et al. (2018). En effet, $\bar{f}_n$ est la moyenne des estimateurs $\widetilde{f}_t$, $1 \leqslant t \leqslant n+1$, où $\widetilde{f}_t$ est le postérieur prédictif Bayésiens calculé à partir des $t-1$ premières observations, soit

$$\widetilde{f}_t(y|x) := \int_{\mathbf{R}^d} \sigma(\langle \beta, x \rangle) \widehat{\pi}_t(\mathrm{d}\beta)\,,$$

où $\widehat{\pi}_t$ est le postérieur $\pi(\cdot|Z_1, \ldots, Z_{t-1})$, dont la densité par rapport à π est donnée par

$$\frac{\prod_{s=1}^{t-1} f_\beta(Y_s|X_s)}{\int_{\mathbf{R}^d} \prod_{s=1}^{t-1} f_\beta(Y_s|X_s)\pi(\mathrm{d}\beta)} = \frac{\prod_{s=1}^{t-1} \sigma(-\langle \beta, Z_s \rangle)}{\int_{\mathbf{R}^d} \prod_{s=1}^{t-1} \sigma(-\langle \beta, Z_s \rangle)\pi(\mathrm{d}\beta)}\,. \tag{1.102}$$

Le dénominateur de (1.102) n'admettant pas d'expression explicite, il est nécessaire de recourir à des techniques de calcul approché de postérieurs. Par exemple, Foster et al. (2018) montrent qu'il est possible de calculer approximativement $\bar{f}_n$ en un temps de $O(B^6 \max(d, BRn)^{12}/\varepsilon^{12})$, où ε désigne le niveau de précision recherché. Bien que ce temps soit polynomial en n, d, cette procédure est trop coûteuse pour être utilisable en pratique. Un problème ouvert posé par Foster et al. (2018) demande s'il est possible d'obtenir un algorithme moins coûteux satisfaisant une vitesse rapide de regret ou d'excès de risque. Comme nous le montrons dans la section suivante (ainsi que dans le Chapitre 7), le SMP apporte une réponse positive (partielle) à cette question.

1.4.7 Application du SMP à la régression logistique (Chapitre 7)

Nous appliquons maintenant le SMP à la régression logistique. Commençons pour simplifier par considérer le SMP non pénalisé. Ses prédictions sont données par

$$\widehat{g}_n^{\mathrm{SMP}}(y|x) = \frac{f_{\widehat{\beta}_n^{(x,y)}}(y|x)}{f_{\widehat{\beta}_n^{(x,1)}}(1|x) + f_{\widehat{\beta}_n^{(x,-1)}}(-1|x)} = \frac{\sigma(\langle \widehat{\beta}_n^{(x,y)}, yx \rangle)}{\sigma(\langle \widehat{\beta}_n^{(x,1)}, x \rangle) + \sigma(\langle \widehat{\beta}_n^{(x,-1)}, -x \rangle)}\,, \tag{1.103}$$

pour tout $(x, y) \in \mathbf{R}^d \times \{-1, 1\}$, où $\widehat{\beta}_n^{(x,y)}$ désigne l'EMV obtenu en rajoutant (x, y) à l'échantillon. De plus, sa borne d'excès de risque (1.93) s'écrit :

$$\mathbb{E}\big[\mathcal{E}(\widehat{g}_n^{\mathrm{SMP}})\big] \leqslant \mathbb{E}_{Z_1^n, Z}\big[\sigma(\langle \widehat{\beta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\beta}_n^{Z}, Z \rangle)\big]\,. \tag{1.104}$$

Intuitivement, la borne d'excès de risque du SMP est plus fine qu'une borne en termes de stabilité de la perte, s'appliquant par exemple à l'EMV (voir la Section 1.1.6). En effet, en notant $v = \langle \widehat{\beta}_n^{-Z}, Z \rangle$ et $u = \langle \widehat{\beta}_n^{Z}, Z \rangle$, on a pour $u \simeq v \gg 1$, $\ell(\langle \widehat{\beta}_n^{-Z}, Z \rangle) - \ell(\langle \widehat{\beta}_n^{Z}, Z \rangle) = \ell(v) - \ell(u) \simeq v - u$, tandis que $\sigma(\langle \widehat{\beta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\beta}_n^{Z}, Z \rangle) \simeq \sigma'(u)(v-u) \simeq e^{-u}(v-u)$. Ainsi, le terme (1.104) peut être exponentiellement plus petit qu'un terme de stabilité de la perte. Ce gain permet d'éviter une dépendance exponentielle en BR dans le pire des cas.

D'un point de vue qualitatif, les prédictions du SMP sont moins "confiantes" (proches de 0 ou 1) que celles de l'EMV. En particulier, elles appartiennent toujours à $(0, 1)$, et sont toujours définies de manière unique même dans le cas séparé (si le jeu de données augmenté de (x, y) est séparé, alors $\widehat{f}_n^{(x,y)}(y|x) = 1$, quel que soit le choix de l'hyperplan séparateur). En particulier, si le point x est tel que la quantité

$$\sigma(\langle \widehat{\beta}_n^{(x,1)}, x \rangle) + \sigma(\langle \widehat{\beta}_n^{(x,-1)}, -x \rangle) = 1 + \sigma(\langle \widehat{\beta}_n^{(x,1)}, x \rangle) - \sigma(\langle \widehat{\beta}_n^{(x,-1)}, x \rangle)\,,$$

qui peut être vue comme un analogue du levier dans le cas logistique (en tant que mesure de l'influence de l'étiquette y du point x sur la prédiction associée), est élevée, alors la prédiction correspondante du SMP est "incertaine" (proche de $1/2$). Dans le cas extrême où le jeu de données est séparé, et où x est tel que le jeu de données reste séparé en y ajoutant $(x,1)$ ou $(x,-1)$, alors $\sigma(\langle\widehat{\beta}_n^{(x,1)},x\rangle)=\sigma(\langle\widehat{\beta}_n^{(x,-1)},-x\rangle)=1$, de sorte que $\widehat{g}_n^{\mathrm{SMP}}(1|x)=1/2$. À l'inverse, l'EMV $f_{\widehat{\beta}_n^{\mathrm{EMV}}}$ satisfait $f_{\widehat{\beta}_n^{\mathrm{EMV}}}(1|x)=0$ ou $f_{\widehat{\beta}_n^{\mathrm{EMV}}}(1|x)=1$, en fonction du choix de l'hyperplan séparateur, les deux cas étant possibles. Ainsi, le SMP corrige l'une des défaillances de l'EMV, qui est de produire des prédictions trop confiantes.

Afin d'obtenir des garanties non asymptotiques, nous considérons le SMP non pénalisé.

Théorème 1.14 (Théorème 7.6, Chapitre 7)**.** *Pour le modèle logistique* $\mathcal{F}=\{f_\beta:\beta\in\mathbf{R}^d\}$ (1.96)*, le SMP avec pénalité* $\beta\mapsto\lambda\|\beta\|^2/2$ *(où* $\lambda>0$*) s'écrit*

$$\widetilde{f}_{\lambda,n}(y|x)=\frac{\sigma(y\langle\widehat{\beta}_{\lambda,n}^{(x,y)},x\rangle)\,e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(x,y)}\|^2/2}}{\sigma(\langle\widehat{\beta}_{\lambda,n}^{(x,1)},x\rangle)\,e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(x,1)}\|^2/2}+\sigma(-\langle\widehat{\beta}_{\lambda,n}^{(x,-1)},x\rangle)\,e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(x,-1)}\|^2/2}} \tag{1.105}$$

où l'on note $\widehat{\beta}_\lambda^{(x,y)}=\widehat{\beta}_\lambda^{(-yx)}$*, avec pour* $z\in\mathbf{R}^d$ *:*

$$\widehat{\beta}_{\lambda,n}^{(z)}:=\operatorname*{arg\,min}_{\beta\in\mathbf{R}^d}\left\{\frac{1}{n+1}\left[\sum_{i=1}^n\ell(\langle\beta,Z_i\rangle)+\ell(\langle\beta,z\rangle)\right]+\frac{\lambda}{2}\|\beta\|^2\right\}.$$

De plus, pour toute loi jointe de (X,Y) *telle que* $\|X\|\leqslant R$ *presque sûrement, l'estimateur* (1.105) *satisfait, pour tout* $\lambda\geqslant 2R^2/(n+1)$*,*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big]\leqslant R(\beta)+3\cdot\frac{\mathrm{Tr}[(\Sigma+4\lambda I_d)^{-1}\Sigma]}{n}+\frac{\lambda}{2}\|\beta\|^2\,. \tag{1.106}$$

Cette borne est utile dans le cas non paramétrique où $d\gg n$. Il s'agit d'une vitesse rapide, similaire à celles obtenues par Bach (2010); Marteau-Ferey et al. (2019) dans le cas bien spécifié. Ces dernières admettent de plus un terme de variance exprimé en fonction de $H(\beta^*)$ (qui est toujours inférieur à $\Sigma/4$), et un terme de biais plus général que $\lambda\|\beta^*\|^2$, exprimé en fonction de différentes normes. En revanche, ces résultats nécessitent des hypothèses plus fortes, et exhibent une dépendance exponentielle en BR dans le cas général. D'un point de vue technique, notre analyse utilise également la notion d'auto-concordance de Bach (2010); Bach and Moulines (2013); Ostrovskii and Bach (2018), en conjonction avec la borne générale du SMP (Théorème 1.12).

Dans le cas paramétrique de dimension $d\ll n$, on en déduit directement le résultat suivant :

Corollaire 1.3 (Corollaire 7.2, Chapitre 7)**.** *Supposons que* $\|X\|\leqslant R$ *presque sûrement. Alors, l'estimateur* (1.105) *avec* $\lambda=2R^2/(n+1)$ *satisfait, pour tout* $B>0$*,*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big]-\inf_{\|\beta\|\leqslant B}R(\beta)\leqslant\frac{3d}{n}+\frac{B^2R^2}{n}\,. \tag{1.107}$$

La borne (1.107) est valable en supposant seulement que X est borné. Une telle garantie n'est pas possible pour un estimateur propre (tel que l'estimateur pénalisé de type Ridge (1.97)) sans hypothèse supplémentaire, par la borne inférieure (1.98) de Hazan et al. (2014). Tout comme la procédure de Foster et al. (2018) obtenue par conversion online-to-batch de stratégies

de mélange Bayésien, le SMP pénalisé admet une vitesse rapide sans dépendance exponentielle en BR.

Notons que la borne (1.101) est *logarithmique* en la norme BR, tandis que la borne (1.107) pour le SMP est quadratique en BR, à l'instar des résultats de Bach (2010); Marteau-Ferey et al. (2019) dans le cas bien spécifié. Dans le cas du modèle Gaussien, nous avions obtenu une borne avec une dépendance logarithmique en BR (Proposition 1.8, en utilisant un paramètre de régularisation λ plus faible lorsque BR est élevé ; cela n'est plus possible ici, car nous avons besoin de "localiser" le paramètre $\widehat{\beta}^{-Z}_{\lambda,n}$ dans un voisinage de diamètre constant de $\widehat{\beta}^{Z}_{\lambda,n}$, afin de pouvoir effectuer une approximation quadratique locale du risque en utilisant l'auto-concordance. Notons toutefois que dans le régime considéré précédemment où $BR = O(\sqrt{d})$, la borne (1.107) du SMP est bien d'ordre optimal $O(d/n)$, tandis que la borne (1.101) est d'ordre $O(d\log(n)/n)$.

D'un point de vue computationnel, la prédiction du SMP s'obtient en calculant les deux solutions $\widehat{\beta}^{(x,1)}_{\lambda,n}$ et $\widehat{\beta}^{(x,-1)}_{\lambda,n}$ de régressions logistiques "perturbées", obtenues en ajoutant un échantillon. Ainsi, comparé à une approche fondée sur le mélange Bayésien (Kakade and Ng, 2005; Foster et al., 2018), le SMP remplace un problème d'échantillonnage selon le postérieur par un problème d'optimisation, qui est typiquement moins coûteux. En ce sens, le SMP apporte une réponse au problème ouvert posé par Foster et al. (2018). Le SMP reste cependant plus coûteux qu'une régression logistique simple (lorsque l'on cherche à calculer la prédiction en plusieurs valeurs de x), car il requiert de calculer les solutions modifiées $\widehat{\beta}^{(x,1)}_{\lambda,n}$ et $\widehat{\beta}^{(x,-1)}_{\lambda,n}$ pour chaque point de test x. Cette différence de complexité s'observe également dans le cas du modèle Gaussien (Section 1.4.5) : une fois calculés $\widehat{\beta}^{\mathrm{LS}}_n$ et $\widehat{\Sigma}^{-1}_n$, avec un coût de $O(nd^2+d^3) = O(nd^2)$ (pour $n \geqslant d$), la prédiction de l'EMV en $x \in \mathbf{R}^d$, soit $\mathcal{N}(\langle\widehat{\beta}^{\mathrm{LS}}_n, x\rangle, 1)$, se calcule en temps $O(d)$, tandis que celle du SMP $\widehat{g}^{\mathrm{SMP}}_n(\cdot|x) = \mathcal{N}(\langle\widehat{\beta}^{\mathrm{LS}}_n, x\rangle, (1+\langle\widehat{\Sigma}^{-1}_n x, x\rangle)^2)$ s'obtient en temps $O(d^2)$. Il y a donc dans ce cas aussi un compromis entre garanties statistiques et coût computationnel.

1.5 Forêts aléatoires

Cette section porte sur les méthodes de forêts aléatoires, étudiées dans la Partie I. Dans la Section 1.5.1, nous décrivons le principe des méthodes de forêts aléatoires, avant de passer en revue les garanties théoriques existantes sur leur performance prédictive en Section 1.5.2. La Section 1.5.3 porte sur notre principale contribution, à savoir l'analyse précise d'une variante de forêts, les *forêts de Mondrian* (introduites par Lakshminarayanan et al., 2014), menée dans le Chapitre 2, ainsi que celle d'une variante séquentielle de l'algorithme (Chapitre 3).

1.5.1 Forêts aléatoires : principes généraux

Méthodes ensemblistes. Dans cette thèse, nous étudions une famille particulière de procédures utilisés tant en classification qu'en régression, à savoir les *forêts aléatoires* (*Random Forests* en anglais). Proposées par Breiman (2001a), et inspirées par les travaux de Amit and Geman (1997); Ho (1998); Dietterich (2000); Breiman (2000); Cutler and Zhao (2001), les forêts aléatoires appartiennent à la famille des méthodes dites d'*ensemble* ou *ensemblistes* (*ensemble methods* en anglais), qui consistent à combiner plusieurs prédicteurs individuels afin d'obtenir un prédicteur "agrégé" de bonne qualité. Les approches ensemblistes les plus courantes sont :

- le *boosting* (Freund and Schapire, 1997; Friedman, 2001; Schapire and Freund, 2012), qui combine des prédicteurs simples de manière séquentielle afin d'obtenir un prédicteur combiné ayant une erreur faible, chaque prédicteur élémentaire étant choisi de manière à corriger les erreurs des précédents. Cette procédure part de prédicteurs simples (ayant une forte erreur d'approximation) et opère une réduction du biais.

- le *bagging* (contraction de *bootstrap aggregating*), introduit par Breiman (1996). À l'inverse du boosting, cette approche part d'un prédicteur complexe caractérisé par une forte attache aux données, ce qui est généralement associé à un faible biais mais à une variance élevée. Le bagging consiste alors à construire plusieurs réalisations aléatoires indépendantes de ce prédicteur, obtenues en l'évaluant sur des sous-échantillons tirés aléatoirement, et à effectuer la moyenne de ces réalisations aléatoires. L'idée sous-jacente est de réduire la variance du prédicteur, en effectuant la moyenne de plusieurs réalisations faiblement ou modérément corrélées.

La philosophie des forêts aléatoires se rapproche davantage de celle du bagging — qui est d'ailleurs l'un des ingrédients de la procédure proposée par Breiman (2001a) — que de celle du boosting : elles partent de prédicteurs individuels complexes (les arbres de décision, décrits plus bas) construits de manière randomisée, dont elles forment la moyenne non pondérée. De plus, à l'instar du bagging et contrairement au boosting, les prédicteurs individuels sont construits en parallèle, indépendamment les uns des autres.

Arbres de décision. Les prédicteurs individuels que combinent les forêts aléatoires sont des *arbres de décision* (Breiman et al., 1984; Devroye et al., 1996). Ces derniers sont construits en partitionnant l'espace de manière récursive, en appliquant successivement des *coupures* qui séparent chaque cellule en deux. Afin de simplifier la discussion, nous supposons que $\mathcal{X} = [0,1]^d$; notons toutefois que les arbres de décision ont la propriété agréable qu'elles permettent de traiter simultanément des variables continues et discrètes (Breiman et al., 1984).

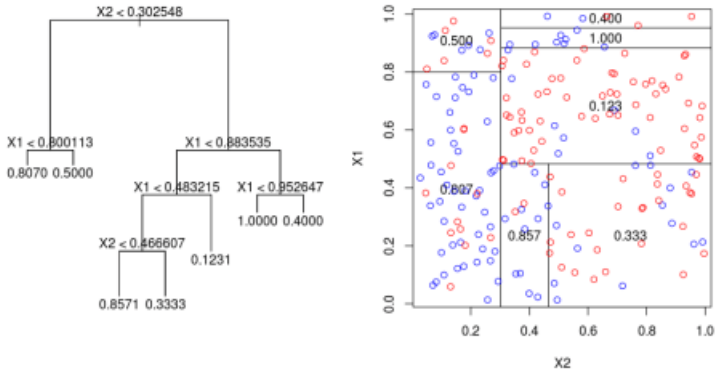


Figure 1.2: Un arbre de décision en dimension $d = 2$. Le schéma de gauche indique la structure d'arbre binaire (ainsi que les coupures associées à chaque nœud), tandis que la partie droite illustre la partition de $[0,1]^2$ correspondante. À chaque nœud de l'arbre, correspondant à une cellule de la partition, est associée une prédiction. Source : https://www.r-bloggers.com/regression-tree-using-ginis-index/.

Un arbre de décision (à valeurs dans $\widehat{\mathcal{Y}}$) est un prédicteur $g : [0,1]^d \to \widehat{\mathcal{Y}}$, caractérisé les éléments suivants (voir la Figure 1.2).

- Une partition finie $\mathscr{C}$ de $[0,1]^d$ dite *arborescente* (en anglais *kd-tree* ou *tree partition*), c'est-à-dire associée à
  - un arbre binaire $\mathcal{T}$ fini, enraciné et ordonné. Ainsi, chaque nœud $\mathbf{v}$ distinct de la racine admet un unique parent, et admet soit deux nœuds fils (un fils gauche noté $\mathbf{v}0$, et un fils droit noté $\mathbf{v}1$), auquel cas $\mathbf{v}$ est dit *intérieur*, soit aucun nœud fils, auquel cas $\mathbf{v}$ est une *feuille* (ou *nœud terminal*) ;
  - ainsi qu'une famille de *coupures* $\Sigma = (\sigma_{\mathbf{v}})_{\mathbf{v}}$ indexée par les nœuds intérieurs de $\mathcal{T}$. Une coupure σ est caractérisée par une coordonnée de coupure $j \in \{1, \dots, d\}$ ainsi que d'un seuil $s \in \mathbf{R}$.

  À partir de la donnée de l'arbre $\mathcal{T}$ et des coupures Σ, il est possible d'associer récursivement à tout nœud $\mathbf{v}$ de $\mathcal{T}$ une *cellule* $C_{\mathbf{v}} \subset [0,1]^d$, de la façon suivante :
  - la cellule associée à la racine ε de $\mathcal{T}$ est $C_\varepsilon = [0,1]^d$;
  - pour tout nœud intérieur $\mathbf{v}$, étant donné sa cellule $C_{\mathbf{v}}$ et la coupure $\sigma_{\mathbf{v}} = (j_{\mathbf{v}}, s_{\mathbf{v}})$, on définit $C_{\mathbf{v}0} = \{x \in C_{\mathbf{v}} : x_{j_{\mathbf{v}}} \leqslant s_{\mathbf{v}}\}$ (où $x = (x_j)_{1 \leqslant j \leqslant d}$), et $C_{\mathbf{v}1} = C_{\mathbf{v}} \setminus C_{\mathbf{v}0}$.

  Ainsi, les cellules $C_{\mathbf{v}}$ des feuilles $\mathbf{v}$ de $\mathcal{T}$ forment une partition $\mathscr{C}$ de $[0,1]^d$ associée à la structure d'arbre et aux coupures.
- Une prédiction $\widehat{y}_{\mathbf{v}} = \widehat{y}_C \in \widehat{\mathcal{Y}}$ associée à toute feuille $\mathbf{v}$ de $\mathcal{T}$, c'est-à-dire à toute cellule $C = C_{\mathbf{v}}$ de $\mathscr{C}$. Ces prédictions définissent une fonction $g : [0,1]^d \to \widehat{\mathcal{Y}}$ par $g(x) = \widehat{y}_{C(x)}$, où $C(x) \in \mathscr{C}$ est la cellule de x, telle que $x \in C(x)$.

Une fois la structure de la partition déterminée, les arbres de décision prédisent généralement en utilisant la minimisation du risque empirique dans chaque cellule ; dans le cas de la classification (Exemple 1.3), cela revient à choisir la classe la plus fréquente dans la cellule,

tandis que dans le cas de la régression (Exemple 1.4), cela correspond à la moyenne des Y_i tels que X_i appartienne à la cellule.

Ainsi, les algorithmes d'arbres de décision sont caractérisés par le choix de la partition de $[0,1]^d$. L'algorithme de référence en classification et régression à partir d'arbres de décision est l'algorithme CART (*Classification And Regression Trees*, Breiman et al., 1984). Il s'agit d'un algorithme *glouton*, qui optimise le choix de la coupure dans chaque nœud en fonction de la réduction immédiate de l'erreur apportée par cette coupure, plutôt que par une optimisation globale prenant en compte l'effet des coupures suivantes. Cela est dû au fait que le problème de la minimisation du risque empirique parmi tous les arbres de décision est NP-complet (Hyafil and Rivest, 1976).

Il reste alors à choisir la complexité des arbres, c'est-à-dire le critère d'arrêt qui détermine quand un nœud n'est plus coupé. L'algorithme CART (Breiman et al., 1984; Friedman et al., 2001) procède de la façon suivante : dans un premier temps, un arbre de décision complètement développé (ne contenant qu'un point par cellule) est formé de manière gloutonne ; dans un second temps, cet arbre est *élagué* (en enlevant certaines coupures), en minimisant l'erreur empirique pénalisée par le nombre de feuilles.

Les arbres de décision, et en particulier l'algorithme CART, sont simples, rapides à évaluer (une fois l'arbre de décision formé, calculer les prédictions ne demande que de comparer les coordonnées de l'entrée x aux seuils des coupures, le long du chemin de x dans l'arbre) et interprétables (l'arbre renseigne sur les variables qui ont déterminé la prédiction). Cependant, du point de vue de la performance prédictive, les arbres individuels souffrent de certaines limitations : en particulier les prédictions de l'algorithme CART s'avèrent instables, car fortement sensibles au choix des coupures (Breiman, 1996).

Forêts aléatoires. Comme nous l'avons annoncé, les forêts aléatoires sont obtenues en combinant des arbres individuels. La procédure proposée par Breiman (2001a) repose sur les ingrédients suivants :

- des arbres de décision construits en parallèle de manière randomisée, et non élagués : chaque arbre est "profond" et contient un faible nombre de points par cellule. Les arbres individuels sont combinés en faisant la moyenne de leurs prédictions (dans le cas de la régression), ou par un vote majoritaire (en classification) ;
- chaque arbre est construit sur un sous-échantillon de $(X_1, Y_1), \ldots, (X_n, Y_n)$, obtenu par exemple en tirant n points avec répétition parmi l'échantillon (on parle alors d'échantillon de type *bootstrap*, utilisé dans le cas du bagging) ;
- le choix de la partition (et en particulier des coupures) est partiellement randomisé, notamment en tirant aux hasard les coordonnées considérées pour chaque coupure potentielle.

Signalons également l'existence d'une variante de forêts aléatoires couramment utilisée, à savoir l'algorithme *Extra-Trees* (Geurts et al., 2006). Cette procédure combine également les ingrédients ci-dessus, à l'exception du sous-échantillonnage (bagging).

Les forêts aléatoires comptent parmi les algorithmes de classification et de régression les plus couramment utilisés en pratique (Fernández-Delgado et al., 2014). Elles combinent en effet une très bonne performance prédictive, un coût d'entraînement et d'évaluation modeste, et le peu (voire l'absence) de paramètres libres à calibrer. Ces succès empiriques contrastent

cependant avec une compréhension et des garanties théoriques limitées pour cette famille de procédures (Arlot and Genuer, 2014; Wager and Walther, 2015; Biau and Scornet, 2016).

1.5.2 Revue des résultats théoriques sur les forêts aléatoires

Dans cette section, nous passons en revue les garanties théoriques disponibles sur la performance prédictive des forêts aléatoires. Nous nous restreignons au cas de la régression avec perte quadratique (pour lequel $\mathcal{Y} = \widehat{\mathcal{Y}} = \mathbf{R}$), ce qui permet d'avoir recours à une décomposition biais-variance précise. Signalons également que des résultats généraux permettent de convertir toute garantie de risque en régression en une garantie en classification (Devroye et al., 1996). Au delà de l'erreur de prédiction, de nombreux autres questions relatives aux forêts, tant théoriques que méthodologiques, ont été étudiées ; nous renvoyons à Criminisi et al. (2012); Boulesteix et al. (2012); Biau and Scornet (2016) pour des revues plus complètes sur ce sujet.

Biais, variance et randomisation. Afin de clarifier la discussion ultérieure, nous commençons par des préliminaires généraux sur le risque, le biais et la variance de prédicteurs randomisés et de leurs ensembles. Ces résultats s'appliquent à tous les estimateurs de type ensembliste (qui effectuent la moyenne simple d'estimateurs randomisés individuels) et donc en particulier aux différentes variantes de forêts aléatoires.

Soit $\widehat{g}_n(\cdot;\Theta) : [0,1]^d \to \mathbf{R}$ un prédicteur randomisé, construit à partir du jeu de données i.i.d. $\mathscr{D}_n = \big((X_1,Y_1),\dots,(X_n,Y_n)\big)$ et de l'aléa Θ introduit par l'algorithme (pour les forêts, à travers le choix aléatoire des coupures, le sous-échantillonnage, etc.). Définissons pour tout $x \in [0,1]^d$:

$$\bar{g}_n(x;\Theta) = \mathbb{E}\big[\widehat{g}_n(x;\Theta)|\Theta\big]\,,$$

où l'espérance conditionnelle par rapport à Θ revient à intégrer sur l'échantillon aléatoire $\mathscr{D}_n$. Ainsi, en notant $g^*(X) = \mathbb{E}[Y|X]$ la fonction de régression (Exemple 1.4), on a pour tout $x \in \mathbf{R}^d$

$$\mathbb{E}\big[(\widehat{g}_n(x;\Theta) - g^*(x))^2|\Theta\big] = \mathbb{E}\big[(\bar{g}_n(x;\Theta) - g^*(x))^2|\Theta\big] + \mathrm{Var}\big(\widehat{g}_n(x;\Theta)|\Theta\big)\,. \tag{1.108}$$

En prenant $x = X$ dans (1.108), en considérant l'espérance sur (X,Θ) et en notant que $R(g) - R(g^*) = \mathbb{E}[(g(X) - g^*(X))^2]$ pour toute fonction $g : [0,1]^d \to \mathbf{R}^d$ (Exemple 1.4), on obtient la *décomposition biais-variance* suivante[17] :

$$\mathbb{E}\big[R(\widehat{g}_n(\cdot;\Theta))\big] - R(g^*) = \mathbb{E}\big[(\bar{g}_n(X;\Theta) - g^*(X))^2\big] + \mathbb{E}\big[\mathrm{Var}(\widehat{g}_n(X;\Theta)|X,\Theta)\big]\,. \tag{1.109}$$

Le premier terme du membre de droite de (1.109) constitue le *biais* de $\widehat{g}_n$; il s'agit de l'espérance sur (X,Θ) d'un terme indépendant de l'échantillon. Le second terme correspond à la *variance* de $\widehat{g}_n$. Notons que le terme de "variance" se réfère ici à la variabilité induite par l'échantillon $\mathscr{D}_n$; l'aléa Θ introduit par l'algorithme contribue quant à lui à la fois au biais et à la variance. Un estimateur de variance élevée est instable, en général car il s'attache trop aux spécificités du jeu de données. À l'inverse, une procédure présentant un biais élevé est insuffisamment flexible pour approcher la fonction de régression g^*.

[17] Cette décomposition diffère légèrement de celle que nous utilisons au Chapitre 2, qui introduit un terme intermédiaire distinct (mais proche) de $\bar{g}_n$.

Risque d'estimateurs ensemblistes. Considérons maintenant $M \geqslant 1$ réalisations i.i.d. $\Theta^{(1)}, \ldots, \Theta^{(M)}$ de Θ, indépendantes de l'échantillon $\mathscr{D}_n$, et définissons l'estimateur d'ensemble (moyenné)

$$\widehat{g}_{n,M}(x;\Theta_M) = \frac{1}{M}\sum_{m=1}^{M} \widehat{g}_n(x;\Theta^{(m)})$$

avec $\Theta_M = (\Theta^{(1)}, \ldots, \Theta^{(M)})$. À $\mathscr{D}_n$ et x fixés, la loi des grands nombres implique que, lorsque $M \to \infty$, $\widehat{g}_{n,M}(x;\Theta_M)$ converge presque sûrement vers $\widehat{g}_{n,\infty}(x) := \mathbb{E}[\widehat{g}_{n,M}(x;\Theta^{(1)})|\mathscr{D}_n]$ dès lors que cette espérance est bien définie. $\widehat{g}_{n,\infty}$ correspond à un estimateur *non-randomisé* de g^*. En écrivant une décomposition "biais-variance" par rapport à Θ_M, conditionnellement à $\mathscr{D}_n$, on obtient pour tout $x \in \mathbf{R}^d$:

$$\begin{aligned}\mathbb{E}\big[(\widehat{g}_{n,M}(x;\Theta_M) - g^*(x))^2|\mathscr{D}_n\big] &= \big(\mathbb{E}[\widehat{g}_{n,M}(x;\Theta_M)|\mathscr{D}_n] - g^*(x)\big)^2 + \mathrm{Var}\big(\widehat{g}_{n,M}(x;\Theta_M)|\mathscr{D}_n\big)\\ &= \big(\widehat{g}_{n,\infty}(x) - g^*(x)\big)^2 + \frac{1}{M}\mathrm{Var}\big(\widehat{g}_n(x;\Theta^{(1)})|\mathscr{D}_n\big)\,;\end{aligned}$$

en considérant l'espérance sur $x = X$ et $\mathscr{D}_n$, on obtient :

$$\mathbb{E}\big[R(\widehat{g}_{n,M})\big] = \mathbb{E}\big[R(\widehat{g}_{n,\infty})\big] + \frac{1}{M}\mathbb{E}\big[\mathrm{Var}\big(\widehat{g}_n(X;\Theta^{(1)})|X,\mathscr{D}_n\big)\big]\,. \tag{1.110}$$

L'expression (1.110) montre que le risque moyen de $\widehat{g}_{n,M}$ décroît avec le nombre de répétitions M, et converge lorsque $M \to \infty$ vers celui de l'estimateur idéalisé $\widehat{g}_{n,\infty}$, correspondant à un ensemble infini. Bien souvent, $\widehat{g}_{n,\infty}$ est complexe et n'est pas calculable explicitement ; $\widehat{g}_{n,M}$ en constitue un approximation de type Monte-Carlo, d'autant plus précise que M est élevé. En pratique, M est contraint par des limites computationnelles, et l'expression (1.110) quantifie le risque additionnel encouru pour une valeur donnée de M.

Remarque 1.8 (Nombre de répétitions). En considérant le cas $M = 1$ dans (1.110), il vient

$$\mathbb{E}\big[\mathrm{Var}\big(\widehat{g}_n(X;\Theta^{(1)})|X,\mathscr{D}_n\big)\big] = \mathbb{E}\big[R(\widehat{g}_{n,M})\big] - \mathbb{E}\big[R(\widehat{g}_{n,\infty})\big] \leqslant \mathbb{E}\big[R(\widehat{g}_{n,M})\big] - R(g^*)\,,$$

de sorte que (1.110) implique, en notant $\mathcal{R}(\widehat{h}_n) = \mathbb{E}[R(\widehat{h}_n)] - R(g^*)$ l'excès de risque en espérance de l'estimateur $\widehat{h}_n$:

$$\mathcal{R}(\widehat{g}_{n,M}) \leqslant \mathcal{R}(\widehat{g}_{n,\infty}) + \frac{1}{M}\cdot\mathcal{R}(\widehat{g}_{n,1})\,.$$

Ainsi, on a $\mathcal{R}(\widehat{g}_{n,M}) \lesssim \mathcal{R}(\widehat{g}_{n,\infty})$ dès lors que $M \gtrsim \mathcal{R}(\widehat{g}_{n,1})/\mathcal{R}(\widehat{g}_{n,\infty})$. Le nombre de prédicteurs individuels M nécessaires à une performance optimale dépend donc de la qualité relative des prédicteurs individuels et ensemblistes.

Considérons à présent l'effet de la moyenne $\widehat{g}_{n,M}$ sur le biais et la variance (au sens de la décomposition (1.109), c'est-à-dire par rapport à l'échantillon aléatoire $\mathscr{D}_n$). Définissons comme précédemment les intermédiaires

$$\bar{g}_{n,M}(x;\Theta_M) = \mathbb{E}[\widehat{g}_{n,M}(x;\Theta_M)|\Theta_M] = \frac{1}{M}\sum_{m=1}^{M}\bar{g}_n(x;\Theta^{(m)})$$

(qui ne dépend que de Θ_M et pas de $\mathscr{D}_n$) et $\bar{g}_{n,\infty}(x) = \mathbb{E}[\widehat{g}_{n,\infty}(x)] = \mathbb{E}[\widehat{g}_n(X;\Theta)]$ (qui est purement déterministe). Le terme de biais dans la décomposition (1.109) vaut alors :

$$\mathbb{E}\big[(\bar{g}_{n,M}(X;\Theta_M) - g^*(X))^2\big] = \mathbb{E}\big[(\bar{g}_{n,\infty}(X) - g^*(X))^2\big] + \frac{1}{M}\mathbb{E}\big[\mathrm{Var}\big(\bar{g}_n(X;\Theta)|X\big)\big]\,; \tag{1.111}$$

le premier terme de la décomposition (1.111) correspond au biais de la procédure idéalisée $\widehat{g}_{n,\infty}$, tandis que le second constitue le biais additionnel de son approximation $\widehat{g}_{n,M}$. De même, le terme de variance dans (1.109) s'écrit

$$\begin{aligned}
&\mathbb{E}\big[\mathrm{Var}(\widehat{g}_{n,M}(X;\Theta_M)|X,\Theta_M)\big] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\big[\mathrm{Var}(\widehat{g}_n(X;\Theta^{(m)})|X,\Theta^{(m)})\big] + \\
&\quad + \frac{1}{M^2}\sum_{m\neq m'}\mathbb{E}\big[\mathrm{Cov}(\widehat{g}_n(X;\Theta^{(m)}),\widehat{g}_n(X;\Theta^{(m')})|X,\Theta^{(m)},\Theta^{(m')})\big] \\
&= \frac{1}{M}\mathbb{E}\big[\mathrm{Var}(\widehat{g}_n(X;\Theta^{(1)})|X,\Theta^{(1)})\big] + \frac{M-1}{M}\mathbb{E}\big[\mathrm{Cov}\big(\widehat{g}_n(X;\Theta^{(1)}),\widehat{g}_n(X;\Theta^{(2)})|X,\Theta^{(1)},\Theta^{(2)}\big)\big].
\end{aligned}$$

En particulier, pour M grand, la variance de l'estimateur moyenné $\widehat{g}_{n,M}$ est proche de celle de l'ensemble infini, qui vaut

$$\mathbb{E}\big[\mathrm{Var}(\widehat{g}_{n,\infty}(X)|X)\big] = \mathbb{E}\big[\mathrm{Cov}\big(\widehat{g}_n(X;\Theta^{(1)}),\widehat{g}_n(X;\Theta^{(2)})|X,\Theta^{(1)},\Theta^{(2)}\big)\big]\,. \tag{1.112}$$

Ainsi, la variance de l'estimateur ensembliste infini dépend de la corrélation (par rapport à l'échantillon $\mathscr{D}_n$) entre les prédictions de deux réalisations indépendantes de l'estimateur randomisé individuel.

Garanties théoriques sur les forêts. L'article original de Breiman (2001a) établit des bornes sur l'erreur des forêts (une pour la classification et une pour la régression), en fonction de la qualité des arbres individuels et de la corrélation entre les différents arbres ; la borne obtenue en régression est similaire à celle issue de la décomposition biais-variance indiquée plus haut.

Un pan important de la littérature étudie les propriétés de *consistance* de forêts aléatoires. Biau et al. (2008) établit la consistance de certains types de forêts en classification en la déduisant de celle des arbres individuels[18], et montre également que certains classifieurs ensemblistes sont consistants même lorsque les classifieurs individuels ne le sont pas.

Pour montrer la consistance d'arbres individuels, il est possible d'utiliser des résultats généraux de consistance d'histogrammes (Devroye et al., 1996). Par exemple, un classifieur en histogramme dont (1) la partition est construite indépendamment des réponses Y_i ; (2) le diamètre de la cellule du point de test X tend en loi vers 0 lorsque $n \to \infty$; et (3) le nombre de points de l'échantillon dans la cellule du point X tend vers l'infini lorsque $n \to \infty$, est consistant indépendamment de la loi de (X, Y) en classification (Devroye et al., 1996). Ces résultats découlent eux-mêmes de résultats généraux de consistance d'estimateurs à moyenne locale dûs à Stone (1977), dont la consistance des k-plus proche voisins (énoncée dans le Théorème 1.2) est un cas particulier. Garantir les conditions nécessaires à la consistance pour des forêts (ou pour des arbres) requiert en général de considérer le mécanisme de construction des arbres, tant du point de vue de la structure combinatoire de l'arbre (en particulier le nombre de cellules) que de la loi des coupures (qui affecte notamment le diamètre des cellules).

À la suite de Biau et al. (2008) (qui étudie des modèles simples de forêts), un axe de recherche récent consiste à établir la consistance de variantes de forêts plus sophistiquées (Denil

[18]Dans le cas de la classification (contrairement à celui de la régression), le risque d'un ensemble de prédicteurs randomisés combinés par vote majoritaire n'est pas nécessairement inférieur à celui des classifieurs individuels, à cause de la non-convexité de l'erreur de classification.

et al., 2014; Scornet et al., 2015; Wager and Walther, 2015; Mentch and Hooker, 2016; Cui et al., 2017; Wager and Athey, 2018; Athey et al., 2019), se rapprochant davantage des forêts utilisées en pratique. Bien que plus complexes (en raison de la complexité des algorithmes considérés), les preuves reposent sur les mêmes principes généraux. Wager and Athey (2018) et Mentch and Hooker (2016) ont indépendamment établi la normalité asymptotique de l'estimateur des forêts aléatoires, lorsque la taille n de l'échantillon tend vers l'infini ; en particulier, ils montrent que la variance asymptotique de l'estimateur des forêts peut être estimée à l'aide du Jackknife. Ces résultats sont utiles dans le cadre de tâches inférentielles plus générales que la prédiction, comme par exemple pour former des intervalles de confiance sur les valeurs de la fonction de régression. En revanche, ils ne fournissent pas de vitesse théorique sur la variance asymptotique de l'estimateur, et donc de garantie sur la qualité de celui-ci.

Difficultés liées au choix glouton des coupures. Dans le cas des forêts de Breiman, l'obtention de garanties théoriques précises est délicate. Biau et al. (2008) montrent notamment que les forêts de Breiman $\widehat{g}_n$ sont en fait *inconsistantes*, au sens où il existe une loi jointe de (X, Y) telle que $R(g^*) = 0$ mais $\mathbb{E}[R(\widehat{g}_n)] \geqslant c > 0$ pour tout n. Cette limitation tient au fait que ces forêts sont construites de manière *gloutonne*, en optimisant le choix de coupures en fonction de la réduction immédiate de l'erreur. Il est en effet possible de construire des lois de (X, Y) pour lesquelles certaines coupures nécessaires (car permettant des coupures ultérieures réduisant fortement l'erreur) mais n'entraînant pas de réduction immédiate de l'erreur ne seront jamais réalisées, l'algorithme lui préférant indéfiniment des coupures induisant un (faible) gain immédiat.

Scornet et al. (2015) démontrent la consistance d'un algorithme de forêts proche de celui de Breiman, mais avec des arbres suffisamment élagués[19]. Cette garantie de consistance est obtenue en régression, en supposant que la loi de Y sachant X suit un modèle additif (Stone, 1985) :

$$Y = \sum_{j=1}^{d} f_j(X^j) + \varepsilon\,,$$

où X^j désigne la $j$ième coordonnée de X, et où $\mathbb{E}[\varepsilon|X] = 0$. En d'autres termes, la fonction de régression $g^*(X) = \mathbb{E}[Y|X]$ se décompose comme une somme de fonctions des coordonnées, et ne présente pas de terme d'interaction. L'absence d'interactions permet d'éviter les configurations mettant en défaut le choix glouton des coupures, où le bénéfice d'une coupure n'apparaît qu'après des coupures ultérieures. Notons cependant que les arbres et les forêts, qui combinent des coupures selon plusieurs variables, permettent en principe d'approcher des fonctions de régression avec des effets d'interaction complexes.

Dans une autre approche, Wager and Walther (2015) établissent un résultat garantissant la validité de l'estimation des moyennes dans les feuilles, uniformément sur une famille de partitions "médianes" où chaque coupure partage les données en deux fractions équilibrées. Ce résultat s'applique donc aux partitions "adaptatives", qui utilisent les données (et en particulier les sorties Y_i) pour choisir les coupures, à condition que les Y_i utilisés pour choisir les coupures soient disjoints de ceux utilisés pour estimer les moyennes. Afin d'obtenir la consistance, Wager and Walther (2015) considèrent également une hypothèse sur la fonction de régression (d'effet minoré de coupures individuelles) qui assure la qualité du choix glouton de coupures.

[19] Dans le cas des forêts de Breiman classiques qui ne sont pas élaguées, Scornet et al. (2015) réduit la consistance à une conjecture relativement délicate à établir.

Vitesses de convergence pour des forêts stylisées. La seule consistance ne donne que peu d'information sur l'effet de différents choix de paramètres de l'algorithme, ainsi que sur les facteurs qui influencent la qualité des prédictions. Il est en effet possible d'établir la consistance même au moyen d'une analyse lâche, qui ne capture que très partiellement le comportement effectif de l'algorithme. Pour cette raison, plusieurs travaux cherchent à quantifier l'erreur de prédiction des forêts, ainsi que sa dépendance en les paramètres de l'algorithme, sous certaines hypothèses sur la fonction de régression.

En raison des difficultés liées au choix glouton des coupures, il est commode de considérer des forêts simplifiées, où le choix des coupures ne se fait plus par minimisation de l'erreur mais par randomisation. Cela correspond dans une certaine mesure à la motivation initiale des forêts, qui cherchent à pallier la trop forte instabilité des arbres CART (due notamment à la sensibilité au choix des coupures, Breiman, 1996, 2001a). Cette approche conduit à considérer des forêts stylisées dites *purement aléatoires* (en anglais *purely random forests*, abrégé PRF), pour lesquelles la partition est générée indépendamment du jeu de données. Des forêts de ce type ont été considérées par Cutler and Zhao (2001); Breiman (2000), et analysées par Breiman (2004); Biau et al. (2008); Biau (2012); Genuer (2012); Arlot and Genuer (2014); Klusowski (2018) ; nous étudions également un tel algorithme dans le Chapitre 2. Signalons également que des idées similaires de partitions aléatoires ont été étudiées par Rahimi and Recht (2008, 2009) ; en particulier, l'algorithme de *random binning* de Rahimi and Recht (2008) repose sur des partitions aléatoires du cube $[0,1]^d$. À l'inverse, chaque arbre aléatoire (en tant qu'histogramme) effectue une régression linéaire avec pour variables (aléatoires) les indicatrices des différentes cellules de la partition, à l'instar de la procédure de *Random Kitchen Sinks* (RKS) de Rahimi and Recht (2009). La façon dont les différentes partitions sont utilisées diffère cependant entre les PRF et les RKS ; tandis que les premières effectuent la moyenne simple des histogrammes associés aux arbres individuels, les seconds les combinent avec des poids eux-mêmes optimisés.

Breiman (2004); Biau (2012) considèrent un algorithme de type PRF, les *forêts centrées* (Biau and Scornet, 2016), obtenues de la façon suivante :

- les arbres sont complets de profondeur p (à 2^p feuilles), où $p = p_n$ est fixé ;
- les coupures sont obtenues dans chaque nœud en coupant au milieu d'une coordonnée j sélectionnée au hasard, uniformément sur $\{1, \dots, d\}$.

Biau (2012) établit que de telles forêts (avec un choix convenable de p_n) atteignent un risque de $O(n^{-1/((4/3\cdot\log 2)d+1)})$ lorsque la fonction de régression $g^* : [0,1]^d \to \mathbf{R}$ est Lipschitz. De plus, si la fonction de régression ne dépend que de $s \leqslant d$ variables, et si l'algorithme choisit en fait chacune de ces variables avec probabilité proche de $1/s$ (et les variables restantes avec une probabilité faible), alors la vitesse de convergence devient $O(n^{-1/((4/3\cdot\log 2)s+1)})$, ce qui apporte un gain significatif lorsque la dimension d est élevée mais le nombre s de variables informatives est faible. Biau (2012) propose également un mécanisme en partie heuristique de sélection de variables, permettant de sélectionner les s variables aléatoires en ayant recours à un échantillon indépendant. Duroux and Scornet (2018); Wager and Walther (2015) considèrent des forêts proches (dites *médianes*), et obtiennent la même vitesse de $O(n^{-1/((4/3\cdot\log 2)d+1)})$, ainsi que $O(n^{-1/((4/3\cdot\log 2)s+1)})$ dans le cas de forêts coupant selon les s variables informatives. Klusowski (2018) complète ces résultats en établissant des bornes inférieures pour les forêts centrées.

Les vitesses de convergences précédentes s'avèrent sous-optimales par rapport à la vitesse minimax sur la classe des fonctions Lipschitz, qui est de $O(n^{-2/(d+2)})$ (Stone, 1980, 1982; Györfi et al., 2002). Intuitivement, cela tient au fait que le choix aléatoire (uniforme) des coordonnées de coupure conduit à des cellules déséquilibrées, où certaines coordonnées ont été choisies moins souvent que d'autres ; le diamètre de ces cellules est donc élevé à cause de ces coordonnées. De manière générale, la nature récursive des arbres aléatoires, qui introduisent de l'aléa à chaque nouvelle coupure, peut conduire à des partitions déséquilibrées ou délicates à contrôler de manière théorique. Arlot and Genuer (2014) obtiennent la vitesse optimale $O(n^{-2/3})$ dans le cas $d = 1$ pour une autre variante de PRF, appelée *forêts purement aléatoires uniformes* (*purely uniformly random forest*, PURF, dont la partition associée de $[0, 1]$ s'obtient en tirant un nombre fixe $k - 1 \geqslant 0$ d'extrémités uniformément dans $[0, 1]$), mais une vitesse sous-optimale en dimension $d \geqslant 2$ pour une autre variante de PRF. Enfin, dans le Chapitre 2, nous étudions une variante de PRF, les *forêts de Mondrian* (Lakshminarayanan et al., 2014), dont nous montrons qu'elles atteignent la vitesse minimax $O(n^{-2/(d+2)})$ (voir également la Section 1.5.3).

Avantage des forêts par rapport aux arbres. Comme nous venons de le voir, les vitesses de convergence évoquées précédemment permettent de distinguer différentes façons de construire les arbres de manière randomisée, en montrant que certaines constructions conduisent à des vitesses moins bonnes que d'autres, en raison du caractère plus au moins bien équilibré des partitions. Cependant, tous les résultats précédents s'appliquent tant aux forêts qu'aux arbres individuels, et ne mettent donc pas en évidence d'avantage des forêts par rapport aux arbres.

Un tel effet a été établi pour la première fois par Arlot and Genuer (2014) : pour l'algorithme PURF décrit plus haut (qui combine des histogrammes aléatoires en dimension 1), lorsque la fonction de régression est deux fois continûment différentiable, la forêt infinie atteint une vitesse améliorée de[20] $O(n^{-4/5})$, qui est optimale pour cette classe de fonctions, à l'inverse des arbres simples qui admettent toujours une vitesse de $O(n^{-2/3})$ dans ce cas. Arlot and Genuer (2014) établissent également des vitesses améliorées pour les forêts dans le cas de PRF en dimension $d \geqslant 2$, bien que les vitesses soient dans ce cas sous-optimales. Dans le Chapitre 2, nous étendons les résultats optimaux de Arlot and Genuer (2014) pour les PURF dans le cas $d = 1$, au cas général $d \geqslant 1$ pour les forêts de Mondrian, en montrant que les forêts infinies (ou avec suffisamment d'arbres) atteignent la vitesse optimale $O(n^{-4/(d+4)})$ sur la classe des fonctions deux fois différentiables, tandis que les arbres atteignent seulement la vitesse en $O(n^{-2/(d+2)})$ du cas Lipschitz.

Dans ces résultats, il est important de noter que les vitesses améliorées pour les forêts proviennent d'une réduction du *biais*, et non de la *variance* (qui n'est réduite que d'un facteur constant). L'effet sous-jacent est un phénomène de lissage : tandis qu'un arbre exhibe des discontinuités aux bords des cellules, une forêt présente plusieurs petites discontinuités provenant de chacun des arbres, ayant lieu à des seuils différents. La fonction de régression associée à une forêt est par conséquent plus lisse que celle d'un arbre, et approche donc mieux les fonctions régulières.

Cet effet de réduction du biais (et non de la variance) va à rebours de la motivation initiale ayant conduit aux forêts, et plus généralement aux méthodes d'agrégation ensembliste de type bagging. En effet, l'objectif de ces méthodes est de partir de prédicteurs individuels complexes

[20] À un effet de bord près, commun à toutes les procédures par moyennes locales (Wasserman, 2006).

(de faible biais mais de variance élevée), et d'en réduire la variance. En particulier, les forêts de Breiman combinent des arbres plus profonds (non élagués) que ceux de l'algorithme CART ; à l'inverse, les résultats obtenus pour les PRF (réduction du biais) suggèrent une profondeur optimale des forêts inférieure à celle des arbres simples. Il y a deux interprétations complémentaires à ces résultats : d'une part, le lissage opéré par les forêts entraîne une réduction du biais, davantage que de la variance (le phénomène de lissage a également été relevé par Bühlmann and Yu, 2002, qui l'associent à une réduction d'un facteur constant de la variance) ; d'autre part, les modèles de forêts *purement aléatoires* ne permettent pas de mettre en évidence un effet de réduction de la variance dans les forêts, qui justifierait d'utiliser des arbres profonds.

Ainsi, à notre connaissance, aucun résultat existant ne permet de justifier l'usage des forêts aléatoires telles que proposées par Breiman (c'est-à-dire constituées d'arbres individuels profonds et non élagués) ou de montrer leur avantage par rapport à des arbres simples, convenablement élagués. Les propriétés des forêts avancées pour justifier leur performance, telles que leur capacité à sélectionner des variables informatives (Breiman, 2004; Biau, 2012; Scornet et al., 2015; Wager and Walther, 2015) ou à s'adapter à la "dimension intrinsèque" (notamment aux corrélations) des variables (Dasgupta and Freund, 2008; Verma et al., 2009), bien que pertinentes, sont également applicables aux arbres individuels. Comme souligné précédemment, l'effet de réduction du biais (par lissage des prédiction) conduit à sélectionner des forêts *moins profondes* que les arbres individuels. Il est bien possible de montrer que le sous-échantillonnage permet de réduire la variance, notamment dans le cas de l'estimateur du plus proche voisin (Biau and Devroye, 2010; Biau et al., 2010; Samworth, 2012) ; toutefois, cette réduction de variance suppose que la taille k_n des sous-échantillons satisfait $k_n/n \to 0$; dans le cas où $k_n \asymp n$, qui correspond au *bagging* utilisé en pratique, les résultats existants ne montrent qu'une réduction d'un facteur constant de la variance.

Une piste possible est d'étudier des partitions aléatoires partiellement adaptatives (contrairement à celles des estimateurs de forêts purement aléatoires, pour lequel la moyennisation ne réduit que le biais), pour lesquelles il est possible d'escompter une réduction de la variance. Une approche plus abordable consisterait à étudier l'effet de la moyenne de prédicteurs individuels "complexes" (qui interpolent le jeu de données), construits par un sous-échantillonnage des observations ou des variables, pour des procédures plus simples que les méthodes de forêts, telles que des méthodes linéaires, pour lesquelles une analyse explicite semble envisageable.

1.5.3 Analyse des forêts de Mondrian (Chapitres 2 et 3)

Notre principale contribution à l'étude des forêts est l'analyse d'une variante de forêts purement aléatoires (PRF), à savoir les *forêts de Mondrian* (Lakshminarayanan et al., 2014). Ces forêts reposent sur une loi particulière sur les partitions arborescentes du cube $[0,1]^d$, appelée *partitions de Mondrian*, introduites par Roy and Teh (2009). Dans la discussion qui suit, nous allons donner une définition récursive informelle de cette loi ; pour une définition rigoureuse, nous renvoyons à Roy (2011) ou à la Section 2.8.1 du Chapitre 2.

Définition des forêts de Mondrian. Le *processus de Mondrian* $\mathsf{MP}([0,1]^d)$ (Roy and Teh, 2009) est en fait (la loi d') un processus $(\Pi_\lambda)_{\lambda \in \mathbf{R}^+}$ de partitions Π_λ de $[0,1]^d$, telles que la partition $\Pi_{\lambda'}$ est un raffinement de la partition Π_λ pour tous $\lambda' \geqslant \lambda \geqslant 0$, obtenue en effectuant d'éventuelles coupures supplémentaires des cellules de Π_λ. Le paramètre λ, appelé *durée de vie* de la partition Π_λ, gouverne la complexité de cet arbre. La partition Π_λ, dont la loi est

notée $\mathsf{MP}(\lambda, [0,1]^d)$ s'obtient de la façon suivante (nous définissons en fait la loi $\mathsf{MP}(\lambda, C)$ pour tout hyperrectangle $C = \prod_{j=1}^d [a_j, b_j] \subset \mathbf{R}^d$) :

En notant ε la racine de l'arbre associé, on pose $C_\varepsilon = C$. En notant $C = \prod_{j=1}^d [a_j, b_j]$, soient E_j, $j = 1, \dots, d$, des variables exponentielles indépendantes d'intensités respectives $b_j - a_j$, et soit $E = \min_{1 \leqslant j \leqslant d} E_j$.

- Si $E > \lambda$, la cellule C_ε n'est pas coupée, la partition Π_λ est simplement $\{C\}$.
- Sinon, soit $J = \arg\min_{1 \leqslant j \leqslant d} E_j$, et S une variable uniforme sur $[a_J, b_J]$. La cellule C_ε est alors coupée selon la coordonnée J au seuil S, en deux cellules $C_0 := \{x \in C_\varepsilon : x_J \leqslant S\}$ et $C_1 := C_\varepsilon \backslash C_0$. Les cellules C_0 et C_1 sont alors elles-mêmes partitionnées indépendamment en répétant ce procédé, mais avec un budget de $\lambda - E$: $\Pi^0 \sim \mathsf{MP}(\lambda - E, C_0)$ et $\Pi^1 \sim \mathsf{MP}(\lambda - E, C_1)$, et la partition Π_λ est $\Pi^0 \cup \Pi^1$.

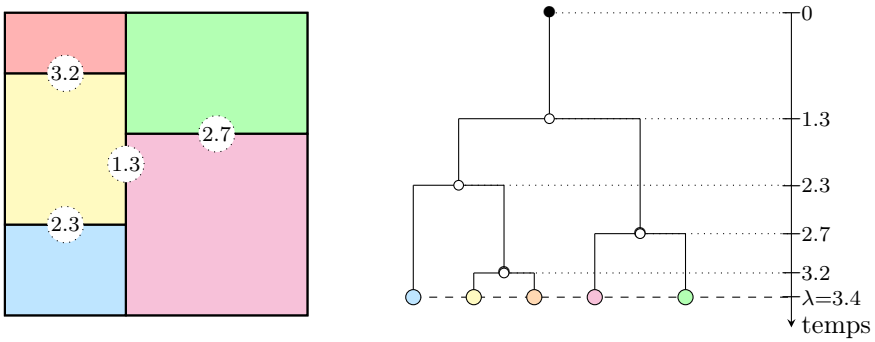


Figure 1.3: Une partition de Mondrian (à gauche), avec la structure d'arbre correspondante (à droite). Les temps des coupures sont indiqués sur l'axe vertical, tandis que les coupures sont signalés par des cercles (∘).

On obtient ainsi une partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$ de C, avec ici $C = [0,1]^d$. En faisant varier le budget λ, c'est-à-dire l'instant à partir duquel on cesse de couper les cellules, on obtient le processus $(\Pi_\lambda)_{\lambda \in \mathbf{R}^+} \sim \mathsf{MP}([0,1]^d)$. Par définition, $\Pi_{\lambda'}$ raffine Π_λ pour tout $\lambda' \geqslant \lambda$. De plus, on vérifie (en utilisant l'absence de mémoire de la loi exponentielle) que $(\Pi_\lambda)_\lambda$ est un processus de Markov, au sens où pour $\lambda \leqslant \lambda'$, $\Pi_{\lambda'}$ est indépendant de $(\Pi_t)_{t<\lambda}$ conditionnellement à Π_λ.

En dimension $d = 1$, il est possible de montrer que $\mathsf{MP}(\lambda, [0,1])$ est la loi de la partition de $[0,1]$ dont les lieux de coupures forment un processus ponctuel de Poisson d'intensité λ (Roy and Teh, 2009; Roy, 2011). En outre, une propriété fondamentale du processus de Mondrian est celle de restriction ; cette propriété découle des propriétés des variables exponentielles.

Proposition 1.9 (Roy, 2011)**.** *Soient $C_0 \subset C_1$ deux hyperrectangles de $\mathbf{R}^d$. Si $\Pi_\lambda \sim \mathsf{MP}(\lambda, C_1)$, alors la partition $\Pi_\lambda|_{C_0} = \{A \cap C_0 : A \in \Pi_\lambda\}$ de C_0 induite par restriction de la partition Π_λ suit la loi $\mathsf{MP}(\lambda, C_0)$.*

Les *forêts de Mondrian* ont été proposées par Lakshminarayanan et al. (2014). Cet algorithme effectue la moyenne d'arbres de décision tirés indépendamment selon la loi $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$; cet estimateur utilise également une forme de régularisation au sein de chaque arbre (contrairement aux estimateurs par histogrammes), au moyen d'une procédure bayésienne hiérarchique sur l'arbre calculée de manière approximative (Lakshminarayanan et al.,

2014). Nous omettons ce second aspect de la procédure, et appelons *forêt de Mondrian* l'estimateur de type PRF obtenu en utilisant des partitions de loi $\mathsf{MP}(\lambda, [0,1]^d)$; dans ce cas, la régularisation est obtenue par le choix du paramètre de complexité λ des partitions. Nous reviendrons sur cet aspect à la fin de cette section et dans le Chapitre 3.

Analyse des forêts de Mondrian (Chapitre 2). Les forêts de Mondrian ont été introduites pour des raisons computationnelles, en tant que procédure calculable de manière séquentielle (en ligne) à partir des propriétés de Markov et de restriction des partitions de Mondrian (Lakshminarayanan et al., 2014).

Dans le Chapitre 2, nous montrons que cette variante de PRF se prête à une analyse théorique précise. Celle-ci repose sur le fait qu'il est possible d'obtenir directement une description exacte des propriétés locales et globales utiles à l'analyse statistique. Le théorème suivant (qui correspond à la Proposition 2.1 du Chapitre 2) fournit en particulier la loi *exacte* de la cellule $C_\lambda(x)$ contenant un point $x \in [0,1]^d$ dans une partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$.

Théorème 1.15 (Proposition 2.1, Chapitre 2)**.** *Soit $x \in [0,1]^d$, et soit $C_\lambda(x)$ la cellule d'une partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ contenant x. Alors, $C_\lambda(x)$ a la même loi que*

$$\prod_{j=1}^{d} \left[(x_j - \lambda^{-1} E_{j,L}) \vee 0, (x_j + \lambda^{-1} E_{j,R}) \wedge 1 \right], \tag{1.113}$$

où $x = (x_j)_{1 \leqslant j \leqslant d}$, et où $E_{1,L}, E_{1,R}, \ldots, E_{d,L}, E_{d,R}$ sont des variables i.i.d. de loi $\mathsf{Exp}(1)$.

Pour les variantes de PRF considérées dans la littérature (Breiman, 2000; Biau et al., 2008; Arlot and Genuer, 2014; Scornet, 2016), la loi des cellules est en général complexe et n'admet pas de description aussi précise. En effet, le contrôle de la partition requiert de considérer l'effet des coupures successives, ce qui conduit à une analyse délicate. Dans le cas des partitions de Mondrian, la caractérisation de la loi des cellules s'obtient directement, en exploitant la propriété de restriction des partitions de Mondrian, sans avoir à raisonner conditionnellement à la structure combinatoire de l'arbre ou au nombre de coupures.

Une seconde quantité importante pour l'analyse est le nombre de cellules des partitions. Pour cette variante de forêts, il est également de possible de calculer l'espérance de cette quantité :

Proposition 1.10 (Proposition 2.2, Chapitre 2)**.** *Si $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$, alors le nombre $|\Pi_\lambda|$ de cellules dans la partition Π_λ satisfait :*

$$\mathbb{E}[|\Pi_\lambda|] = (1+\lambda)^d . \tag{1.114}$$

Une esquisse de la preuve de la Proposition 1.10 figure en Section 2.7, tandis que la preuve complète de ce résultat se trouve dans la Section 2.8.2. L'idée de la preuve consiste à construire une version modifiée du processus de Mondrian, pour laquelle l'espérance du nombre de coupures est inchangée, et dont les partitions sont des "produits" de partitions de Mondrian unidimensionnelles ; on conclut alors en utilisant le lien entre ces dernières et les processus de Poisson sur $[0,1]$.

Comme première application du Théorème 1.15 et de l'équation (1.114), nous obtenons une vitesse de convergence minimax sur la classe des fonctions Lipschitz :

Théorème 1.16 (Théorème 1.16, Chapitre 2). *Supposons la fonction de régression* $g^*(x) := \mathbb{E}[Y|X=x]$ *Lipschitz, et la variance conditionnelle* $\mathrm{Var}(Y|X)$ *bornée. Alors, l'estimateur des forêts de Mondrian* $\widehat{g}_{\lambda,M,n}$ *avec* $M \geqslant 1$ *arbres et de paramètre* $\lambda \asymp n^{1/(d+2)}$ *satisfait*

$$\mathbb{E}\big[(\widehat{g}_{\lambda,M,n}(X) - g^*(X))^2\big] = O(n^{-2/(d+2)}),$$

qui est la vitesse optimale au sens minimax sur cette classe de fonctions (Stone, 1982; Györfi et al., 2002).

En effet, le Théorème 1.15 implique que le diamètre de la cellule $C_\lambda(X)$ du point de test $X \sim P_X$ est d'ordre $O(1/\lambda)$, ce qui permet de contrôler le biais de l'estimateur $\widehat{g}_{\lambda,M,n}$. De plus, la formule (1.114) sur le nombre de cellules permet de contrôler la variance de cet estimateur. Le Théorème 1.16 en découle. Comme signalé précédemment, les variantes de forêts purement aléatoires considérées dans la littérature (telles les forêts centrées) n'atteignent pas la vitesse minimax $O(n^{-2/(d+2)})$ (Breiman, 2004; Biau, 2012; Arlot and Genuer, 2014; Klusowski, 2018), hormis en dimension 1 (Arlot and Genuer, 2014). Cela tient au choix uniforme de la coordonnée de coupure à chaque étape, qui conduit à des cellules déséquilibrées.

Le Théorème 1.16 est valable quel que soit le nombre d'arbres $M \geqslant 1$, et en particulier pour des arbres de Mondrian ($M = 1$). L'avantage des forêts par rapports aux arbres individuels se manifeste dans le cas d'une fonction de régression plus régulière que simplement Lipschitz :

Théorème 1.17 (Théorème 2.3, Chapitre 2). *Supposons que la fonction de régression* g^* *est de classe* $\mathscr{C}^2$, *que* $\mathrm{Var}(Y|X)$ *bornée, et que* X *admet une densité Lipschitz et positive. Alors, l'estimateur des forêts de Mondrian* $\widehat{g}_{\lambda,M,n}$ *avec* $\lambda \asymp n^{1/(d+4)}$ *et* $M \gtrsim n^{2/(d+4)}$ *satisfait*

$$\mathbb{E}\big[(\widehat{g}_{\lambda,M,n}(X) - g^*(X))^2 | X \in B_\varepsilon\big] = O(n^{-4/(d+4)}),$$

où $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$, *pour tout* $\varepsilon > 0$.

Par ailleurs, on vérifie directement pour $d = 1$ qu'un arbre de Mondrian calibré de manière optimale admet un risque d'au moins $\Theta(n^{-2/3})$, correspondant à la vitesse Lipschitz, même dans le cas $\mathscr{C}^2$ (Proposition 2.3). La vitesse en $O(n^{-4/(d+4)})$ correspond à la vitesse minimax d'estimation des fonctions de classe $\mathscr{C}^2$ (Györfi et al., 2002). La restriction à B_ε permet d'éviter un effet de bord, commun aux estimateurs par moyennes locales (Györfi et al., 2002; Wasserman, 2006) ; sans cette restriction, la vitesse obtenue est de $O(n^{-3/(d+3)})$, qui est plus lente mais meilleure que celle du cas Lipschitz.

Le Théorème 1.17 repose sur un contrôle plus précis du biais de forêts avec $M \gg 1$ (c'est-à-dire des forêts infinies) sous l'hypothèse $g^* \in \mathscr{C}^2([0,1]^d)$. Pour obtenir ce contrôle, le seul diamètre de $C_\lambda(x)$ (pour $x \in [0,1]^d$) n'est pas assez précis, nous utilisons donc la loi exacte de $C_\lambda(x)$ décrite par le Théorème 1.15. Par la propriété d'absence de mémoire des lois exponentielles, on déduit notamment de celle-ci la loi de $C_\lambda(x)$ conditionnellement à $z \in C_\lambda(x)$, pour tous $x, z \in [0,1]^d$. Le biais des forêts infinies s'étudie alors en considérant le "noyau" :

$$F_{p,\lambda}(x,z) = e^{-\lambda\|x-z\|_1}\,\mathbb{E}\left[\left\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\right\}^{-1}\right],$$

pour $x, z \in [0,1]^d$, où p désigne la densité de X, et où

$$C_\lambda(x,z) := \prod_{j=1}^{d} \left[(x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1\right]$$

avec $E_{1,L}, E_{1,R}, \ldots, E_{d,L}, E_{d,R}$ des variables i.i.d. de loi $\mathsf{Exp}(1)$.

Forêts de Mondrian en ligne par agrégation (Chapitre 3). Ce chapitre complète le précédent d'un point de vue méthodologique. Nous revenons à la motivation initiale des forêts de Mondrian (Lakshminarayanan et al., 2014), qui est de fournir un algorithme de forêt calculable en ligne, c'est-à-dire pouvant être mis à jour de manière efficace à l'arrivée d'un nouvel échantillon (X_t, Y_t). Comme indiqué précédemment, l'algorithme proposé par Lakshminarayanan et al. (2014) utilise les propriétés des processus de Mondrian afin d'obtenir un tel algorithme en ligne ; cet algorithme admet cependant deux limitations principales. D'une part, il ne s'agit pas d'un algorithme exact, mais d'une procédure qui calcule un estimateur bayésien sur chaque arbre de manière approchée. D'autre part, cette procédure n'admet pas de garantie théorique.

Nous introduisons un algorithme de forêts de Mondrian, constitué par un ensemble d'arbres individuels. Chaque arbre s'obtient en considérant une réalisation d'un processus de Mondrian $\Pi_\infty = (\Pi_\lambda)_{\lambda \in \mathbf{R}^+}$, dont la donnée équivaut à celle d'un arbre binaire complet infini, où le temps, la coordonnée et le seuil de coupure sont indiqués à chaque nœud et aléatoires. L'estimateur associé à un tel arbre infini correspond à l'agrégation à poids exponentiels de tous les prédicteurs constitués par des sous-arbres de décision finis T de Π_∞, avec une prédiction $\widehat{y}_{\mathbf{v}}$ associée à chaque feuille $\mathbf{v}$ de T. D'un point de vue computationnel, un tel estimateur pose a priori deux difficultés :

- il nécessite a priori de tirer une réalisation du processus de Mondrian infini, ce qui n'est pas possible avec des moyens finis ;
- même pour un arbre fini à n feuilles (par exemple avec un point par feuille), le nombre de sous-arbres est exponentiel en n ; le coût d'une agrégation à poids exponentiels naïve avec un poids par sous-arbre est donc prohibitif.

La première difficulté est levée en ne tirant que les coupures nécessaires à séparer les points du jeu de données ; une telle partition se met à jour en utilisant les propriétés de restriction des partitions de Mondrian (Lakshminarayanan et al., 2014). Afin de contourner la seconde difficulté, nous utilisons une loi a priori d'une certaine forme sur les sous-arbres (un processus de branchement), telle que le postérieur soit également de cette forme. Le calcul de l'agrégation à poids exponentiels se "factorise" alors, de sorte qu'il suffit de maintenir un poids par nœud plutôt qu'un poids par sous-arbre, ce qui permet une réduction exponentielle de la complexité. La procédure d'agrégation correspond alors à celle des *arbres experts* introduite par Helmbold and Schapire (1997), intimement liée à l'algorithme du *Context Tree Weighting* (Willems et al., 1995; Willems, 1998) de compression de données, qui repose sur la même factorisation.

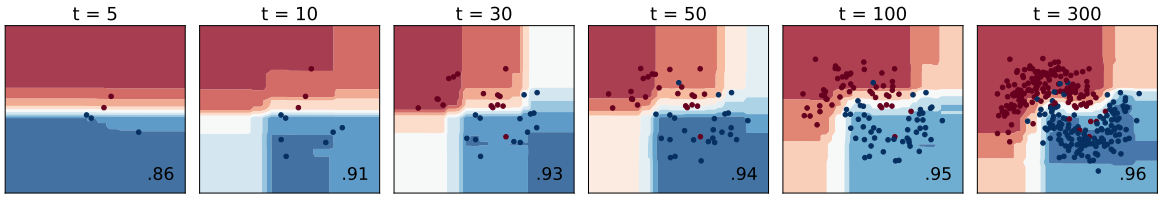


Figure 1.4: Évolution de la fonction de décision en classification, pour l'algorithme AMF introduit, en fonction du nombre de points (Chapitre 3).

L'estimateur $\widehat{g}_n$ ainsi obtenu satisfait la garantie de risque suivante, par exemple en régression bornée avec $\mathcal{Y} = [-B, B]$ (Exemple 1.4) : si Π est une réalisation aléatoire du processus

de Mondrian (c'est-à-dire un arbre infini, avec une coupure associée à chaque nœud), alors

$$\mathbb{E}[R(\widehat{g}_n)] \leqslant \mathbb{E}_\Pi\Big[\inf_{\mathcal{T},g_\mathcal{T}}\Big\{R(g_\mathcal{T}) + \frac{4B^2(|\mathcal{T}|+1)\log n}{n}\Big\}\Big] \tag{1.115}$$

où l'infimum porte sur les sous-arbres finis $\mathcal{T}$ (à $|\mathcal{T}|$ feuilles) de Π et les fonctions $g_\mathcal{T} : [0,1]^d \to \mathbf{R}$ constantes sur les cellules de $\mathcal{T}$, et l'espérance du membre de droite porte sur le tirage de Π. Ce résultat découle du Corollaire 3.2 et de la conversion online-to-batch. En particulier, en considérant les sous-arbres $\mathcal{T}$ de la forme Π_λ ($\lambda > 0$), et en utilisant les résultats du Chapitre 2, il est possible d'obtenir des vitesses de convergence non paramétriques sur les classes de régularité Hölder (Théorème 3.2). Nous renvoyons au Chapitre 3 pour des expériences numériques sur cette procédure ; bien que non compétitive avec les forêts de Breiman pour des tâches de classification pures, elle offre une bonne calibration des probabilités des différentes classes (mesurée par la perte logarithmique ou l'AUC), tout en étant implémentable de manière séquentielle et efficace, et moins sensible au nombre d'arbres.

Par rapport à l'estimateur des forêts de type PRF considéré au Chapitre 2, la régularisation n'est plus assurée par le paramètre λ de complexité des arbres (les arbres considérés ne sont pas élagués), mais par l'agrégation à poids exponentiels sur tous les sous-arbres. L'estimateur ainsi obtenu est donc compétitif avec le meilleur élagage $\mathcal{T}$ de l'arbre infini ; en pratique, cela permet de sélectionner des partitions plus adaptatives, qui approchent mieux la fonction de régression lorsque celle-ci varie davantage dans certaines régions que d'autres. Cependant, contrairement au cas des forêts de Mondrian de type PRF, l'estimateur n'admet a priori pas de vitesses minimax dans le cas $\mathscr{C}^2$: cela tient au fait que la complexité des arbres est optimisée individuellement pour chaque arbre. Cet estimateur peut être vu comme un algorithme de plus proches voisins, mais où le nombre de voisins dépend du point considéré et est choisi de manière adaptative.

1.6 Annexe technique

Dans cette section figurent des définitions et résultats techniques mentionnés ou utilisés dans cette introduction.

1.6.1 Variables sous-Gaussiennes et inégalités de concentration

Dans cette section, nous collectons quelques définitions et résultats élémentaires de concentration de variables i.i.d. ; nous renvoyons à Boucheron et al. (2013) pour davantage de détails sur ce sujet.

Nous commençons par rappeler l'inégalité de Hoeffding (Hoeffding, 1963), qui est utilisée dans la preuve de la Proposition 1.5.

Lemme 1.2 (Boucheron et al., 2013, Lemme 2.2)**.** *Soit X une variable aléatoire à valeurs dans $[0,1]$. Alors, pour tout $\lambda \in \mathbf{R}$, $\log \mathbb{E}[e^{\lambda X}] \leqslant \lambda \mathbb{E}[X] + \lambda^2/8$.*

Proof. Soit $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$ pour tout $\lambda \in \mathbf{R}$. Définissons pour $\lambda \in \mathbf{R}$ la mesure de probabilité $\mathbb{P}_\lambda := \{e^{\lambda X}/\mathbb{E}[e^{\lambda X}]\} \cdot \mathbb{P}$, et notons $\mathbb{E}_\lambda[Z]$, $\mathrm{Var}_\lambda(Z)$ l'espérance et la variance d'une variable aléatoire Z selon $\mathbb{P}_\lambda$. Par dérivation sous le signe intégral (en utilisant le fait que $0 \leqslant X \leqslant 1$), la fonction ψ est deux fois dérivable, et vérifie $\psi'(\lambda) = \mathbb{E}[Xe^{\lambda X}]/\mathbb{E}[e^{\lambda X}] = \mathbb{E}_\lambda[X]$ et $\psi''(\lambda) = \mathrm{Var}_\lambda(X)$.

Or, la variable X appartient à $[0,1]$ $\mathbb{P}$-presque sûrement, donc $\mathbb{P}_\lambda$-presque sûrement, de sorte que $\psi''(\lambda) = \mathrm{Var}_\lambda(X) \leqslant 1/4$. L'inégalité de Taylor montre alors que, pour tout $\lambda \in \mathbf{R}$, $\psi(\lambda) \leqslant \psi(0) + \psi'(0) \cdot \lambda + (1/4) \cdot \lambda^2/2$, ce qui établit le Lemme 1.2 puisque $\psi(0) = 0$ et $\psi'(0) = \mathbb{E}_0[X] = \mathbb{E}[X]$. □

L'inégalité de Hoeffding équivaut à dire que $X - \mathbb{E}[X]$ est $1/4$-sous-Gaussienne, au sens de la définition suivante :

Définition 1.6 (Variable sous-Gaussienne)**.** Soit $\sigma^2 > 0$. Une variable aléatoire réelle centrée X est dite *σ^2-sous-Gaussienne* si, pour tout $\lambda \in \mathbf{R}$, $\mathbb{E}[e^{\lambda X}] \leqslant e^{\sigma^2\lambda^2/2}$.

Le terme *sous-Gaussien* vient du fait que, si $X \sim \mathcal{N}(0, \sigma^2)$, alors $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$ pour tout $\lambda \in \mathbf{R}$. Si X est σ^2-sous-Gaussienne, alors pour tout $t \geqslant 0$,

$$\mathbb{P}(X \geqslant t) \vee \mathbb{P}(X \leqslant -t) \leqslant e^{-t^2/(2\sigma^2)}\,, \tag{1.116}$$

en appliquant l'inégalité de Markov aux variables $e^{\lambda X}$ et $e^{-\lambda X}$ et en optimisant le choix de λ. En outre, il existe une constante universelle C telle que pour tout $p \geqslant 1$:

$$\|X\|_{L^p} = \mathbb{E}[|X|^p]^{1/p} \leqslant C\sigma\sqrt{p}\,. \tag{1.117}$$

Réciproquement, pour une variable centrée X, les inégalités (1.116) et (1.117) impliquent que X est $c\sigma^2$-sous-Gaussiennes pour une certaine constante absolue c (Vershynin, 2012; Boucheron et al., 2013). La condition (1.117) permet d'étendre la définition de variables sous-Gaussiennes aux variables non centrées, et revient à dire que $X - \mathbb{E}[X]$ est $C'\sigma^2$-sous-Gaussienne et que $|\mathbb{E}[X]| \leqslant C''\sigma$.

Plus généralement, on dit qu'un vecteur aléatoire X à valeurs dans $\mathbf{R}^d$ est σ^2-sous-Gaussien ($\sigma > 0$) si $\langle \theta, X \rangle$ est σ^2-sous-Gaussien pour tout $\theta \in S^{d-1}$.

Si $X_1, \dots, X_M$ sont des variables centrées σ^2-sous-Gaussiennes, alors pour tout $\lambda > 0$:

$$\mathbb{E}[e^{\lambda \max(X_1,\dots,X_M)}] = \mathbb{E}[\max(e^{\lambda X_1}, \dots, e^{\lambda X_M})] \leqslant \sum_{i=1}^{M} \mathbb{E}[e^{\lambda X_i}] \leqslant M e^{\sigma^2\lambda^2/2}. \tag{1.118}$$

Par convexité de la fonction exponentielle, on en déduit que $\exp(\lambda \mathbb{E}[\max(X_1, \dots, X_M)]) \leqslant M e^{\sigma^2\lambda^2/2}$, c'est-à-dire (en optimisant λ) que $\mathbb{E}[\max(X_1, \dots, X_M)] \leqslant \sigma\sqrt{2\log M}$. De même, l'inégalité de Markov implique que, pour tout $t > 0$,

$$\mathbb{P}\big(\max(X_1, \dots, X_M) \geqslant \sigma\sqrt{2\log M} + \sigma t\big) \leqslant e^{-t^2/2}. \tag{1.119}$$

Enfin, soient $X_1, \dots, X_n$ des variables indépendantes centrées et σ^2-sous-Gaussiennes. Il découle de la Définition 1.6 que la moyenne $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ est σ^2/n-sous-Gaussienne. En particulier, si $(X_{i,j})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant M}$ sont des variables centrées à valeurs dans $[-1, 1]$, telles que les lignes $X_{i,\cdot} = (X_{i,j})_{1\leqslant j\leqslant M}$ sont indépendantes, alors en notant $\bar{X}_i := n^{-1}\sum_{i=1}^n X_{i,j}$, on a

$$\mathbb{E}\big[\max_{1\leqslant j\leqslant M} \bar{X}_j\big] \leqslant \sqrt{\frac{2\log M}{n}}, \quad \text{et} \quad \mathbb{P}\Big(\max_{1\leqslant j\leqslant M} \bar{X}_j \geqslant \sqrt{\frac{2\log M}{n}} + \sqrt{\frac{2t}{n}}\Big) \leqslant e^{-t} \tag{1.120}$$

pour tout $t > 0$. Ce résultat s'applique notamment à $X_{i,j} = \ell(f_j, Z_i) - R(f_j)$, où $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$ est une fonction de perte, $\mathcal{F} = \{f_1, \dots, f_M\}$ est une classe finie et $Z_1, \dots, Z_n$ sont des observations i.i.d., et permet dans ce cas de contrôler l'erreur de généralisation et l'excès de risque (Section 1.1.4).

1.6.2 Entropie relative et dualité

Nous décrivons à présent l'entropie relative ainsi qu'un résultat fondamental de dualité, qui est utilisé à plusieurs reprises dans ce texte.

Définition 1.7. Soit Θ un espace mesurable, et ρ, π deux mesures de probabilité sur Θ. L'*entropie relative*, ou *divergence de Kullback-Leibler* entre ρ et π, notée $\mathrm{KL}(\rho, \pi)$, est définie par

$$\mathrm{KL}(\rho, \pi) := \int_\Theta \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho \tag{1.121}$$

lorsque ρ est absolument continue par rapport à π, et $+\infty$ dans le cas contraire.

L'entropie relative $\mathrm{KL}(\rho, \pi)$ mesure la différence entre les mesures π et ρ, ou la qualité de π en tant qu'approximation de ρ ; cette quantité n'est pas symétrique en ρ, π. Plus précisément, l'entropie relative constitue l'excès de risque de π par rapport à ρ en estimation de densité avec perte logarithmique, lorsque la vraie loi est ρ (Exemple 1.2). Cette quantité est toujours positive, avec égalité si et seulement si $\pi = \rho$; cela se vérifie en appliquant l'inégalité de Jensen à la fonction $-\log$ et le fait que $\rho(\mathrm{d}\rho/\mathrm{d}\pi = 0) = 0$:

$$\begin{aligned}\mathrm{KL}(\rho, \pi) &= \int_\Theta -\log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\rho}\Big)\mathbf{1}\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} > 0\Big)\mathrm{d}\rho \\ &\geqslant -\log\Big(\int_\Theta \frac{\mathrm{d}\pi}{\mathrm{d}\rho}\mathbf{1}\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} > 0\Big)\mathrm{d}\rho\Big) \geqslant -\log\Big(\int_\Theta \mathrm{d}\pi\Big) = 0\end{aligned}$$

avec égalité si et seulement si $\mathrm{d}\rho/\mathrm{d}\pi \equiv 1$, c'est-à-dire $\rho = \pi$.

Notons également, pour toute mesure finie π sur Θ et toute fonction $f : \Theta \to \mathbf{R}$ mesurable, $\langle \pi, f \rangle := \int_\Theta f \mathrm{d}\pi$ dès lors que l'intégrale est bien définie dans $\mathbf{R} \cup \{+\infty, -\infty\}$. Pour toute fonction positive $h : \Theta \to \mathbf{R}^+$ telle que $\langle \pi, h \rangle \in (0, +\infty)$, notons $\pi_h := \frac{h}{\langle \pi, h \rangle} \cdot \pi$ la mesure de probabilité sur Θ de densité $h/\langle \pi, h \rangle$ par rapport à π.

Théorème 1.18 (Donsker-Varadhan). *Soit π, ρ deux mesures de probabilité sur Θ ; pour toute fonction bornée $f : \Theta \to \mathbf{R}$,*

$$\log \left(\int_\Theta \exp(f) \, \mathrm{d}\pi \right) + \mathrm{KL}(\rho, \pi) - \int_\Theta f \mathrm{d}\rho = \mathrm{KL}(\rho, \pi_{\exp(f)}) \, . \tag{1.122}$$

En particulier, pour toute fonction mesurable $f : \Theta \to \mathbf{R}$,

$$\log \langle \pi, \exp(f) \rangle = \sup_\rho \left\{ \langle \rho, f \rangle - \mathrm{KL}(\rho, \pi) \right\} , \tag{1.123}$$

le supremum étant atteint pour $\rho = \pi_{\exp(f)}$ lorsque le terme de gauche est fini.

Proof. Commençons par supposer f bornée, de sorte que les intégrales sont bien définies. Si ρ n'est pas absolument continue par rapport à π, il ne l'est pas non plus par rapport à $\pi_{\exp(f)} = [\exp(f)/\langle \pi, \exp(f) \rangle]\pi$, et donc les membres de gauche et de droite de (1.122) sont infinis. Si ρ est absolument continue par rapport à π, alors $\rho = [\mathrm{d}\rho/\mathrm{d}\pi]\pi = [\mathrm{d}\rho/\mathrm{d}\pi] \times [\exp(f)/\langle \pi, \exp(f) \rangle]^{-1} \pi_{\exp(f)}$, de sorte que

$$\begin{aligned} \mathrm{KL}(\rho, \pi_{\exp(f)}) &= \int_\Theta \log \Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \cdot \frac{\langle \pi, \exp(f) \rangle}{\exp(f)} \Big) \mathrm{d}\rho \\ &= \int_\Theta \log \Big(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \Big) \mathrm{d}\rho + \log \big(\langle \pi, \exp(f) \rangle \big) - \int_\Theta f \mathrm{d}\rho \, , \end{aligned}$$

qui coïncide précisément avec (1.122). Il en découle que, pour tous ρ, π, f (avec f bornée), par positivité de l'entropie relative,

$$\langle \rho, f \rangle - \mathrm{KL}(\rho, \pi) = \log \langle \pi, \exp(f) \rangle - \mathrm{KL}(\rho, \pi_{\exp(f)}) \leqslant \log \langle \pi, \exp(f) \rangle \, ,$$

avec égalité si et seulement si $\rho = \pi_{\exp(f)}$. Le résultat dans le cas général où f n'est pas bornée s'en déduit en considérant $f^B := \min(f, B)$ avec $B \to +\infty$. □

Le Théorème 1.18 affirme que, pour toute mesure de probabilité π, l'entropie relative $\rho \mapsto \mathrm{KL}(\rho, \pi)$ (définie sur l'espace des mesures de probabilité sur Θ, et à valeurs dans $[0, +\infty]$) est la transformée de Fenchel-Legendre (Boyd and Vandenberghe, 2004) de la transformée de Laplace logarithmique $f \mapsto \log \langle \pi, \exp(f) \rangle \in \mathbf{R} \cup \{+\infty\}$ (définie sur les fonctions $\Theta \to \mathbf{R}$). Ce résultat (ou une variante "inversée" de (1.123), qui découle de la même manière de (1.122)) est parfois appelé *formule variationnelle de Donsker-Varadhan.*

1.6.3 Convexité et forte convexité

Nous indiquons dans cette section la définition de la (forte) convexité (Boyd and Vandenberghe, 2004), ainsi que le lien entre la forte convexité et la stabilité des minima.

Définition 1.8 (Convexité, forte convexité)**.** Soit E un espace vectoriel normé (de norme notée $\|\cdot\|$), et $f : E \to \mathbf{R} \cup \{+\infty\}$. On dit que f est *convexe* si $\Omega := \{x \in E : f(x) \in \mathbf{R}\}$ est non vide et convexe, et si pour tous $x, x' \in \Omega$ et $t \in [0,1]$,

$$f\big(tx + (1-t)x'\big) \leqslant t \cdot f(x) + (1-t) \cdot f(x')\,.$$

Supposons également f différentiable sur Ω, et notons $\nabla f(x) \in E^*$ (où E^* est le dual de E) le gradient de f en x. Pour tout $\lambda > 0$, on dit que f est *λ-fortement convexe* si, pour tous $x, x' \in \Omega$,

$$f(x') - f(x) - \langle \nabla f(x), x' - x\rangle \geqslant \frac{\lambda}{2}\|x - x'\|^2\,. \tag{1.124}$$

La convexité est équivalente à la condition de λ-forte convexité avec $\lambda = 0$. On vérifie directement que la fonction $x \mapsto \|x\|^2/2$ est 1-fortement convexe sur $\mathbf{R}^d$ (la condition (1.124) est alors une égalité), et que la somme d'une fonction λ-fortement convexe et d'une fonction convexe est λ-fortement convexe (Boyd and Vandenberghe, 2004). Notons que, si $f : \Omega \to \mathbf{R}$ (où $\Omega \subset E$ est convexe) est différentiable et $x^* \in \arg\min f$, alors $\nabla f(x^*) = 0$. Réciproquement, si f est convexe et $\nabla f(x^*) = 0$, alors x^* est un minimiseur de f. Si de plus f est λ-fortement convexe, alors pour tout $x \in \Omega$ (puisque $\nabla f(x^*) = 0$) :

$$f(x) - f(x^*) \geqslant \frac{\lambda}{2}\|x - x^*\|^2\,. \tag{1.125}$$

De plus, en inversant x et x' dans l'inégalité (1.124) et en sommant l'inégalité obtenue avec (1.125), on obtient, pour tous $x, x' \in \Omega$,

$$\langle \nabla f(x') - \nabla f(x), x' - x\rangle \geqslant \lambda\|x - x'\|^2\,. \tag{1.126}$$

De ce qui précède, nous déduisons le résultat de stabilité suivant (rappelons qu'une fonction $f : \mathbf{R}^d \to \mathbf{R}$ est dite *L-Lipschitz* si $|f(x') - f(x)| \leqslant L\|x' - x\|$ pour tous $x, x' \in \mathbf{R}^d$) :

Lemme 1.3. *Soit $F : \mathbf{R}^d \to \mathbf{R}$ une fonction λ-fortement convexe, et $f : \mathbf{R}^d \to \mathbf{R}$ une fonction L-Lipschitz. Soient*[21] *$x^* = \arg\min F$, et $\widetilde{x} \in \arg\min(F + f)$. Alors*

$$\|x^* - \widetilde{x}\| \leqslant \frac{L}{\lambda}\,, \qquad f(x^*) - f(\widetilde{x}) \leqslant \frac{L^2}{\lambda}\,.$$

Proof. L'inégalité $(F + f)(\widetilde{x}) \leqslant (F + f)(x^*)$ et le caractère L-Lipschitz de f impliquent que

$$F(\widetilde{x}) - F(x^*) \leqslant f(x^*) - f(\widetilde{x}) \leqslant L\|\widetilde{x} - x^*\|\,.$$

De plus, par λ-forte convexité de F et par l'inégalité (1.125), on a $F(\widetilde{x}) - F(x^*) \geqslant \lambda\|\widetilde{x} - x^*\|^2/2$. Il en découle que $\|\widetilde{x} - x^*\| \leqslant 2L/\lambda$ et donc $f(x^*) - f(\widetilde{x}) \leqslant 2L^2/\lambda$. Cela établit la borne du Lemme 1.3, au facteur 2 près.

Pour éliminer le facteur 2, procédons comme suit. Pour toute fonction $g : \mathbf{R}^d \to \mathbf{R}$, soit $\phi_g : [0,1] \to \mathbf{R}$ la fonction définie par $\phi_g(t) = g(\widetilde{x} + t(x^* - \widetilde{x}))$. Par définition de $\widetilde{x}$, $\phi_{F+f} = \phi_F + \phi_f$ atteint son minimum en 0, de sorte que, pour tout $t \in (0,1]$,

$$0 \leqslant \frac{\phi_{F+f}(t) - \phi_{F+f}(0)}{t} = \Delta_F(t) + \Delta_f(t)\,, \tag{1.127}$$

[21] L'existence de $x^*, \widetilde{x}$ provient du fait que $F, F + f$ tendent vers $+\infty$ en $+\infty$ (par forte convexité de F et en utilisant le fait que f est Lipschitz), et que ces deux fonctions sont continues (une fonction convexe sur $\mathbf{R}^d$ étant continue). L'unicité de x^* provient de la forte convexité de F et de (1.125) ; $\widetilde{x}$ n'est pas nécessairement unique (il l'est cependant si f est convexe), mais le résultat s'applique à tout choix possible de $\widetilde{x}$.

où l'on note $\Delta_g(t) = (\phi_g(t) - \phi_g(0))/t$. Or, la fonction f est L-Lipschitz, de sorte que $\phi_f(t) - \phi_f(0) \leqslant L\|t \cdot (\widetilde{x} - x^*)\|$, c'est-à-dire $\Delta_f(t) \leqslant L\|\widetilde{x} - x^*\|$ pour tout $t \in (0,1)$. En outre, la fonction F est différentiable, donc ϕ_F également ; ainsi, lorsque $t \to 0^+$, $\Delta_F(t) \to \phi'_F(0) = \langle \nabla F(\widetilde{x}), x^* - \widetilde{x} \rangle$. Or, la λ-forte convexité de F implique, par l'inégalité (1.126), que

$$\langle \nabla F(\widetilde{x}), x^* - \widetilde{x} \rangle = -\langle \nabla F(\widetilde{x}) - \nabla F(x^*), \widetilde{x} - x^* \rangle \leqslant -\lambda \|\widetilde{x} - x^*\|^2 ,$$

où l'on a utilisé que $\nabla F(x^*) = 0$. Ainsi, en prenant $t \to 0^+$, l'inégalité (1.127) implique que $0 \leqslant -\lambda\|\widetilde{x} - x^*\|^2 + L\|\widetilde{x} - x^*\|$, c'est-à-dire $\|\widetilde{x} - x^*\| \leqslant L/\lambda$ et donc $f(x^*) - f(\widetilde{x}) \leqslant L^2/\lambda$ comme annoncé. □

1.6.4 Regret de la descente de gradient en ligne

Dans cette annexe, nous établissons une borne de regret pour l'algorithme de *descente de gradient en ligne* (en anglais *Online gradient descent*, OGD, Zinkevich, 2003). Cet algorithme coïncide avec l'algorithme de descente de gradient stochastique décrit dans la Proposition 1.2, mais considéré comme algorithme d'optimisation en ligne. Le cadre du problème est ici celui de l'optimisation convexe en ligne (voir la Section 1.2.1)

Proposition 1.11 (Zinkevich, 2003)**.** *Soit Θ une partie convexe fermée de $\mathbf{R}^d$. Supposons que, pour tout $t = 1, \dots, n$, la fonction $\ell_t : \Theta \to \mathbf{R}$ est convexe, différentiable et L-Lipschitz. Considérons l'algorithme de* descente de gradient en ligne projetée :

- $\widehat{\theta}_1 := \theta_1 \in \Theta$ *fixe ;*
- *pour $t = 1, \dots, n$, $\widehat{\theta}_{t+1} := \mathrm{proj}_\Theta(\widehat{\theta}_t - \eta \nabla \ell_t(\widehat{\theta}_t))$.*

Soit $B > 0$; posons $\eta = B/(L\sqrt{n})$, et notons $\Theta_B := \{\theta \in \Theta : \|\theta - \theta_1\| \leqslant B\}$. Alors, on a la borne de regret suivante :

$$\sum_{t=1}^n \ell_t(\widehat{\theta}_t) - \inf_{\theta \in \Theta_B} \sum_{t=1}^n \ell_t(\theta) \leqslant BL\sqrt{n} . \tag{1.128}$$

Proof. Posons $h_t := \nabla \ell_t(\widehat{\theta}_t)$ pour tout $t = 1, \dots, n$; puisque ℓ_t est L-Lipschitz, on a $\|h_t\| \leqslant L$. De plus, l'inégalité (1.27) implique qu'il suffit de contrôler le regret sur la suite de pertes linéaires $\langle h_t, \cdot \rangle$. Comme $\widehat{\theta}_{t+1} = \mathrm{proj}_\Theta(\widehat{\theta}_t - \eta h_t)$, on a pour tout $\theta \in \Theta$:

$$\|\widehat{\theta}_{t+1} - \theta\|^2 \leqslant \|\widehat{\theta}_t - \eta h_t\|^2 = \|\widehat{\theta}_t - \theta\|^2 - 2\eta \langle h_t, \widehat{\theta}_t - \theta \rangle + \eta^2 \|h_t\|^2$$

de sorte que

$$\langle h_t, \widehat{\theta}_t - \theta \rangle \leqslant \frac{1}{2\eta}\big(\|\widehat{\theta}_t - \theta\|^2 - \|\widehat{\theta}_{t+1} - \theta\|^2\big) + \frac{\eta}{2}\|h_t\|^2 .$$

En sommant l'inégalité précédente sur $t = 1, \dots, n$, on obtient pour tout $\theta \in \Theta_B$:

$$\sum_{t=1}^n \ell_t(\widehat{\theta}_t) - \sum_{t=1}^n \ell_t(\theta) \leqslant \sum_{t=1}^n \langle h_t, \widehat{\theta}_t - \theta \rangle \leqslant \frac{\|\theta_1 - \theta\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^n \|h_t\|^2 \leqslant \frac{B^2}{2\eta} + \frac{\eta L^2 n}{2} ,$$

qui vaut $BL\sqrt{n}$ pour $\eta = B/(L\sqrt{n})$. □

Part I

Mondrian Random forests: theory and methodology

Chapter 2

Minimax optimal rates for Mondrian trees and forests

Abstract. Introduced by Breiman (2001a), Random Forests are widely used classification and regression algorithms. While being initially designed as batch algorithms, several variants have been proposed to handle online learning. One particular instance of such forests is the *Mondrian Forest* Lakshminarayanan et al. (2014, 2016), whose trees are built using the so-called Mondrian process, therefore allowing to easily update their construction in a streaming fashion. In this chapter, we provide a thorough theoretical study of Mondrian Forests in a batch learning setting, based on new results about Mondrian partitions. Our results include consistency and convergence rates for Mondrian Trees and Forests, that turn out to be minimax optimal on the set of s-Hölder function with $s \in (0, 1]$ (for trees and forests) and $s \in (1, 2]$ (for forests only), assuming a proper tuning of their complexity parameter in both cases. Furthermore, we prove that an adaptive procedure (to the unknown $s \in (0, 2]$) can be constructed by combining Mondrian Forests with a standard model aggregation algorithm. These results are the first demonstrating that some particular random forests achieve minimax rates *in arbitrary dimension*. Owing to their remarkably simple distributional properties, which lead to minimax rates, Mondrian trees are a promising basis for more sophisticated yet theoretically sound random forests variants.

Contents

2.1 Introduction

Introduced by Breiman (2001a), *Random Forests* (RF) are state-of-the-art classification and regression algorithms that proceed by averaging the forecasts of a number of randomized decision trees grown in parallel. Many extensions of RF have been proposed to tackle quantile estimation problems (Meinshausen, 2006), survival analysis (Ishwaran et al., 2008) and ranking (Clémençon et al., 2013); improvements of original RF are provided in literature, to cite but a few, better sampling strategies (Geurts et al., 2006), new splitting methods (Menze et al., 2011) or Bayesian alternatives (Chipman et al., 2010). Despite their widespread use and remarkable success in practical applications, the theoretical properties of such algorithms are still not fully understood (for an overview of theoretical results on RF, see Biau and Scornet, 2016). As a result of the complexity of the procedure, which combines sampling steps and feature selection, Breiman's original algorithm has proved difficult to analyze. A recent line of research (Scornet et al., 2015; Wager and Walther, 2015; Mentch and Hooker, 2016; Cui et al., 2017; Wager and Athey, 2018; Athey et al., 2019) has sought to obtain some theoretical guarantees for RF variants that closely resembled the algorithm used in practice. It should be noted, however, that most of these theoretical guarantees only offer limited information on the quantitative behavior of the algorithm (guidance for parameter tuning is scarce) or come at the price of conjectures on the true behavior of the RF algorithm itself, being thus still far from explaining the excellent empirical performance of it.

In order to achieve a better understanding of the random forest algorithm, another line of research focuses on modified and stylized versions of RF. Among these methods, *Purely Random Forests* (PRF) (Breiman, 2000; Biau et al., 2008; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Klusowski, 2018) grow the individual trees independently of the sample, and are thus particularly amenable to theoretical analysis. The consistency of such algorithms (as well as other idealized RF procedures) was first obtained by Biau et al. (2008), as a byproduct of the consistency of individual tree estimates. These results aim at quantifying the performance guarantees by analyzing the bias/variance of simplified versions of RF, such as PRF models (Genuer, 2012; Arlot and Genuer, 2014). In particular, Genuer (2012) shows that some PRF variant achieves the minimax rate for the estimation of a Lipschitz regression function in dimension one. The bias-variance analysis is extended by Arlot and Genuer (2014), showing that PRF can also achieve minimax rates for $\mathscr{C}^2$ regression functions in dimension one. These results are much more precise than mere consistency, and offer insights on the proper tuning of the procedure. Quite surprisingly, these optimal rates are only obtained in the one-dimensional case (where decision trees reduce to histograms). In the multi-dimensional setting, where trees exhibit an intricate recursive structure, only suboptimal rates are derived. As shown by lower bounds from Klusowski (2018), this is not merely a limitation from the analysis: centered forests, a standard variant of PRF, exhibit suboptimal rates under nonparametric assumptions.

From a more practical perspective, an important limitation of the most commonly used RF algorithms, such as Breiman's Random Forests (Breiman, 2001a) and the Extra-Trees algorithm (Geurts et al., 2006), is that they are typically trained in a batch manner, where the whole dataset, available at once, is required to build the trees. In order to allow their use in situations where large amounts of data have to be analyzed in a streaming fashion, several online variants of decision trees and RF algorithms have been proposed (Domingos

and Hulten, 2000; Saffari et al., 2009; Taddy et al., 2011; Denil et al., 2013, 2014).

Of particular interest in this article is the *Mondrian Forest* (MF) algorithm, an efficient and accurate online random forest classifier introduced by Lakshminarayanan et al. (2014), see also Lakshminarayanan et al. (2016). This algorithm is based on the Mondrian process (Roy and Teh, 2009; Roy, 2011; Orbanz and Roy, 2015), a natural probability distribution on the set of recursive partitions of the unit cube $[0,1]^d$. An appealing property of Mondrian processes is that they can be updated in an online fashion. In Lakshminarayanan et al. (2014), the use of the *conditional Mondrian* process enables the authors to design an online algorithm which matches its batch counterpart: training the algorithm one data point at a time leads to the same randomized estimator as training the algorithm on the whole dataset at once. The algorithm proposed in Lakshminarayanan et al. (2014) depends on a lifetime parameter $\lambda > 0$ that guides the complexity of the trees by stopping their building process. However, a theoretical analysis of MF is lacking, in particular, the tuning of λ is unclear from a theoretical perspective. In this chapter, we show that, aside from their appealing computational properties, Mondrian Forests are amenable to a precise theoretical analysis. We study MF in a batch setting and provide theoretical guidance on the tuning of λ.

Based on a detailed analysis of Mondrian partitions, we prove consistency and convergence rates for MF *in arbitrary dimension*, that turn out to be minimax optimal on the set of s-Hölder function with $s \in (0,2]$, assuming that λ and the number of trees in the forest (for $s \in (1,2]$) are properly tuned. Furthermore, we construct a procedure that adapts to the unknown smoothness $s \in (0,2]$ by combining Mondrian Forests with a standard model aggregation algorithm. To the best of our knowledge, such results have only been proved for very specific purely random forests, where the covariate space is of dimension one (Arlot and Genuer, 2014). Our analysis also sheds light on the benefits of Mondrian Forests compared to single Mondrian Trees: the bias reduction of Mondrian Forests allow them to be minimax for $s \in (1,2]$, while a single tree fails to be minimax in this case.

Agenda. This chapter is organized as follows. In Section 2.2, we describe the considered setting and set the notations for trees and forests. Section 2.3 defines the Mondrian process introduced by Roy and Teh (2009) and describes the MF algorithm. Section 2.4 provides new sharp properties for Mondrian partitions: cells distribution in Proposition 2.1 and a control of the cells diameter in Corollary 2.1, while the expected number of cells is provided in Proposition 2.2. Building on these properties, we provide, in Section 2.5, statistical guarantees for MF: Theorem 2.1 proves consistency, while Theorems 2.2 and 2.3 provide minimax rates for $s \in (0,1]$ and $s \in (1,2]$ respectively. Finally, Proposition 2.4 proves that a combination of MF with a model aggregation algorithm adapts to the unknown smoothness $s \in (0,2]$.

2.2 Setting and notations

We first describe the setting of the chapter and set the notations related to the Mondrian tree structure. For the sake of conciseness, we consider the regression setting, and show how to extend the results to classification in Section 2.5.5.

Setting. We consider a regression framework, where the dataset $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consists of i.i.d. $[0,1]^d \times \mathbf{R}$-valued random variables. We assume throughout the chapter that

the dataset is distributed as a generic pair (X, Y) such that $\mathbb{E}[Y^2] < \infty$. This unknown distribution, characterized by the distribution μ of X on $[0,1]^d$ and by the conditional distribution of $Y|X$, can be written as

$$Y = f(X) + \varepsilon, \tag{2.1}$$

where $f(X) = \mathbb{E}[Y|X]$ is the conditional expectation of Y given X, and ε is a noise satisfying $\mathbb{E}[\varepsilon|X] = 0$. Our goal is to output a *randomized estimate* $\widehat{f}_n(\cdot, Z, \mathscr{D}_n) : [0,1]^d \to \mathbf{R}$, where Z is a random variable that accounts for the randomization procedure. To simplify notation, we will denote $\widehat{f}_n(x, Z) = \widehat{f}_n(x, Z, \mathscr{D}_n)$. The quality of a randomized estimate $\widehat{f}_n$ is measured by its quadratic risk

$$R(\widehat{f}_n) = \mathbb{E}[(\widehat{f}_n(X, Z) - f(X))^2]$$

where the expectation is taken with respect to $(X, Z, \mathscr{D}_n)$. We say that a sequence $(\widehat{f}_n)_{n \geqslant 1}$ is *consistent* whenever $R(\widehat{f}_n) \to 0$ as $n \to \infty$.

Trees and Forests. A regression tree is a particular type of partitioning estimate. First, a recursive partition Π of $[0,1]^d$ is built by performing successive axis-aligned splits (see Section 2.3), then the regression tree prediction is computed by averaging the labels Y_i of observations falling in the same cell as the query point $x \in [0,1]^d$, that is

$$\widehat{f}_n(x, \Pi) = \sum_{i=1}^{n} \frac{\mathbf{1}(X_i \in C_\Pi(x))}{N_n(C_\Pi(x))} Y_i, \tag{2.2}$$

where $C_\Pi(x)$ is the cell of the tree partition containing x and $N_n(C_\Pi(x))$ is the number of observations falling into $C_\Pi(x)$, with the convention that the estimate returns 0 if the cell $C_\Pi(x)$ is empty.

A random forest estimate is obtained by averaging the predictions of M randomized decision trees; more precisely, we will consider purely random forests, where the randomization of each tree (denoted above by Z) comes exclusively from the random partition, which is independent of $\mathscr{D}_n$. Let $\Pi_M = (\Pi^{(1)}, \ldots, \Pi^{(M)})$, where $\Pi^{(m)}$ (for $m = 1, \ldots, M$) are i.i.d. random partitions of $[0,1]^d$. The random forest estimate is thus defined as

$$\widehat{f}_{n,M}(x, \Pi_M) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_n(x, \Pi^{(m)}), \tag{2.3}$$

where $\widehat{f}_n(x, \Pi^{(m)})$ is the prediction, at point x, of the tree with random partition $\Pi^{(m)}$, defined in (2.2).

The Mondrian Forest, whose construction is described below, is a particular instance of (2.3), in which the Mondrian process plays a crucial role by specifying the randomness Π of tree partitions.

2.3 The Mondrian Forest algorithm

Given a rectangular box $C = \prod_{j=1}^{d} [a_j, b_j] \subseteq \mathbf{R}^d$, we denote $|C| := \sum_{j=1}^{d} (b_j - a_j)$ its *linear dimension*. The Mondrian process $\mathsf{MP}(C)$ is a distribution on (infinite) tree partitions of C introduced by Roy and Teh (2009), see also Roy (2011) for a rigorous construction. Mondrian

partitions are built by iteratively splitting cells at some random time, which depends on the linear dimension of the cell; the splitting probability on each side is proportional to the side length of the cell, and the position is drawn uniformly.

The Mondrian process distribution $\mathsf{MP}(\lambda, C)$ is a distribution on tree partitions of C, resulting from the pruning of partitions drawn from $\mathsf{MP}(C)$. The pruning is done by removing all splits occurring after time $\lambda > 0$. In this perspective, λ is called the lifetime parameter and controls the complexity of the partition: large values of λ corresponds to deep trees (complex partitions).

Sampling from the distribution $\mathsf{MP}(\lambda, C)$ can be done efficiently by applying the recursive procedure $\texttt{SampleMondrian}(C, \tau = 0, \lambda)$ described in Algorithm 1. Figure 2.1 below shows a particular instance of Mondrian partition on a square box, with lifetime parameter $\lambda = 3.4$. In what follows, $\mathsf{Exp}(\lambda)$ stands for the exponential distribution with intensity $\lambda > 0$.

Algorithm 1 $\texttt{SampleMondrian}(C, \tau, \lambda)$: samples a Mondrian partition of C, starting from time τ and until time λ.

Inputs: A cell $C = \prod_{1 \leqslant j \leqslant d} [a_j, b_j]$, starting time τ and lifetime parameter λ.
Sample a random variable $E_C \sim \mathsf{Exp}(|C|)$
if $\tau + E_C \leqslant \lambda$ **then**
    Sample a split dimension $J \in \{1, \ldots, d\}$, with $\mathbb{P}(J = j) = (b_j - a_j)/|C|$
    Sample a split threshold S_J uniformly in $[a_J, b_J]$
    Split C along the split (J, S_J): let $C_0 = \{x \in C : x_J \leqslant S_J\}$ and $C_1 = C \setminus C_0$
    return $\texttt{SampleMondrian}(C_0, \tau + E_C, \lambda) \cup \texttt{SampleMondrian}(C_1, \tau + E_C, \lambda)$
else
    return $\{C\}$ (*i.e.*, do not split C).
end if

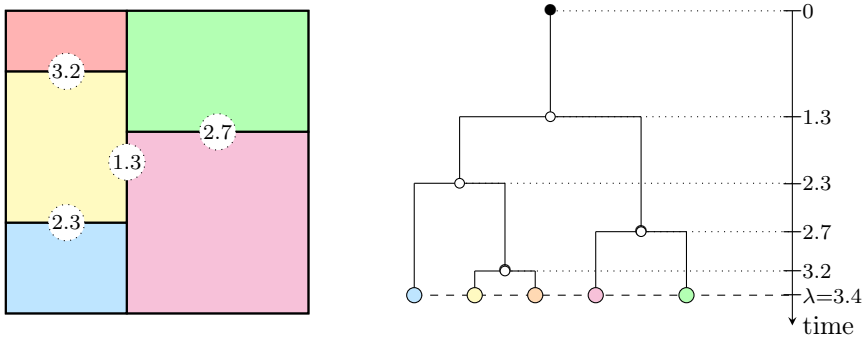


Figure 2.1: A Mondrian partition (left) with corresponding tree structure (right), which shows the evolution of the tree over time. The split times are indicated on the vertical axis, while the splits are denoted with bullets ($\circ$).

Remark 2.1. Using the fact that Exp is memoryless (if $E \sim \mathsf{Exp}(\lambda)$ and $u > 0$ then $E - u | E > u \sim \mathsf{Exp}(\lambda)$), it is possible to efficiently sample $\Pi_{\lambda'} \sim \mathsf{MP}(\lambda', C)$ given its pruning $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$ at time $\lambda \leqslant \lambda'$.

A Mondrian Tree estimator is given by Equation (2.2) where the partition $\Pi^{(m)}$ is sampled from the distribution $\mathsf{MP}(\lambda, [0,1]^d)$. The Mondrian Forest grows randomized tree partitions

$\Pi_\lambda^{(1)}, \ldots, \Pi_\lambda^{(M)}$, fits each one with the dataset $\mathscr{D}_n$ by averaging the labels falling into each leaf, then combines the resulting Mondrian Tree estimates by averaging their predictions. In accordance with Equation (2.3), we let

$$\widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)}) \tag{2.4}$$

be the Mondrian Forest estimate described above, where $\widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)})$ denotes the Mondrian Tree based on the random partition $\Pi_\lambda^{(m)}$ and $\Pi_{\lambda,M} = (\Pi_\lambda^{(1)}, \ldots, \Pi_\lambda^{(M)})$. To ease notation, we will write $\widehat{f}_{\lambda,n}^{(m)}(x)$ instead of $\widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)})$. Although we use the standard definition of Mondrian processes, the way we compute the prediction in a Mondrian Tree differs from the original one. Indeed, in Lakshminarayanan et al. (2014), prediction is given by the expectation over a posterior distribution, where a hierarchical prior is assumed on the label distribution of each cell of the tree. In this chapter, we simply compute the average of the observations falling into a given cell.

2.4 Local and global properties of the Mondrian process

In this Section, we show that the properties of the Mondrian process enable us to compute explicitly some local and global quantities related to the structure of Mondrian partitions. To do so, we will need the following two facts, exposed by Roy and Teh (2009).

Fact 2.1 (Dimension 1). *For $d = 1$, the splits from a Mondrian process $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1])$ form a subset of $[0,1]$, which is distributed as a Poisson point process of intensity $\lambda \mathrm{d}x$.*

Fact 2.2 (Restriction). *Let $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ be a Mondrian partition, and $C = \prod_{j=1}^d [a_j, b_j] \subset [0,1]^d$ be a box. Consider the* restriction $\Pi_\lambda|_C$ *of Π_λ on C, i.e. the partition on C induced by the partition Π_λ of $[0,1]^d$. Then $\Pi_\lambda|_C \sim \mathsf{MP}(\lambda, C)$.*

Fact 2.1 deals with the one-dimensional case by making explicit the distribution of splits for Mondrian process, which follows a Poisson point process. The restriction property stated in Fact 2.2 is fundamental, and enables one to precisely characterize the behavior of the Mondrian partitions.

Given any point $x \in [0,1]^d$, Proposition 2.1 below is a sharp result giving the exact distribution of the cell $C_\lambda(x)$ containing x from the Mondrian partition. Such a characterization is typically unavailable for other randomized trees partitions involving a complex recursive structure.

Proposition 2.1 (Cell distribution). *Let $x \in [0,1]^d$ and denote by*

$$C_\lambda(x) = \prod_{1 \leqslant j \leqslant d} [L_{j,\lambda}(x), R_{j,\lambda}(x)]$$

the cell containing x in a partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ (this cell corresponds to a leaf). Then, the distribution of $C_\lambda(x)$ is characterized by the following properties:

(*i*) *$L_{1,\lambda}(x), R_{1,\lambda}(x), \ldots, L_{d,\lambda}(x), R_{d,\lambda}(x)$ are independent*;

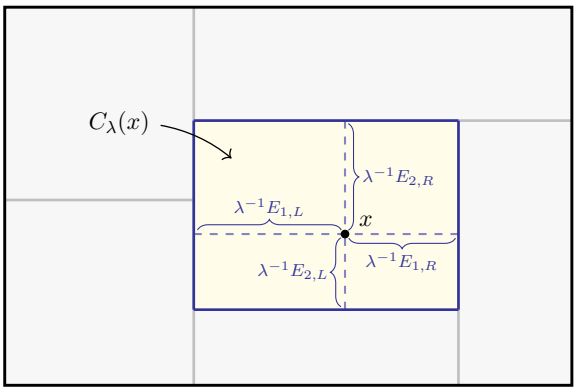


Figure 2.2: Cell distribution in a Mondrian partition (Proposition 2.1).

(*ii*) *For each* $j = 1, \dots, d$, $L_{j,\lambda}(x)$ *is distributed as* $(x - \lambda^{-1}E_{j,L}) \vee 0$ *and* $R_{j,\lambda}(x)$ *as* $(x + \lambda^{-1}E_{j,R}) \wedge 1$, *where* $E_{j,L}, E_{j,R} \sim \mathsf{Exp}(1)$.

The proof of Proposition 2.1 is given in Section 2.7. Figure 2.2 is a graphical representation of Proposition 2.1. A consequence of Proposition 2.1 is the next Corollary 2.1, which gives a precise upper bound on the diameter of the cells. In particular, this result is used in the proofs of the theoretical guarantees for Mondrian Trees and Forests from Section 2.5 below.

Corollary 2.1 (Cell diameter)**.** *Set* $\lambda > 0$ *and* $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ *be a Mondrian partition. Let* $x \in [0,1]^d$ *and let* $D_\lambda(x)$ *be the* ℓ^2*-diameter of the cell* $C_\lambda(x)$ *containing* x *in* Π_λ. *For every* $\delta > 0$, *we have*

$$\mathbb{P}(D_\lambda(x) \geqslant \delta) \leqslant d\Big(1 + \frac{\lambda\delta}{\sqrt{d}}\Big) \exp\Big(-\frac{\lambda\delta}{\sqrt{d}}\Big) \tag{2.5}$$

and

$$\mathbb{E}\big[D_\lambda(x)^2\big] \leqslant \frac{4d}{\lambda^2}\,. \tag{2.6}$$

In order to control the risk of Mondrian Trees and Forests, we need an upper bound on the number of cells in a Mondrian partition. Quite surprisingly, the expectation of this quantity can be computed exactly, as shown in Proposition 2.2.

Proposition 2.2 (Number of cells)**.** *Set* $\lambda > 0$ *and* $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ *be a Mondrian partition. If* K_λ *denotes the number of cells in* Π_λ, *we have* $\mathbb{E}[K_\lambda] = (1+\lambda)^d$.

The proof of Proposition 2.2 is given in Section 2.8.2, while a sketch of proof is provided in Section 2.7. Although the proof is technically involved, it relies on a natural coupling argument: we introduce a recursive modification of the construction of the Mondrian process which keeps the expected number of leaves unchanged, and for which this quantity can be computed directly using the Mondrian-Poisson equivalence in dimension one (Fact 2.1). A much simpler result is $\mathbb{E}[K_\lambda] \leqslant (e(1+\lambda))^d$, which was previously obtained in Mourtada et al. (2017). By contrast, Proposition 2.2 provides the *exact* value of this expectation, which removes a superfluous e^d factor.

Remark 2.2. Proposition 2.2 naturally extends (with the same proof) to the more general case of a Mondrian process with finite measures with no atoms $\nu_1, \dots, \nu_d$ on the sides $C^1, \dots, C^d$ of a box $C \subseteq \mathbf{R}^d$ (for a definition of the Mondrian process in this more general case, see Roy, 2011). In this case, we have $\mathbb{E}[K_\lambda] = \prod_{1 \leqslant j \leqslant d}(1 + \nu_j(C^j))$.

As illustrated in this Section, a remarkable fact with the Mondrian Forest is that the quantities of interest for the statistical analysis of the algorithm can be made explicit. In particular, we have seen in this Section that, roughly speaking, a Mondrian partition is balanced enough that it contains $O(\lambda^d)$ cells of diameter $O(1/\lambda)$, which is the minimal number of cells to cover $[0,1]^d$.

2.5 Minimax theory for Mondrian Forests

This Section gathers several theoretical guarantees for Mondrian Trees and Forests. Section 2.5.1 states the universal consistency of the procedure, provided that the lifetime λ_n belongs to an appropriate range. We provide convergence rates which turn out to be minimax optimal for s-Hölder regression functions with $s \in (0,1]$ in Section 2.5.2 and with $s \in (1,2]$ in Section 2.5.3, provided in both cases that λ_n is properly tuned. Note that in particular, we illustrate in Section 2.5.3 the fact that Mondrian Forests improve over Mondrian trees, when $s \in (1,2]$. In Section 2.5.4, we prove that a combination of MF with a model aggregation algorithm adapts to the unknown $s \in (0,2]$. Finally, results for classification are given in Section 2.5.5.

2.5.1 Consistency of Mondrian Forests

The consistency of the Mondrian Forest estimator is established in Theorem 2.1 below, assuming a proper tuning of the lifetime parameter λ_n.

Theorem 2.1 (Universal consistency)**.** *Let $M \geqslant 1$. Consider Mondrian Trees $\widehat{f}^{(m)}_{\lambda_n,n}$ (for $m = 1,\dots,M$) and Mondrian Forest $\widehat{f}_{\lambda_n,n,M}$ given by Equation (2.4) for a sequence $(\lambda_n)_{n\geqslant 1}$ satisfying $\lambda_n \to \infty$ and $\lambda_n^d/n \to 0$. Then, under the setting described in Section 2.2 above, the individual trees $\widehat{f}^{(m)}_{\lambda_n,n}$ (for $m = 1,\dots,M$) are consistent, and as a consequence, the forest $\widehat{f}_{\lambda_n,n,M}$ is consistent for any $M \geqslant 1$.*

The proof of Theorem 2.1 is given in Section 2.8.3. It uses the properties of Mondrian partitions established in Section 2.4 together with general consistency results for histograms. This result is universal, in the sense that it makes no assumption on the joint distribution of (X,Y), apart from $\mathbb{E}[Y^2] < \infty$ in order to ensure that the quadratic risk is well-defined (see Section 2.2).

The only tuning parameter of a Mondrian Tree is the lifetime λ_n, which encodes the complexity of the trees. Requiring an assumption on this parameter is natural, and confirmed by the well-known fact that the tree-depth is an important tuning parameter for Random Forests, see Biau and Scornet (2016). However, Theorem 2.1 does not address the question of a theoretically optimal tuning of λ_n under additional assumptions on the regression function f, which we consider in the following sections.

2.5.2 Mondrian Trees and Forests are minimax over s-Hölder classes for $s \in (0,1]$

The bounds obtained in Corollary 2.1 and Proposition 2.2 are explicit and sharp in their dependency on λ. Based on these properties, we now establish a theoretical upper bound on the risk of Mondrian Trees, which gives the optimal theoretical tuning of the lifetime parameter λ_n. To pursue the analysis, we need the following assumption.

Assumption 2.1. Consider (X, Y) from the setting described in Section 2.2 and assume also that $\mathbb{E}[\varepsilon|X] = 0$ and $\mathrm{Var}(\varepsilon|X) \leqslant \sigma^2 < \infty$ almost surely, where ε is given by Equation (2.1).

Our minimax results hold for a class of s-Hölder regression functions defined below.

Definition 2.1. Let $p \in \mathbf{N}$, $\beta \in (0, 1]$ and $L > 0$. The (p, β)-Hölder ball of norm L, denoted $\mathscr{C}^{p,\beta}(L) = \mathscr{C}^{p,\beta}([0,1]^d, L)$, is the set of p times differentiable functions $f : [0,1]^d \to \mathbf{R}$ such that

$$\|\nabla^p f(x) - \nabla^p f(x')\| \leqslant L\|x - x'\|^\beta \quad \text{and} \quad \|\nabla^k f(x)\| \leqslant L$$

for every $x, x' \in [0,1]^d$ and $k \in \{1, \dots, p\}$. Whenever $f \in \mathscr{C}^{p,\beta}(L)$, we say that f is s-Hölder with $s = p + \beta$.

Note that in what follows we will assume $s \in (0, 2]$, so that $p \in \{0, 1\}$. Theorem 2.2 below states an upper bound on the risk of Mondrian Trees and Forests, which explicitly depends on the lifetime parameter λ. Selecting λ that minimizes this bound leads to a convergence rate which turns out to be minimax optimal over the class of s-Hölder functions for $s \in (0, 1]$ (see for instance Stone, 1982, Chapter I.3 in Nemirovski, 2000 or Theorem 3.2 in Györfi et al., 2002).

Theorem 2.2. *Grant Assumption 2.1 and assume that $f \in \mathscr{C}^{0,\beta}(L)$, where $\beta \in (0, 1]$ and $L > 0$. Let $M \geqslant 1$. The quadratic risk of the Mondrian Forest $\widehat{f}_{\lambda,n,M}$ with lifetime parameter $\lambda > 0$ satisfies*

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2\big] \leqslant \frac{(4d)^\beta L^2}{\lambda^{2\beta}} + \frac{(1+\lambda)^d}{n}\big(2\sigma^2 + 9\|f\|_\infty^2\big). \tag{2.7}$$

In particular, as $n \to \infty$, the choice $\lambda := \lambda_n \asymp L^{2/(d+2\beta)} n^{1/(d+2\beta)}$ gives

$$\mathbb{E}\big[(\widehat{f}_{\lambda_n,n,M}(X) - f(X))^2\big] = O(L^{2d/(d+2\beta)} n^{-2\beta/(d+2\beta)}), \tag{2.8}$$

which corresponds to the minimax rate over the class $\mathscr{C}^{0,\beta}(L)$.

The proof of Theorem 2.2 is given in Section 2.7. It relies on the properties about Mondrian partitions stated in Section 2.4. Namely, Corollary 2.1 allows to control the bias of Mondrian Trees (first term on the right-hand side of Equation 2.7), while Proposition 2.2 helps in controlling the variance of Mondrian Trees (second term on the right-hand side of Equation 2.7).

To the best of our knowledge, Theorem 2.2 is the first to prove that a purely random forest (Mondrian Forest in this case) can be minimax optimal *in arbitrary dimension*. Minimax optimal upper bounds are obtained for $d = 1$ in Genuer (2012) and Arlot and Genuer (2014) for models of purely random forests such as Toy-PRF (where the individual partitions correspond to random shifts of the regular partition of $[0, 1]$ in k intervals) and PURF (Purely Uniformly Random Forests, where the partitions are obtained by drawing k random thresholds uniformly in $[0, 1]$). However, for $d = 1$, tree partitions reduce to partitions of $[0, 1]$ in intervals, and do not possess the recursive structure that appears in higher dimensions, which makes their analysis challenging. For this reason, the analysis of purely random forests for $d > 1$ has typically produced sub-optimal results: for example, Biau (2012) exhibit an upper bound on the risk of the centered random forests (a particular instance of PRF) which turns out to be much slower than the minimax rate for Lipschitz regression functions. A more in-depth

analysis of the same random forest model in Klusowski (2018) exhibits a new upper and lower bound of the risk, which is still slower than minimax rates for Lipschitz functions. A similar result was proved by Arlot and Genuer (2014), who studied the BPRF (Balanced Purely Random Forests algorithm, where all leaves are split, so that the resulting tree is complete), and obtained suboptimal rates. In our approach, the convenient properties of the Mondrian process enable us to bypass the inherent difficulties met in previous attempts. One specificity of Mondrian forests compared to other PRF variants is that the largest sides of cells are more likely to be split. By contrast, variants of PRF (such as centered forests) where the coordinate of the split is chosen with equal probability, may give rise to unbalanced cells with large diameter.

Theorem 2.2 provides theoretical guidance on the choice of the lifetime parameter, and suggests to set $\lambda := \lambda_n \asymp n^{1/(d+2)}$. Such an insight cannot be gleaned from an analysis that focuses on consistency alone. Theorem 2.2 is valid for Mondrian Forests with any number of trees, and thus in particular for a Mondrian Tree (this is also true for Theorem 2.1). However, it is a well-known fact that forests outperform single trees in practice (Fernández-Delgado et al., 2014). Section 2.5.3 proposes an explanation for this phenomenon, by assuming $f \in \mathscr{C}^{1,\beta}(L)$.

2.5.3 Improved rates for Mondrian Forests compared to a Mondrian Tree

The convergence rate stated in Theorem 2.2 for $f \in \mathscr{C}^{0,\beta}(L)$ is valid for both trees and forests, and the risk bound does not depend on the number M of trees that compose the forest. In practice, however, forests exhibit much better performances than individual trees. In this Section, we provide a result that illustrates the benefits of forests over trees by assuming that $f \in \mathscr{C}^{1,\beta}(L)$. As the counterexample in Proposition 2.3 below shows, single Mondrian trees do not benefit from this additional smoothness assumption, and achieve the same rate as in the Lipschitz case. This comes from the fact that the bias of trees is highly sub-optimal for such functions.

Proposition 2.3. *Assume that $Y = f(X) + \varepsilon$ with $f(x) = 1 + x$, where $X \sim \mathcal{U}([0,1])$ and ε is independent of X with variance σ^2. Consider a single Mondrian Tree estimate $\widehat{f}^{(1)}_{\lambda,n}$. Then, there exists a constant $C_0 > 0$ such that*

$$\inf_{\lambda \in \mathbf{R}_+^*} \mathbb{E}\big[(\widehat{f}^{(1)}_{\lambda,n}(X) - f(X))^2\big] \geqslant C_0 \wedge \frac{1}{4}\Big(\frac{3\sigma^2}{n}\Big)^{2/3}$$

for any $n \geqslant 18$.

The proof of Proposition 2.3 is given in Section 2.8.4. Since the minimax rate over $\mathscr{C}^{1,1}$ in dimension 1 is $O(n^{-4/5})$, Proposition 2.3 proves that a single Mondrian Tree is not minimax optimal over this function class. However, it turns out that large enough Mondrian Forests, which average Mondrian trees, are minimax optimal over $\mathscr{C}^{1,1}$. Therefore, Theorem 2.3 below highlights the benefits of a forest compared to a single tree.

Theorem 2.3. *Grant Assumption 2.1 and assume that $f \in \mathscr{C}^{1,\beta}(L)$, with $\beta \in (0,1]$ and $L > 0$. In addition, assume that X has a positive and C_p-Lipschitz density p w.r.t the Lebesgue measure on $[0,1]^d$. Let $\widehat{f}_{\lambda,n,M}$ be the Mondrian Forest estimate given by (2.4). Set $\varepsilon \in (0, 1/2)$*

and $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$. *Then, we have*

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \frac{2(1+\lambda)^d}{n} \frac{2\sigma^2 + 9\|f\|_\infty^2}{p_0(1-2\varepsilon)^d} + \frac{144L^2 d p_1}{p_0(1-2\varepsilon)^d} \frac{e^{-\lambda\varepsilon}}{\lambda^3} + \\ + \frac{72L^2 d^3}{\lambda^4} \Big(\frac{p_1 C_p}{p_0^2}\Big)^2 + \frac{16L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}} \Big(\frac{p_1}{p_0}\Big)^2 + \frac{8dL^2}{M\lambda^2}, \quad (2.9)$$

where $p_0 = \inf_{x\in[0,1]^d} p(x)$ *and* $p_1 = \sup_{x\in[0,1]^d} p(x)$. *In particular, letting* $s = 1+\beta$, *the choices*

$$\lambda_n \asymp L^{2/(d+2s)} n^{1/(d+2s)} \quad \text{and} \quad M_n \gtrsim L^{4\beta/(d+2s)} n^{2\beta/(d+2s)}$$

give

$$\mathbb{E}\big[(\widehat{f}_{\lambda_n,n,M_n}(X) - f(X))^2 | X \in B_\varepsilon\big] = O(L^{2d/(d+2s)} n^{-2s/(d+2s)}), \quad (2.10)$$

which corresponds to the minimax risk over the class $\mathscr{C}^{1,\beta}(L)$.

In the case where $\varepsilon = 0$, *which corresponds to integrating over the whole hypercube, the bound* (2.10) *holds if* $2s \leqslant 3$. *On the other hand, if* $2s > 3$, *letting*

$$\lambda_n \asymp L^{2/(d+3)} n^{1/(d+3)} \quad \text{and} \quad M_n \gtrsim L^{4/(d+3)} n^{2/(d+3)}$$

yields the following upper bound on the integrated risk of the Mondrian Forest estimate over B_0

$$\mathbb{E}\big[(\widehat{f}_{\lambda_n,n,M_n}(X) - f(X))^2\big] = O(L^{2d/(d+3)} n^{-3/(d+3)}). \quad (2.11)$$

The proof of Theorem 2.3 is given in Section 2.7 below. It relies on an improved control of the bias, compared to the one used in Theorem 2.2 in the Lipschitz case: it exploits the knowledge of the distribution of the cell $C_\lambda(x)$ given in Proposition 2.1 instead of merely the cell diameter given in Corollary 2.1 (which was enough for Theorem 2.2). The improved rate for Mondrian Forests compared to Mondrian Trees comes from the fact that large enough forests have a smaller bias than single trees for smooth regression functions. This corresponds to the fact that averaging randomized trees tends to smooth the decision function of single trees, which are discontinuous piecewise constant functions that approximate smooth functions sub-optimally. Such an effect was already noticed by Arlot and Genuer (2014) for purely random forests.

Remark 2.3. While (2.10) gives the minimax rate for $\mathscr{C}^{1,1}$ functions, it suffers from an unavoidable standard artifact, namely a boundary effect which impacts local averaging estimates, such as kernel estimators (Wasserman, 2006; Arlot and Genuer, 2014). It is however possible to set $\varepsilon = 0$ in (2.9), which leads to the sub-optimal rate stated in (2.11).

2.5.4 Adaptation to smoothness

The minimax rates of Theorems 2.2 and 2.3 for trees and forests are achieved through a specific tuning of the lifetime parameter λ, which depends on the considered smoothness class $\mathscr{C}^{p,\beta}(L)$ through $s = p+\beta$ and $L > 0$, while on the other hand, the number of trees M simply needs to be large enough in the statement of Theorem 2.3. Since in practice such smoothness parameters are unknown, it is of interest to obtain a single method that *adapts* to them.

In order to achieve this, we adopt a standard approach based on model aggregation (Nemirovski, 2000). More specifically, we split the dataset into two part: the first is used to fit

Mondrian Forest estimators with λ varying in an exponential grid, while the second part is used to fit the STAR procedure for model aggregation, introduced by Audibert (2008). The appeals of this aggregation procedure are its simplicity, its optimal guarantee and the lack of parameter to tune.

Let $n_0 = \lfloor n/2 \rfloor$, $\mathscr{D}_{n_0} = \{(X_i, Y_i) : 1 \leqslant i \leqslant n_0\}$ and $\mathscr{D}_{n_0+1:n} = \{(X_i, Y_i) : n_0 + 1 \leqslant i \leqslant n\}$. Also, let $I_\varepsilon = \{i \in \{n_0 + 1, \dots, n\} : X_i \in [\varepsilon, 1-\varepsilon]^d\}$ for some $\varepsilon \in (0, 1/2)$. If I_ε is empty, we let the estimator be $\widehat{g}_n = 0$. We define $A = \lfloor \log_2(n^{1/d}) \rfloor$ and $M = \lceil n^{2/d} \rceil$ and consider the geometric grid $\Lambda = \{2^\alpha : \alpha = 0, \dots, A\}$. Now, let

$$\Pi^{(1)}_{n^{1/d}}, \dots, \Pi^{(M)}_{n^{1/d}} \sim \mathsf{MP}(n^{1/d}, [0,1]^d)$$

be i.i.d. Mondrian partitions. For $m = 1, \dots, M$, we let $\Pi^{(m)}_\lambda$ be the pruning of $\Pi^{(m)}_{n^{1/d}}$ in which only splits occurring before time λ have been kept. We consider now the Mondrian Forest estimators

$$\widehat{f}_\alpha = \widehat{f}_{2^\alpha, n_0, M}$$

for every $\alpha = 0, \dots, A$, where we recall that these estimators are given by (2.4). The estimators $\widehat{f}_\alpha$ are computed using the sample $\mathscr{D}_{n_0}$ and the Mondrian partitions $\Pi^{(m)}_{2^\alpha}$, $1 \leqslant m \leqslant M$. Let

$$\widehat{\alpha} = \operatorname*{arg\,min}_{\alpha=0,\dots,A} \frac{1}{|I_\varepsilon|} \sum_{i \in I_\varepsilon} (\widehat{f}_\alpha(X_i) - Y_i)^2$$

be a risk minimizer and let $\widehat{\mathcal{G}} = \bigcup_\alpha [\widehat{f}_{\widehat{\alpha}}, \widehat{f}_\alpha]$ where $[f, g] = \{(1-t)f + tg : t \in [0,1]\}$. Note that $\widehat{\mathcal{G}}$ is a star domain with origin at the empirical risk minimizer $\widehat{f}_{\widehat{\alpha}}$, hence the name STAR (Audibert, 2008). Then, the adaptive estimator is a convex combination of two Mondrian forests estimates with different lifetime parameters, given by

$$\widehat{g}_n = \operatorname*{arg\,min}_{g \in \widehat{\mathcal{G}}} \left\{ \frac{1}{|I_\varepsilon|} \sum_{i \in I_\varepsilon} (g(X_i) - Y_i)^2 \right\}. \tag{2.12}$$

Proposition 2.4. *Grant Assumption 2.1, with $|Y| \leqslant B$ almost surely and $f \in \mathscr{C}^{p,\beta}(L)$ with $p \in \{0, 1\}$, $\beta \in (0, 1]$ and $L > 0$. Also, assume that the density p of X is C_p-Lipschitz and satisfies $p_0 \leqslant p \leqslant p_1$. Then, the estimator $\widehat{g}_n$ defined by (2.12) satisfies:*

$$\begin{aligned} \mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \min_{\alpha=0,\dots,A} \mathbb{E}\big[(\widehat{f}_\alpha(X) - f(X))^2 | X \in B_\varepsilon\big] \\ + 4B^2 e^{-c_1 n/4} + \frac{600 B^2 (\log(1 + \log_2 n) + 1)}{c_1 n} \end{aligned} \tag{2.13}$$

where $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$ and $c_1 = p_0(1-2\varepsilon)^d/4$. In particular, we have

$$\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 | X \in B_\varepsilon\big] = O\big(L^{2d/(d+2s)} n^{-2s/(d+2s)}\big), \tag{2.14}$$

where $s = p + \beta$.

The proof of Proposition 2.4 is to be found in the Supplementary Material. Proposition 2.4 proves that the estimator $\widehat{g}_n$, which is a STAR aggregation of Mondrian Forests, is adaptive to the smoothness of f, whenever f is s-Hölder with $s \in (0, 2]$.

2.5.5 Results for binary classification

We now consider, as a by-product of the analysis conducted for regression estimation, the setting of binary classification. Assume that we are given a dataset $\mathscr{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of i.i.d. random variables with values in $[0,1]^d \times \{0,1\}$, distributed as a generic pair (X, Y) and define $\eta(x) = \mathbb{P}[Y = 1 | X = x]$. We define the Mondrian Forest classifier $\widehat{g}_{\lambda,n,M}$ as a plug-in estimator of the regression estimator. Namely, we introduce

$$\widehat{g}_{\lambda,n,M}(x) = \mathbf{1}(\widehat{f}_{\lambda,n,M}(x) \geqslant 1/2)$$

for all $x \in [0,1]^d$, where $\widehat{f}_{\lambda,n,M}$ is the Mondrian Forest estimate defined in the regression setting. The performance of $\widehat{g}_{\lambda,n,M}$ is assessed by the 0-1 classification error defined as

$$L(\widehat{g}_{\lambda,n,M}) = \mathbb{P}(\widehat{g}_{\lambda,n,M}(X) \neq Y), \tag{2.15}$$

where the probability is taken with respect to $(X, Y, \Pi_{\lambda,M}, \mathscr{D}_n)$, where $\Pi_{\lambda,M}$ is the set sampled Mondrian partitions, see (2.4). Note that (2.15) is larger than the Bayes risk defined as

$$L(g^*) = \mathbb{P}(g^*(X) \neq Y),$$

where $g^*(x) = \mathbf{1}(\eta(x) \geqslant 1/2)$. A general theorem (Devroye et al., 1996, Theorem 6.5) allows us to derive an upper bound on the distance between the classification risk of $\widehat{g}_{\lambda,n,M}$ and the Bayes risk, based on Theorem 2.2.

Corollary 2.2. *Let $M \geqslant 1$ and assume that $\eta \in \mathscr{C}^{0,1}(L)$. Then, the Mondrian Forest classifier $\widehat{g}_n = \widehat{g}_{\lambda_n,n,M}$ with parameter $\lambda_n \asymp n^{1/(d+2)}$ satisfies*

$$L(\widehat{g}_n) - L(g^*) = o(n^{-1/(d+2)}).$$

The rate of convergence $o(n^{-1/(d+2)})$ for the error probability with a Lipschitz conditional probability η is optimal (Yang, 1999). In the same way, Theorem 2.3 extends to the context of classification:

Corollary 2.3. *Assume that X has a positive and Lipschitz density p w.r.t the Lebesgue measure on $[0,1]^d$ and that $\eta \in \mathscr{C}^{1,1}(L)$. Let $\widehat{g}_n = \widehat{g}_{\lambda_n,n,M_n}$ be the Mondrian Forest classifier composed of $M_n \gtrsim n^{2/(d+4)}$ trees, with lifetime $\lambda_n \asymp n^{1/(d+4)}$. Then, we have*

$$\mathbb{P}(\widehat{g}_n(X) \neq Y | X \in B_\varepsilon) - \mathbb{P}(g^*(X) \neq Y | X \in B_\varepsilon) = o(n^{-2/(d+4)}) \tag{2.16}$$

for all $\varepsilon \in (0, 1/2)$, where $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$.

This shows that Mondrian Forests achieve an improved rate compared to Mondrian trees for classification.

2.6 Conclusion

Despite their widespread use in practice, the theoretical understanding of Random Forests is still incomplete. In this chapter, we show that the Mondrian Forest, originally introduced to provide an efficient online algorithm, leads to an algorithm that is not only consistent, but in fact minimax optimal under nonparametric assumptions in arbitrary dimension. This

provides, to the best of our knowledge, the first results of this nature for a random forest method in arbitrary dimension. Besides, our analysis allows to illustrate improved rates for forests compared to individual trees. Mondrian partitions possess nice geometric properties, which can be controlled in an exact and direct fashion, while previous approaches (Biau et al., 2008; Arlot and Genuer, 2014) require arguments that work conditionally on the structure of the tree. Since Random forests are usually black-box procedures that are hard to analyze, it would be interesting to see whether the simple properties of the Mondrian process could be leveraged to design more sophisticated variants of RF that remain amenable to precise analysis.

The minimax rate $O(n^{-2s/(2s+d)})$ for a s-Hölder regression with $s \in (0, 2]$ obtained in this study is very slow when the number of features d is large. This comes from the well-known curse of dimensionality phenomenon, a problem affecting all fully nonparametric algorithms. A standard approach used in high-dimensional settings is to work under a sparsity assumption, where only $s \ll d$ features are informative. In this case, a procedure such as the Mondrian forests estimator should be used after a variable selection step. From a theoretical perspective, it would be interesting to see if the variable selection and function estimation steps could be combined, using results on the ability of forests to select informative variables (see, for instance, Scornet et al., 2015).

2.7 Proofs

This Section gathers the proofs of Proposition 2.1 and Corollary 2.1 (cell distribution and cell diameter). Then, a sketch of the proof of Proposition 2.2 is described in this Section (the full proof, which involves some technicalities, can be found in the Supplementary Material). Finally, we provide the proofs of Theorem 2.2 and Theorem 2.3.

Proof of Proposition 2.1. Let $0 \leqslant a_1, \dots, a_n, b_1, \dots, b_n \leqslant 1$ be such that $a_j \leqslant x_j \leqslant b_j$ for $1 \leqslant j \leqslant d$. Let $C := \prod_{j=1}^{d} [a_j, b_j]$. Note that the event

$$E_\lambda(C, x) = \left\{ L_{1,\lambda}(x) \leqslant a_1, R_{1,\lambda}(x) \geqslant b_1, \dots, L_{d,\lambda}(x) \leqslant a_d, R_{d,\lambda}(x) \geqslant b_d \right\}$$

coincides — up to the negligible event that one of the splits of Π_λ occurs on coordinate j at a_j or b_j — with the event that Π_λ does not cut C, *i.e.* that the restriction $\Pi_\lambda|_C$ of Π_λ to C contains no split. Now, by the restriction property of the Mondrian process (Fact 2.2), $\Pi_\lambda|_C$ is distributed as $\mathsf{MP}(\lambda, C)$; in particular, the probability that $\Pi_\lambda|_C$ contains no split is $\exp(-\lambda|C|)$. Hence, we have

$$\mathbb{P}(E_\lambda(C, x)) = e^{-\lambda(x-a_1)} e^{-\lambda(b_1-x)} \times \cdots \times e^{-\lambda(x-a_d)} e^{-\lambda(b_d-x)} . \tag{2.17}$$

In particular, setting $a_j = b_j = x$ in (2.17) except for one a_j or b_j, and using that $L_{j,\lambda}(x) \leqslant x$ and $R_{j,\lambda}(x) \geqslant x$, we obtain

$$\mathbb{P}(R_{j,\lambda}(x) \geqslant b_j) = e^{-\lambda(b_j-x)} \quad \text{and} \quad \mathbb{P}(L_{j,\lambda}(x) \leqslant a_j) = e^{-\lambda(x-a_j)}. \tag{2.18}$$

Since clearly $R_{j,\lambda}(x) \leqslant 1$ and $L_{j,\lambda}(x) \geqslant 0$, Equation (2.18) implies (ii). Additionally, plugging (2.18) back into Equation (2.17) shows that $L_{1,\lambda}(x), R_{1,\lambda}(x), \dots, L_{d,\lambda}(x), R_{d,\lambda}(x)$ are independent, *i.e.* point (i). This completes the proof. □

Proof of Corollary 2.1. Using Proposition 2.1, for $1 \leqslant j \leqslant d$, $D_{j,\lambda}(x) = R_{j,\lambda}(x) - x_j + x_j - L_{j,\lambda}(x)$ is stochastically upper bounded by $\lambda^{-1}(E_1 + E_2)$ with E_1, E_2 two independent $\mathsf{Exp}(1)$ random variables, which is distributed as $\mathsf{Gamma}(2, \lambda)$. This implies that

$$\mathbb{P}(D_{j,\lambda}(x) \geqslant \delta) \leqslant (1 + \lambda\delta)e^{-\lambda\delta} \tag{2.19}$$

for every $\delta > 0$ (with equality if $\delta \leqslant x_j \wedge (1 - x_j)$) and $\mathbb{E}[D_{j,\lambda}(x)^2] \leqslant \lambda^{-2}(\mathbb{E}[E_1^2] + \mathbb{E}[E_2^2]) = 4/\lambda^2$. The bound (2.5) for the diameter $D_\lambda(x) = [\sum_{j=1}^d D_{j,\lambda}(x)^2]^{1/2}$ is obtained by noting that

$$\mathbb{P}(D_\lambda(x) \geqslant \delta) \leqslant \mathbb{P}\Big(\exists j : D_{j,\lambda}(x) \geqslant \frac{\delta}{\sqrt{d}}\Big) \leqslant \sum_{j=1}^d \mathbb{P}\Big(D_{j,\lambda}(x) \geqslant \frac{\delta}{\sqrt{d}}\Big),$$

while (2.6) follows from the identity $\mathbb{E}[D_\lambda(x)^2] = \sum_{j=1}^d \mathbb{E}[D_{j,\lambda}(x)^2]$. □

Sketch of Proof of Proposition 2.2. Let us provide here an outline of the argument; a fully detailed proof is available in the Supplementary Material. The general idea of the proof is to modify the construction of Mondrian partitions (and hence their distribution) in a way that leaves the expected number of cells unchanged, while making this quantity directly computable.

Consider a Mondrian partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ and a cell C formed at time τ in it (e.g., $C = [0,1]^d$ for $\tau = 0$). By the properties of exponential distributions, the split of C (if it exists) from Algorithm 1 can be obtained as follows. Sample independent variables E_j, U_j with $E_j \sim \mathsf{Exp}(1)$ and $U_j \sim \mathcal{U}([0,1])$ for $j = 1, \dots, d$. Let $T_j = (b_j - a_j)^{-1} E_j$ and $S_j = a_j + (b_j - a_j)U_j$, where $C = \prod_{j=1}^d [a_j, b_j]$, and set $J = \arg\min_{1 \leqslant j \leqslant d} T_j$. If $\tau + T_J > \lambda$ then C is not split (and is thus a cell of Π_λ). On the other hand, if $\tau + T_J \leqslant \lambda$ then C is split along coordinate J at S_J (and at time $\tau + T_J$) into $C' = \{x \in C : x_J \leqslant S_J\}$ and $C'' = C \setminus C'$. This process is then repeated for the cells C' and C'', by using independent random variables E'_j, U'_j and E''_j, U''_j respectively.

Now, note that the number of cells $K_\lambda(C)$ in Π_λ contained in C is the sum of the number of cells in C' and C'', namely $K_\lambda(C')$ and $K_\lambda(C'')$. Hence, the expectation of $K_\lambda(C)$ (conditionally on previous splits) only depends on the distribution of the split (J, S_J, T_J), as well as on the marginal distributions of $K_\lambda(C')$ and $K_\lambda(C'')$, but not on the joint distribution of $(K_\lambda(C'), K_\lambda(C''))$.

Consider the following change: instead of splitting C' and C'' based on the independent random variables E'_j, U'_j and E''_j, U''_j respectively, we reuse for both C' and C'' the variables E_j, U_j (and thus S_j, T_j) for $j \neq J$, which were not used to split C. It can be seen that, for both C' and C'', these variables have the same conditional distribution given J, S_J, T_J as the independent ones. One can then form the modified random partition $\widetilde{\Pi}_\lambda$ by recursively applying this change to the construction of Π_λ, starting with the root and propagating the unused variables at each split. By the above outlined argument, its number of cells $\widetilde{K}_\lambda$ satisfies $\mathbb{E}[\widetilde{K}_\lambda] = \mathbb{E}[K_\lambda]$.

On the other hand, one can show that the partition $\widetilde{\Pi}_\lambda$ is a "product" of independent one-dimensional Mondrian partition $\Pi_\lambda^j \sim \mathsf{MP}(\lambda, [0,1])$ along the coordinates $j = 1, \dots, d$ (this means that the cells of $\widetilde{\Pi}_\lambda$ are the Cartesian products of cells of the Π_λ^j). Since the splits of a one-dimensional Mondrian partition of $[0,1]$ form a Poisson point process of intensity $\lambda \mathrm{d}x$ (Fact 2.1), the expected number of cells of Π_λ^j is $1 + \lambda$. Since the Π_λ^j for $j = \{1, \dots, d\}$ are independent, this implies that $\mathbb{E}[\widetilde{K}_\lambda] = (1 + \lambda)^d$. Once again, the full proof is provided in the Supplementary Material. □

Proof of Theorem 2.2. Recall that the Mondrian Forest estimate at x is given by

$$\widehat{f}_{\lambda,n,M}(x) = \frac{1}{M}\sum_{m=1}^{M} \widehat{f}^{(m)}_{\lambda,n}(x)\,.$$

By convexity of the function $y' \mapsto (y-y')^2$ for any $y \in \mathbf{R}$, we have

$$R(\widehat{f}_{\lambda,n,M}) \leqslant \frac{1}{M}\sum_{m=1}^{M} R(\widehat{f}^{(m)}_{\lambda,n}) = R(\widehat{f}^{(1)}_{\lambda,n}),$$

since the random trees estimators $\widehat{f}^{(m)}_{\lambda,n}$ have the same distribution for $m = 1 \dots M$. Hence, it suffices to prove Theorem 2.2 for the tree estimator $\widehat{f}^{(1)}_{\lambda,n}$. We will denote for short $\widehat{f}_\lambda := \widehat{f}^{(1)}_{\lambda,n}$ all along this proof.

Bias-variance decomposition. We establish a *bias-variance* decomposition of the risk of a Mondrian tree, akin to the one stated for purely random forests by Genuer (2012). Denote $\bar{f}_\lambda(x) := \mathbb{E}[f(X)|X \in C_\lambda(x)]$ (which depends on Π_λ) for every x in the support of μ. Given Π_λ, the function $\bar{f}_\lambda$ is the orthogonal projection of $f \in L^2([0,1]^d, \mu)$ on the subspace of functions that are constant on the cells of Π_λ. Since $\widehat{f}_\lambda$ belongs to this subspace given $\mathscr{D}_n$, we have conditionally on $(\Pi_\lambda, \mathscr{D}_n)$:

$$\mathbb{E}_X\big[(f(X) - \widehat{f}_\lambda(X))^2\big] = \mathbb{E}_X\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}_X\big[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2\big].$$

This gives the following decomposition of the risk of $\widehat{f}_\lambda$ by taking the expectation over $(\Pi_\lambda, \mathscr{D}_n)$:

$$R(\widehat{f}_\lambda) = \mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2\big]\,. \tag{2.20}$$

The first term of the sum, the *bias*, measures how close f is to its best approximation $\bar{f}_\lambda$ that is constant on the leaves of Π_λ (on average over Π_λ). The second term, the *variance*, measures how well the expected value $\bar{f}_\lambda(x)$ is estimated by the empirical average $\widehat{f}_\lambda(x)$ (on average over $\mathscr{D}_n, \Pi_\lambda$).

Note that (2.20) holds for the estimation risk *integrated over the hypercube* $[0,1]^d$, and not for the pointwise estimation risk. This is because in general, we have $\mathbb{E}_{\mathscr{D}_n}\big[\widehat{f}_\lambda(x)\big] \neq \bar{f}_\lambda(x)$: indeed, the cell $C_\lambda(x)$ may contain no data point in $\mathscr{D}_n$, in which case the estimate $\widehat{f}_\lambda(x)$ equals 0. It seems that a similar difficulty occurs for the decomposition in Genuer (2012); Arlot and Genuer (2014), which should only hold for the integrated risk.

Bias term. For each $x \in [0,1]^d$ in the support of μ, we have

$$|f(x) - \bar{f}_\lambda(x)| = \Big|\frac{1}{\mu(C_\lambda(x))}\int_{C_\lambda(x)} (f(x) - f(z))\mu(\mathrm{d}z)\Big| \leqslant \sup_{z \in C_\lambda(x)} |f(x) - f(z)| \leqslant L D_\lambda(x)^\beta,$$

where $D_\lambda(x)$ is the ℓ^2-diameter of $C_\lambda(x)$, since $f \in \mathscr{C}^{0,\beta}(L)$. By concavity of $x \mapsto x^\beta$ for $\beta \in (0,1]$ and Corollary 2.1, this implies

$$\mathbb{E}\big[(f(x) - \bar{f}_\lambda(x))^2\big] \leqslant L^2\mathbb{E}[D_\lambda(x)^{2\beta}] \leqslant L^2\mathbb{E}[D_\lambda(x)^2]^\beta \leqslant L^2\Big(\frac{4d}{\lambda^2}\Big)^\beta. \tag{2.21}$$

Integrating (2.21) with respect to μ yields the following bound on the bias:

$$\mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] \leqslant \frac{(4d)^\beta L^2}{\lambda^{2\beta}}\,. \tag{2.22}$$

Variance term. In order to bound the variance term, we use Proposition 2 in Arlot and Genuer (2014): if Π is a random tree partition of the unit cube in k cells (with $k \in \mathbf{N}^*$ deterministic) formed independently of the dataset $\mathscr{D}_n$, then

$$\mathbb{E}\big[(\bar{f}_\Pi(X) - \widehat{f}_\Pi(X))^2\big] \leqslant \frac{k}{n}(2\sigma^2 + 9\|f\|_\infty^2)\,. \tag{2.23}$$

Note that Proposition 2 in Arlot and Genuer (2014), stated in the case where the noise variance is constant, still holds when the noise variance is just upper bounded, based on Proposition 1 in Arlot (2008). For every $k \in \mathbf{N}^*$, applying (2.23) to the random partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ conditionally on the event $\{K_\lambda = k\}$, we get

$$\begin{aligned}
\mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_{\lambda,n}(X))^2\big] &= \sum_{k=1}^{\infty} \mathbb{P}(K_\lambda = k)\, \mathbb{E}[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2 | K_\lambda = k] \\
&\leqslant \sum_{k=1}^{\infty} \mathbb{P}(K_\lambda = k)\, \frac{k}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right) \\
&= \frac{\mathbb{E}[K_\lambda]}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right).
\end{aligned}$$

Using Proposition 2.2, we obtain an upper bound of the variance term:

$$\mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2\big] \leqslant \frac{(1+\lambda)^d}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right). \tag{2.24}$$

Combining (2.22) and (2.24) leads to (2.7). Finally, the bound (2.8) follows by using $\lambda = \lambda_n$ in (2.7), which concludes the proof of Theorem 2.2. □

Proof of Theorem 2.3. Consider a Mondrian Forest

$$\widehat{f}_{\lambda,M}(x) = \frac{1}{M}\sum_{m=1}^{M} \widehat{f}_\lambda^{(m)}(x),$$

where the Mondrian Trees $\widehat{f}_\lambda^{(m)}$ for $m = 1, \ldots, M$ are based on independent partitions $\Pi_\lambda^{(m)} \sim \mathsf{MP}(\lambda, [0,1]^d)$. Also, for x in the support of μ let

$$\bar{f}_\lambda^{(m)}(x) = \mathbb{E}_X[f(X) | X \in C_\lambda^{(m)}(x)],$$

which depends on $\Pi_\lambda^{(m)}$. Let $\widetilde{f}_\lambda(x) = \mathbb{E}[\bar{f}_\lambda^{(m)}(x)]$, which is deterministic and does not depend on m. Denoting $\bar{f}_{\lambda,M}(x) = \frac{1}{M}\sum_{m=1}^{M} \bar{f}_\lambda^{(m)}(x)$, we have

$$\mathbb{E}[(\widehat{f}_{\lambda,M}(x) - f(x))^2] \leqslant 2\mathbb{E}[(\widehat{f}_{\lambda,M}(x) - \bar{f}_{\lambda,M}(x))^2] + 2\mathbb{E}[(\bar{f}_{\lambda,M}(x) - f(x))^2].$$

In addition, Jensen's inequality implies that

$$\mathbb{E}\big[(\widehat{f}_{\lambda,M}(x) - \bar{f}_{\lambda,M}(x))^2\big] \leqslant \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\big[(\widehat{f}_\lambda^{(m)}(x) - \bar{f}_\lambda^{(m)}(x))^2\big] = \mathbb{E}\big[(\widehat{f}_\lambda^{(1)}(x) - \bar{f}_\lambda^{(1)}(x))^2\big].$$

For every x we have that $\bar{f}_\lambda^{(m)}(x)$ are i.i.d. for $m = 1, \dots, M$ with expectation $\widetilde{f}_\lambda(x)$, so that

$$\mathbb{E}\big[(\bar{f}_{\lambda,M}(x) - f(x))^2\big] = (\widetilde{f}_\lambda(x) - f(x))^2 + \frac{\mathrm{Var}(\bar{f}_\lambda^{(1)}(x))}{M}\,.$$

Since $f \in \mathscr{C}^{1,\beta}(L)$ we have in particular that f is L-Lipschitz, hence

$$\mathrm{Var}(\bar{f}_\lambda^{(1)}(x)) \leqslant \mathbb{E}\big[(\bar{f}_\lambda^{(1)}(x) - f(x))^2\big] \leqslant L^2\mathbb{E}[D_\lambda(x)^2] \leqslant \frac{4dL^2}{\lambda^2}$$

for all $x \in [0,1]^d$, where we used Corollary 2.1 and where $D_\lambda(x)$ stands for the diameter of $C_\lambda(x)$. Consequently, taking the expectation with respect to X, we obtain

$$\mathbb{E}\big[(\widehat{f}_{\lambda,M}(X) - f(X))^2\big] \leqslant \frac{8dL^2}{M\lambda^2} + 2\mathbb{E}\big[(\widehat{f}_\lambda^{(1)}(X) - \bar{f}_\lambda^{(1)}(X))^2\big] + 2\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2\big].$$

The same upper bound holds also conditionally on $X \in B_\varepsilon := [\varepsilon, 1-\varepsilon]^d$:

$$\begin{aligned}\mathbb{E}\big[(\widehat{f}_{\lambda,M}(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \frac{8dL^2}{M\lambda^2} &+ 2\mathbb{E}\big[(\widehat{f}_\lambda^{(1)}(X) - \bar{f}_\lambda^{(1)}(X))^2 | X \in B_\varepsilon\big] \\ &+ 2\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2 | X \in B_\varepsilon\big].\end{aligned} \tag{2.25}$$

Variance term. Recall that the distribution μ of X has a positive density $p : [0,1]^d \to \mathbf{R}_+^*$ which is C_p-Lipschitz, and recall that $p_0 = \inf_{x\in[0,1]^d} p(x)$ and $p_1 = \sup_{x\in[0,1]^d} p(x)$, both of which are positive and finite, since the continuous function p reaches its maximum and minimum over the compact set $[0,1]^d$. As shown in the proof of Theorem 2.2, the variance term satisfies

$$\mathbb{E}\big[(\bar{f}_\lambda^{(1)}(X) - \widehat{f}_{\lambda,n}^{(1)}(X))^2\big] \leqslant \frac{(1+\lambda)^d}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right).$$

Hence, the conditional variance in the decomposition (2.25) satisfies

$$\begin{aligned}\mathbb{E}\big[(\bar{f}_\lambda^{(1)}(X) - \widehat{f}_\lambda^{(1)}(X))^2 | X \in B_\varepsilon\big] &\leqslant \mathbb{P}(X \in B_\varepsilon)^{-1}\mathbb{E}\big[(\bar{f}_\lambda^{(1)}(X) - \widehat{f}_\lambda^{(1)}(X))^2\big] \\ &\leqslant p_0^{-1}(1-2\varepsilon)^{-d}\frac{(1+\lambda)^d}{n}(2\sigma^2 + 9\|f\|_\infty^2).\end{aligned} \tag{2.26}$$

Expression of $\widetilde{f}_\lambda$. It remains to control the bias term in the decomposition (2.25), which is the most involved part of the proof. Let us recall that $C_\lambda(x)$ stands for the cell of Π_λ which contains $x \in [0,1]^d$. We have

$$\widetilde{f}_\lambda(x) = \mathbb{E}\Big[\frac{1}{\mu(C_\lambda(x))}\int_{[0,1]^d} f(z)p(z)\mathbf{1}(z \in C_\lambda(x))\,\mathrm{d}z\Big] = \int_{[0,1]^d} f(z)\,F_{p,\lambda}(x,z)\,\mathrm{d}z, \tag{2.27}$$

where we defined

$$F_{p,\lambda}(x,z) = \mathbb{E}\left[\frac{p(z)\mathbf{1}(z \in C_\lambda(x))}{\mu(C_\lambda(x))}\right].$$

In particular, $\int_{[0,1]^d} F_{p,\lambda}(x,z)\mathrm{d}z = 1$ for any $x \in [0,1]^d$ (letting $f \equiv 1$ above). Let us also define the function F_λ, which corresponds to the case $p \equiv 1$:

$$F_\lambda(x,z) = \mathbb{E}\Big[\frac{\mathbf{1}(z \in C_\lambda(x))}{\mathrm{vol}(C_\lambda(x))}\Big],$$

where $\mathrm{vol}(C)$ stands for the volume of a box C.

Second order expansion. Assume that $f \in \mathscr{C}^{1+\beta}([0,1]^d)$ for some $\beta \in (0,1]$. This implies that

$$|f(z) - f(x) - \nabla f(x)^\top (z-x)| = \Big| \int_0^1 [\nabla f(x + t(z-x)) - \nabla f(x)]^\top (z-x) \mathrm{d}t \Big| \\ \leqslant \int_0^1 L(t\|z-x\|)^\beta \|z-x\| \mathrm{d}t \leqslant L\|z-x\|^{1+\beta} .$$

Now, by the triangle inequality,

$$\Big| \Big| \int_{[0,1]^d} (f(z) - f(x)) F_{p,\lambda}(x,z) \mathrm{d}z \Big| - \Big| \int_{[0,1]^d} \nabla f(x)^\top (z-x) F_{p,\lambda}(x,z) \mathrm{d}z \Big| \Big| \\ \leqslant \Big| \int_{[0,1]^d} (f(z) - f(x) - \nabla f(x)^\top (z-x)) F_{p,\lambda}(x,z) \mathrm{d}z \Big| \\ \leqslant L \int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z) \mathrm{d}z ,$$

so that, using together $\int F_{p,\lambda}(x,z)\mathrm{d}z = 1$ and (2.27), we obtain

$$|\widetilde{f}_\lambda(x) - f(x)| \leqslant \Big| \nabla f(x)^\top \underbrace{\int_{[0,1]^d} (z-x) F_{p,\lambda}(x,z) \mathrm{d}z}_{:=A} \Big| + L \underbrace{\int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z) \mathrm{d}z}_{:=B} . \quad (2.28)$$

Hence, it remains to control the two terms A, B from Equation (2.28). We will start by expressing $F_{p,\lambda}$ in terms of p, using the distribution of the cell $C_\lambda(x)$ given by Proposition 2.1 above. Next, both terms will be bounded by approximating $F_{p,\lambda}$ by F_λ and controlling these terms for F_λ (this is done in Technical Lemma 2.1 below).

Explicit form of $F_{p,\lambda}$. First, we provide an explicit form of $F_{p,\lambda}$ in terms of p. We start by determining the distribution of the cell $C_\lambda(x)$ conditionally on the event $z \in C_\lambda(x)$. Let $C = C(x,z) = \prod_{1 \leqslant j \leqslant d} [x_j \wedge z_j, x_j \vee z_j] \subseteq [0,1]^d$ be the smallest box containing both x and z; also, let $a_j = x_j \wedge z_j$, $b_j = x_j \vee z_j$, $a = (a_j)_{1 \leqslant j \leqslant d}$ and $b = (b_j)_{1 \leqslant j \leqslant d}$. Note that $z \in C_\lambda(x)$ if and only if Π_λ does not cut C. Since $C = C(x,z) = C(a,b)$, we have that $z \in C_\lambda(x)$ if and only if $b \in C_\lambda(a)$, and in this case $C_\lambda(x) = C_\lambda(a)$. In particular, the conditional distribution of $C_\lambda(x)$ given $z \in C_\lambda(x)$ equals the conditional distribution of $C_\lambda(a)$ given $b \in C_\lambda(a)$.

Write $C_\lambda(a) = \prod_{j=1}^d [L_{\lambda,j}(a), R_{\lambda,j}(a)]$; by Proposition 2.1, we have $L_{\lambda,j}(a) = (a_j - \lambda^{-1} E_{j,L}) \vee 0$, $R_{\lambda,j}(a) = (a_j + \lambda^{-1} E_{j,R}) \wedge 1$, where $E_{j,L}, E_{j,R}$, $1 \leqslant j \leqslant d$ are i.i.d. $\mathsf{Exp}(1)$ random variables. Note that $b \in C_\lambda(a)$ is equivalent to $R_{\lambda,j}(a) \geqslant b_j$ for $j = 1, \dots, d$, *i.e.* to $E_{j,R} \geqslant \lambda(b_j - a_j)$. By the memory-less property of the exponential distribution, the distribution of $E_{j,R} - \lambda(b_j - a_j)$ conditionally on $E_{j,R} \geqslant \lambda(b_j - a_j)$ is $\mathsf{Exp}(1)$. As a result (using the independence of the variables $E_{j,L}, E_{j,R}$), we obtain the following statement:

> Conditionally on $b \in C_\lambda(a)$, the coordinates $L_{\lambda,j}(a), R_{\lambda,j}(a)$, $1 \leqslant j \leqslant d$, are distributed as $(a_j - \lambda^{-1} E'_{j,L}) \vee 0, (b_j + \lambda^{-1} E'_{j,R}) \wedge 1$, where $E'_{j,L}, E'_{j,R}$ are i.i.d. $\mathsf{Exp}(1)$ random variables.

Hence, the distribution of $C_\lambda(x)$ conditionally on $z \in C_\lambda(x)$ has the same distribution as

$$C_\lambda(x,z) := \prod_{j=1}^d \left[(x_j \wedge z_j - \lambda^{-1} E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1} E_{j,R}) \wedge 1 \right] \quad (2.29)$$

where $E_{1,L}, E_{1,R}, \dots, E_{d,L}, E_{d,R}$ are i.i.d. $\mathsf{Exp}(1)$ random variables. In addition, note that $z \in C_\lambda(x)$ if and only if the restriction of Π_λ to $C(x,z)$ has no split (*i.e.*, its first sampled split occurs after time λ). Since this restriction is distributed as $\mathsf{MP}(\lambda, C(x,z))$ using Fact 2.2, this occurs with probability $\exp(-\lambda|C(x,z)|) = \exp(-\lambda\|x-z\|_1)$. Therefore,

$$\begin{aligned} F_{p,\lambda}(x,z) &= \mathbb{P}(z \in C_\lambda(x))\,\mathbb{E}\Big[\frac{p(z)}{\mu(C_\lambda(x))}\,\Big|\, z \in C_\lambda(x)\Big] \\ &= e^{-\lambda\|x-z\|_1}\,\mathbb{E}\bigg[\Big\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\Big\}^{-1}\bigg], \end{aligned} \tag{2.30}$$

where $C_\lambda(x,z)$ is as in (2.29). In addition, applying (2.30) to $p \equiv 1$ yields

$$F_\lambda(x,z) = \lambda^d e^{-\lambda\|x-z\|_1} \prod_{1\leqslant j\leqslant d} \mathbb{E}\Big[\big\{\lambda|x_j - z_j| + E_{j,L} \wedge \lambda(x_j \wedge z_j) + E_{j,R} \wedge \lambda(1 - x_j \vee z_j)\big\}^{-1}\Big]. \tag{2.31}$$

The following technical Lemma, whose proof is given in Section 2.8.6, will prove useful in what follows.

Lemma 2.1. *The function $F_{p,\lambda}$ given by (2.31) satisfies, for any $x \in [0,1]^d$,*

$$\begin{aligned} \Big\|\int_{[0,1]^d} (z-x)F_\lambda(x,z)\mathrm{d}z\Big\|^2 &\leqslant \frac{9}{\lambda^2}\sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} \\ \int_{[0,1]^d} \frac{1}{2}\|z-x\|^2 F_\lambda(x,z)\mathrm{d}z &\leqslant \frac{d}{\lambda^2}\,. \end{aligned}$$

Control of the term B in Equation (2.28). It follows from (2.30) and from the bound $p(y)/p(z) \geqslant p_0/p_1$ that

$$F_{p,\lambda}(x,z) \leqslant \frac{p_1}{p_0} F_\lambda(x,z), \tag{2.32}$$

so that

$$\begin{aligned} \int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z)\mathrm{d}z &\leqslant \frac{p_1}{p_0}\int_{[0,1]^d} \|z-x\|^{1+\beta} F_\lambda(x,z)\mathrm{d}z \\ &\leqslant \frac{p_1}{p_0}\Big(\int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z)\mathrm{d}z\Big)^{(1+\beta)/2} \qquad (2.33) \\ &\leqslant \frac{p_1}{p_0}\Big(\frac{2d}{\lambda^2}\Big)^{(1+\beta)/2}, \qquad (2.34) \end{aligned}$$

where (2.33) follows from the concavity of $x \mapsto x^{(1+\beta)/2}$ for $\beta \in (1,2]$, while (2.34) comes from Lemma 2.1.

Control of the term A in Equation (2.28). It remains to control $A = \int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\mathrm{d}z$. Again, this quantity is controlled in the case of a uniform density ($p \equiv 1$) in Lemma 2.1. However, this time the crude bound (2.32) is no longer sufficient, since we need first-order terms to compensate in order to obtain the optimal rate. Rather, we will show that $F_{p,\lambda}(x,z) = (1 + O(\|x-z\|) + O(1/\lambda))F_\lambda(x,z)$.

A first upper bound on $|F_{p,\lambda}(x,z) - F_\lambda(x,z)|$. Since p is C_p-Lipschitz and lower bounded by p_0, we have

$$\left|\frac{p(y)}{p(z)} - 1\right| = \frac{|p(y) - p(z)|}{p(z)} \leqslant \frac{C_p}{p_0}\|y - z\| \leqslant \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z) \tag{2.35}$$

for every $y \in C_\lambda(x,z)$, so that

$$1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z) \leqslant \frac{p(y)}{p(z)} \leqslant 1 + \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z).$$

Integrating over $C_\lambda(x,z)$ and using $p(y)/p(z) \geqslant p_0/p_1$ gives

$$\begin{aligned}\left\{1 + \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right\}^{-1} \operatorname{vol} C_\lambda(x,z)^{-1} &\leqslant \left\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\right\}^{-1} \\ &\leqslant \left\{\left(1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right) \vee \frac{p_0}{p_1}\right\}^{-1} \operatorname{vol} C_\lambda(x,z)^{-1}. \end{aligned} \tag{2.36}$$

In addition, since $(1+u)^{-1} \geqslant 1 - u$ for $u \geqslant 0$, we have

$$\left\{1 + \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right\}^{-1} \geqslant 1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\,,$$

so that setting $a := \left(1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right) \vee \frac{p_0}{p_1} \in (0,1]$ gives

$$a^{-1} - 1 = \frac{1-a}{a} \leqslant \frac{(C_p/p_0)\operatorname{diam} C_\lambda(x,z)}{p_0/p_1} = \frac{p_1 C_p}{p_0^2}\operatorname{diam} C_\lambda(x,z).$$

Now, Equation (2.36) implies that

$$\begin{aligned}-\frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1} &\leqslant \left\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\right\}^{-1} - \operatorname{vol} C_\lambda(x,z)^{-1} \\ &\leqslant \frac{p_1 C_p}{p_0^2}\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\,.\end{aligned}$$

Taking the expectation over $C_\lambda(x,z)$ and using (2.30) leads to

$$\begin{aligned}-\frac{C_p}{p_0}\mathbb{E}\left[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\right] &\leqslant e^{\lambda\|x-z\|_1}(F_{p,\lambda}(x,z) - F_\lambda(x,z)) \\ &\leqslant \frac{p_1 C_p}{p_0^2}\mathbb{E}\left[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\right]\end{aligned}$$

so that

$$|F_{p,\lambda}(x,z) - F_\lambda(x,z)| \leqslant \frac{p_1 C_p}{p_0^2} e^{-\lambda\|x-z\|_1} \times \mathbb{E}\left[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\right]. \tag{2.37}$$

Control of $\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z)\operatorname{vol} C_\lambda(x,z)^{-1}\big]$. Let us define the interval

$$C_\lambda^j(x,z) := \big[(x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1\big]$$

and let $|C_\lambda^j(x,z)| = (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1 - (x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0$ be its length. We have $\operatorname{diam} C_\lambda(x,z) \leqslant \operatorname{diam}_{\ell^1} C_\lambda(x,z)$ using the triangular inequality, so that

$$\begin{aligned}
\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z)\operatorname{vol} C_\lambda(x,z)^{-1}\big] &\leqslant \mathbb{E}\Big[\sum_{j=1}^d |C_\lambda^j(x,z)|\operatorname{vol} C_\lambda(x,z)^{-1}\Big] \\
&= \sum_{j=1}^d \mathbb{E}\Big[|C_\lambda^j(x,z)|\prod_{l=1}^d |C_\lambda^l(x,z)|^{-1}\Big] = \sum_{j=1}^d \mathbb{E}\Big[\prod_{l\neq j}|C_\lambda^l(x,z)|^{-1}\Big] \\
&\leqslant \sum_{j=1}^d \mathbb{E}\Big[|C_\lambda^j(x,z)|\Big]\mathbb{E}\Big[|C_\lambda^j(x,z)|^{-1}\Big]\mathbb{E}\Big[\prod_{l\neq j}|C_\lambda^l(x,z)|^{-1}\Big] \qquad (2.38)\\
&= \sum_{j=1}^d \mathbb{E}\Big[|C_\lambda^j(x,z)|\Big] \times \mathbb{E}\Big[\prod_{l=1}^d |C_\lambda^l(x,z)|^{-1}\Big] \qquad (2.39)\\
&= \mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] \times \exp(\lambda\|x-z\|_1)F_\lambda(x,z). \qquad (2.40)
\end{aligned}$$

Inequality (2.38) relies on the fact that $\mathbb{E}[X]\mathbb{E}[X^{-1}] \geqslant 1$ for any positive random variable X with $X = |C_\lambda^j(x,z)|$. Equality (2.39) comes from the independence of $|C_\lambda^1(x,z)|, \dots, |C_\lambda^d(x,z)|$. Multiplying both sides of (2.40) by $e^{-\lambda\|x-z\|_1}$ leads to

$$e^{-\lambda\|x-z\|_1}\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z)\operatorname{vol} C_\lambda(x,z)^{-1}\big] \leqslant \mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big]F_\lambda(x,z)\,. \qquad (2.41)$$

In addition,

$$\begin{aligned}
\mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] &\leqslant \sum_{j=1}^d \mathbb{E}\big[|x_j - z_j| + \lambda^{-1}(E_{j,R}+E_{j,L})\big] \\
&= \|x-z\|_1 + \frac{2d}{\lambda}\,. \qquad (2.42)
\end{aligned}$$

Finally, combining Equations (2.37), (2.41) and (2.42) gives

$$|F_{p,\lambda}(x,z) - F_\lambda(x,z)| \leqslant \frac{p_1 C_p}{p_0^2}\Big(\|x-z\|_1 + \frac{2d}{\lambda}\Big)F_\lambda(x,z)\,. \qquad (2.43)$$

Control of A. From (2.43), we can control $\int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\mathrm{d}z$ by approximating $F_{p,\lambda}$ by F_λ. Indeed, we have

$$\Big\|\int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\mathrm{d}z - \int_{[0,1]^d}(z-x)F_\lambda(x,z)\mathrm{d}z\Big\| \leqslant \int_{[0,1]^d}\|z-x\|\cdot|F_{p,\lambda}(x,z)-F_\lambda(x,z)|\mathrm{d}z, \qquad (2.44)$$

with

$$\begin{aligned}
&\int_{[0,1]^d} \|z-x\| \times |F_{p,\lambda}(x,z) - F_\lambda(x,z)| \mathrm{d}z \\
&\leqslant \frac{p_1 C_p}{p_0^2} \int_{[0,1]^d} \|z-x\| \Big[\|x-z\|_1 + \frac{2d}{\lambda} \Big] F_\lambda(x,z) \mathrm{d}z \qquad \text{(by (2.43))} \\
&\leqslant \frac{p_1 C_p}{p_0^2} \Big[\sqrt{d} \int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z) \mathrm{d}z + \frac{2d}{\lambda} \int_{[0,1]^d} \|z-x\| F_\lambda(x,z) \mathrm{d}z \Big] \\
&\leqslant \frac{p_1 C_p}{p_0^2} \Big[\frac{d\sqrt{d}}{\lambda^2} + \frac{2d}{\lambda} \Big(\int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z) \mathrm{d}z \Big)^{1/2} \Big],
\end{aligned}$$

where we used the inequalities $\|v\| \leqslant \|v\|_1 \leqslant \sqrt{d}\|v\|$ as well as the Cauchy-Schwarz inequality. Hence, using Lemma 2.1, we end up with

$$\begin{aligned}
\int_{[0,1]^d} \|z-x\| \times |F_{p,\lambda}(x,z) - F_\lambda(x,z)| \mathrm{d}z &\leqslant \frac{p_1 C_p}{p_0^2} \Big[\frac{d\sqrt{d}}{\lambda^2} + \frac{2d}{\lambda} \sqrt{\frac{d}{\lambda^2}} \Big] \\
&= \frac{p_1 C_p}{p_0^2} \frac{3d\sqrt{d}}{\lambda^2} \,. \qquad (2.45)
\end{aligned}$$

Inequalities (2.44) and (2.45) together with Lemma 2.1 entail that

$$\begin{aligned}
\Big\| \int_{[0,1]^d} (z-x) F_{p,\lambda}(x,z) \mathrm{d}z \Big\|^2 &\leqslant 2 \Big\| \int_{[0,1]^d} (z-x) F_\lambda(x,z) \mathrm{d}z \Big\|^2 \\
&\quad + 2 \Big(\int_{[0,1]^d} \|z-x\| |F_{p,\lambda}(x,z) - F_\lambda(x,z)| \mathrm{d}z \Big)^2 \\
&\leqslant \frac{18}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} + 2 \Big(\frac{p_1 C_p}{p_0^2} \frac{3d\sqrt{d}}{\lambda^2} \Big)^2 . \qquad (2.46)
\end{aligned}$$

Control of the bias. The upper bound (2.28) on the bias writes

$$(\widetilde{f}_\lambda(x) - f(x))^2 \leqslant (|\nabla f(x)^\top A| + LB)^2 \leqslant 2(|\nabla f(x)\|^2 \times \|A\|^2 + L^2 B^2) \,,$$

so that plugging the bounds (2.34) of B and (2.46) of $\|A\|$ gives

$$\begin{aligned}
(\widetilde{f}_\lambda(x) - f(x))^2 &\leqslant 2L^2 \Big[\frac{18}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} + 2 \Big(\frac{p_1 C_p}{p_0^2} \frac{3d\sqrt{d}}{\lambda^2} \Big)^2 \Big] + 2L^2 \frac{p_1}{p_0} \Big(\frac{2d}{\lambda^2} \Big)^{(1+\beta)/2} \\
&\leqslant \frac{36L^2}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} + \frac{36L^2 d^3}{\lambda^4} \Big(\frac{p_1 C_p}{p_0^2} \Big)^2 + \frac{8L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}} \Big(\frac{p_1}{p_0} \Big)^2 .
\end{aligned}$$

By integrating over X conditionally on $X \in B_\varepsilon$, this implies

$$\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2 | X \in B_\varepsilon \big] \leqslant \frac{36L^2}{\lambda^2} \psi_\varepsilon(\lambda) + \frac{36L^2 d^3}{\lambda^4} \Big(\frac{p_1 C_p}{p_0^2} \Big)^2 + \frac{8L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}} \Big(\frac{p_1}{p_0} \Big)^2 , \qquad (2.47)$$

where we have, using the fact that $p_0 \leqslant p(x) \leqslant p_1$ for any $x \in [0,1]$,

$$\psi_\varepsilon(\lambda) := \sum_{j=1}^{d} \mathbb{E}\big[e^{-\lambda[X_j \wedge (1-X_j)]} | X \in B_\varepsilon\big] \leqslant \frac{dp_1}{p_0(1-2\varepsilon)^d} \int_\varepsilon^{1-\varepsilon} e^{-\lambda[u\wedge(1-u)]} \mathrm{d}u$$
$$= \frac{dp_1}{p_0(1-2\varepsilon)^d} \times 2\int_\varepsilon^{1/2} e^{-\lambda u} \mathrm{d}u \leqslant \frac{e^{-\lambda\varepsilon}}{\lambda} \frac{2dp_1}{p_0(1-2\varepsilon)^d}\,.$$

Conclusion. The decomposition (2.25), together with the bounds (2.26) on the variance and (2.47) on the bias lead to inequality (2.9) from the statement of Theorem 2.3. In particular, if $\varepsilon \in (0, \frac{1}{2})$ is fixed, inequality (2.9) writes

$$\mathbb{E}\big[(\widehat{f}_{\lambda,M}(X) - f(X))^2 | X \in B_\varepsilon\big] = O\Big(\frac{\lambda^d}{n} + \frac{L^2}{\lambda^{2(1+\beta)}} + \frac{L^2}{M\lambda^2}\Big).$$

One can optimize the right-hand side by setting $\lambda = \lambda_n \asymp L^{2/(d+2s)} n^{1/(d+2s)}$ and $M = M_n \gtrsim \lambda_n^{2\beta} \asymp L^{4\beta/(d+2s)} n^{2\beta/(d+2s)}$ with $s = 1+\beta \in (1,2]$. This leads to the minimax rate $O(L^{2d/(d+2s)} n^{-2s/(d+2s)})$ for $f \in \mathscr{C}^{1,\beta}(L)$ as announced in the statement of Theorem 2.3.

On the other hand, we have $e^{-\lambda\varepsilon} = 1$ whenever $\varepsilon = 0$, so that inequality (2.9) becomes in this case

$$\mathbb{E}[(\widehat{f}_{\lambda,M}(X) - f(X))^2] \leqslant O\Big(\frac{\lambda^d}{n} + \frac{L^2}{\lambda^{3\wedge(2s)}} + \frac{L^2}{M\lambda^2}\Big).$$

When $2s \leqslant 3$ (*i.e.* $\beta \leqslant 1/2$), this leads to the same rate as above, with the same choice of parameters. When $2s > 3$, this leads to the suboptimal rate $O(L^{2d/(d+3)} n^{-3/(d+3)})$ with the choice $M_n \gtrsim \lambda_n \asymp L^{2/(d+3)} n^{1/(d+3)}$. This concludes the proof of all the claims from Theorem 2.3. □

Notation or formula	Description
$\mathbf{v} \in \{0,1\}^*$	A node
$\mathscr{D}_n$	Data set
μ	Distribution of X on $[0,1]^d$
C, resp. $\lvert C\rvert$	A generic cell $C \subset [0,1]^d$, resp. half-perimeter of C
λ	Lifetime parameter of Mondrian process
$\mathsf{MP}(\lambda, C)$	Distribution of a Mondrian process defined on cell C with lifetime parameter λ
Π_λ, resp. $\Pi_\lambda \lvert C$	Partition drawn from $\mathsf{MP}(\lambda,[0,1]^d)$, resp. $\mathsf{MP}(\lambda, C)$
$C_\lambda(x)$	Cell of a Mondrian tree with parameter λ containing x
$D_\lambda(x)$	Diameter of $C_\lambda(x)$
K_λ	Number of cells in a Mondrian Tree partition Π_λ
f	True regression function: $f(X) = \mathbb{E}[Y\lvert X]$
$\widehat{f}^{(m)}_{\lambda,n}(x)$	Mondrian Tree estimate at query point x based on the Mondrian partition $\Pi^{(m)}_\lambda$
$\widehat{f}_{\lambda,n,M}(x)$	Mondrian Forest estimate at query point x based on the Mondrian partitions $\Pi_{\lambda,M} = (\Pi^{(1)}_\lambda, \dots, \Pi^{(M)}_\lambda)$
$\bar{f}^{(m)}_\lambda(x)$	Expected value of f inside the cell $C^{(m)}_\lambda(x)$
$\widetilde{f}_\lambda(x)$	Expected value of $\bar{f}^{(m)}_\lambda(x)$ over $\Pi^{(m)}_\lambda \sim \mathsf{MP}(\lambda,[0,1]^d)$
$\mathcal{N}(\mathcal{T}), \mathcal{N}^\circ(\mathcal{T}), \mathcal{L}(\mathcal{T})$	Nodes, interior nodes and leaves of a tree $\mathcal{T}$
$\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}^\circ(\mathcal{T})}$	Set of splits for all nodes in the tree $\mathcal{T}$
$\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$	A split at node $\mathbf{v}$ characterized by its split dimension $j_\mathbf{v} \in \{1,\dots,d\}$ and threshold $s_\mathbf{v} \in [0,1]$
$\tau_\mathbf{v}$	Birth time of a node $\mathbf{v}$

Table 2.1: Notations and definitions used in this section

2.8 Remaining proofs

In this section, we gather several proofs and technical details and definitions that were omitted in the rest of the chapter. Namely, we start with a glossary of notations (Table 2.1), then give extra definitions and notations for trees and nested trees partitions in Section 2.8.1. Then, we provide proofs that were omitted in the chapter by order of appearance, namely the proofs of Proposition 2.2, Theorem 2.1, Proposition 2.3, Proposition 2.4 and Lemma 2.1.

2.8.1 Specific notations

Let us now introduce some specific notations to describe the decision tree structure and the Mondrian Process.

2.8.1.1 Trees and nested tree partitions

A decision tree $(\mathcal{T}, \Sigma)$ is composed of the following components:

- A finite rooted ordered binary tree $\mathcal{T}$, with nodes $\mathcal{N}(\mathcal{T})$, interior nodes $\mathcal{N}^\circ(\mathcal{T})$ and leaves $\mathcal{L}(\mathcal{T})$ (so that $\mathcal{N}(\mathcal{T})$ is the disjoint union of $\mathcal{N}^\circ(\mathcal{T})$ and $\mathcal{L}(\mathcal{T})$). The nodes $\mathbf{v} \in \mathcal{N}(\mathcal{T})$ are finite words on the alphabet $\{0,1\}$, that is elements of the set $\{0,1\}^* = \bigcup_{n\geqslant 0}\{0,1\}^n$:

the root ϵ of $\mathcal{T}$ is the empty word, and for every interior $\mathbf{v} \in \{0,1\}^*$, its left child is $\mathbf{v}0$ (obtained by adding a 0 at the end of $\mathbf{v}$) while its right child is $\mathbf{v}1$ (obtained by adding a 1 at the end of $\mathbf{v}$).

- A family of *splits* $\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}^\circ(\mathcal{T})}$ at each interior node, where each split $\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$ is characterized by its split dimension $j_\mathbf{v} \in \{1,\dots,d\}$ and its threshold $s_\mathbf{v} \in [0,1]$.

We associate to $\Pi = (\mathcal{T}, \Sigma)$ a partition $(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(\mathcal{T})}$ of the unit cube $[0,1]^d$, called a *tree partition* (or *guillotine partition*). For each node $\mathbf{v} \in \mathcal{N}(\mathcal{T})$, we define a hyper-rectangular region $C_\mathbf{v}$ recursively:

- The cell associated to the root of $\mathcal{T}$ is $[0,1]^d$;
- For each $\mathbf{v} \in \mathcal{N}^\circ(\mathcal{T})$, we define

$$C_{\mathbf{v}0} := \{x \in C_\mathbf{v} : x_{j_\mathbf{v}} \leqslant s_{j_\mathbf{v}}\} \quad \text{and} \quad C_{\mathbf{v}1} := C_\mathbf{v} \setminus C_{\mathbf{v}0}.$$

The leaf cells $(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(\mathcal{T})}$ form a partition of $[0,1]^d$ by construction. In what follows, we will identify a tree with splits $(\mathcal{T}, \Sigma)$ with its associated tree partition, and a node $\mathbf{v} \in \mathcal{N}(\mathcal{T})$ with the cell $C_\mathbf{v} \subset [0,1]^d$. The Mondrian process, described in the next Section, defines a distribution over nested tree partitions, defined below.

Definition 2.2 (Nested tree partitions)**.** A tree partition $\Pi' = (\mathcal{T}', \Sigma')$ is a *refinement* of the tree partition $\Pi = (\mathcal{T}, \Sigma)$ if $\mathcal{T}$ is a subtree of $\mathcal{T}'$ and, for every $\mathbf{v} \in \mathcal{N}(\mathcal{T}) \subseteq \mathcal{N}(\mathcal{T}')$, $\sigma_\mathbf{v} = \sigma'_\mathbf{v}$. A *nested tree partition* is a family $(\Pi_t)_{t\geqslant 0}$ of tree partitions such that, for every $t, t' \in \mathbf{R}^+$ with $t \leqslant t'$, $\Pi_{t'}$ is a refinement of Π_t. Such a family can be described as follows: let $\mathbf{T}$ be the (in general infinite, and possibly complete) rooted binary tree, such that $\mathcal{N}(\mathbf{T}) = \bigcup_{t\geqslant 0} \mathcal{N}(\mathcal{T}_t) \subseteq \{0,1\}^*$. For each $\mathbf{v} \in \mathcal{N}(\mathcal{T})$, let $\tau_\mathbf{v} = \inf\{t \geqslant 0 \mid \mathbf{v} \in \mathcal{N}(\mathcal{T}_t)\} < \infty$ denote the *birth time* of the node $\mathbf{v}$. Additionally, let $\sigma_\mathbf{v}$ be the value of the split $\sigma_{\mathbf{v},t}$ in Π_t for $t > \tau_\mathbf{v}$ (which does not depend on t by the refinement property). Then, Π is completely characterized by $\mathbf{T}$, $\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}(\mathbf{T})}$ and $\mathfrak{T} = (\tau_\mathbf{v})_{\mathbf{v}\in\mathcal{N}(\mathbf{T})}$.

2.8.1.2 Mondrian Process

To define rigorously the Mondrian Process, we introduce the function Φ_C, which maps any family of couples $(e^j_\mathbf{v}, u^j_\mathbf{v}) \in \mathbf{R}^+ \times [0,1]$ indexed by the coordinates $j \in \{1,\dots,d\}$ and the nodes $\mathbf{v} \in \{0,1\}^*$ to a nested tree partition $\Pi = \Phi_C((e^j_\mathbf{v}, u^j_\mathbf{v})_{\mathbf{v},j})$ of C. The splits $\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$ and birth times $\tau_\mathbf{v}$ of the nodes $\mathbf{v} \in \{0,1\}^*$ are defined recursively, starting from the root ϵ:

- For the root node ϵ, we let $\tau_\epsilon = 0$ and $C_\epsilon = C$.
- At each node $\mathbf{v} \in \{0,1\}^*$, given the labels of all its ancestors $\mathbf{v}' \sqsubset \mathbf{v}$ (so that in particular $\tau_\mathbf{v}$ and $C_\mathbf{v}$ are determined), denote $C_\mathbf{v} = \prod_{j=1}^d [a^j_\mathbf{v}, b^j_\mathbf{v}]$. Then, select the split dimension $j_\mathbf{v} \in \{1,\dots,d\}$ and its location $s_\mathbf{v}$ as follows:

$$j_\mathbf{v} = \operatorname*{arg\,min}_{j=1,\dots,d} \frac{e^j_\mathbf{v}}{b^j_\mathbf{v} - a^j_\mathbf{v}}, \qquad s_\mathbf{v} = a^{j_\mathbf{v}}_\mathbf{v} + (b^{j_\mathbf{v}}_\mathbf{v} - a^{j_\mathbf{v}}_\mathbf{v}) \cdot u^{j_\mathbf{v}}_\mathbf{v}, \tag{2.48}$$

where we break ties in the choice of $j_\mathbf{v}$ e.g., by choosing the smallest index j in the arg min. The node $\mathbf{v}$ is then split at time $\tau_\mathbf{v} + e^{j_\mathbf{v}}_\mathbf{v}/(b^{j_\mathbf{v}}_\mathbf{v} - a^{j_\mathbf{v}}_\mathbf{v}) = \tau_{\mathbf{v}0} = \tau_{\mathbf{v}1}$, we let $C_{\mathbf{v}0} = \{x \in C_\mathbf{v} : x_{j_\mathbf{v}} \leqslant s_\mathbf{v}\}$, $C_{\mathbf{v}1} = C_\mathbf{v} \setminus C_{\mathbf{v}0}$ and recursively apply the procedure to its children $\mathbf{v}0$ and $\mathbf{v}1$.

For each $\lambda \in \mathbf{R}^+$, the tree partition $\Pi_\lambda = \Phi_{\lambda,C}((e^j_{\mathbf{v}}, u^j_{\mathbf{v}})_{\mathbf{v},j})$ is the *pruning of* Π *at time* λ, obtained by removing all the splits in Π that occurred strictly after λ, so that the leaves of the tree are the maximal nodes (in the prefix order) $\mathbf{v}$ such that $\tau_{\mathbf{v}} \leqslant \lambda$.

Definition 2.3 (Mondrian process)**.** Let $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j}$ be a family of independent random variables, with $E^j_{\mathbf{v}} \sim \mathsf{Exp}(1)$, $U^j_{\mathbf{v}} \sim \mathcal{U}([0,1])$. The *Mondrian process* $\mathsf{MP}(C)$ on C is the distribution of the random nested tree partition $\Phi_C((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$. In addition, we denote $\mathsf{MP}(\lambda, C)$ the distribution of $\Phi_{\lambda,C}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$.

2.8.2 Proof of Proposition 2.2

At a high level, the idea of the proof is to modify the construction of the Mondrian partition (and hence, the distribution of the underlying process) without affecting the expected number of cells. More precisely, we show a recursive way to transform the Mondrian process that leaves $\mathbb{E}[K_\lambda]$ unchanged, and which eventually leads to a random partition $\widetilde{\Pi}_\lambda$ for which this quantity can be computed directly and equals $(1+\lambda)^d$. We will in fact show the result for a general box C (not just the unit cube). The proof proceeds in two steps:

1. Define a modified process $\widetilde{\Pi}$, and show that $\mathbb{E}[\widetilde{K}_\lambda] = \prod_{j=1}^d (1 + \lambda |C^j|)$.

2. It remains to show that $\mathbb{E}[K_\lambda] = \mathbb{E}[\widetilde{K}_\lambda]$. For this, it is sufficient to show that the distribution of the birth times $\tau_{\mathbf{v}}$ and $\widetilde{\tau}_{\mathbf{v}}$ of the node $\mathbf{v}$ is the same for both processes. This is done by induction on $\mathbf{v}$, by showing that the splits at one node of both processes have the same conditional distribution given the splits at previous nodes.

Let $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}\in\{0,1\}^*, 1\leqslant j\leqslant d}$ be a family of independent random variables with $E^j_{\mathbf{v}} \sim \mathsf{Exp}(1)$ and $U^j_{\mathbf{v}} \sim \mathcal{U}([0,1])$. By definition, $\Pi = \Phi_C((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$ (Φ_C being defined in Section 2.3) follows a Mondrian process distribution $\mathsf{MP}(C)$. Denote for every node $\mathbf{v} \in \{0,1\}^*$ $C_{\mathbf{v}}$ the cell of $\mathbf{v}$, $\tau_{\mathbf{v}}$ its birth time, as well as its split time $T_{\mathbf{v}}$, dimension $J_{\mathbf{v}}$, and threshold $S_{\mathbf{v}}$ (note that $T_{\mathbf{v}} = \tau_{\mathbf{v}0} = \tau_{\mathbf{v}1}$). In addition, for $\lambda \in \mathbf{R}^+$, denote $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$ the tree partition restricted to time λ, and $K_\lambda \in \mathbf{N} \cup \{+\infty\}$ its number of nodes.

Construction of the modified process. Now, consider the following modified nested partition of C, denoted $\widetilde{\Pi}$, and defined through its split times, dimension and threshold $\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}}$ (which determine the birth times $\tau_{\mathbf{v}}$ and cells $C_{\mathbf{v}}$), and *current* j*-dimensional node* $\mathbf{v}_j(\mathbf{v}) \in \{0,1\}^*$ $(1 \leqslant j \leqslant d)$ at each node $\mathbf{v}$. First, for every $j = 1, \dots, d$, let $\Pi'^j = \Phi_{C^j}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}\in\{0,1\}^*}) \sim \mathsf{MP}(C^j)$ be the nested partition of the interval C^j determined by $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}}$; its split times and thresholds are denoted $(S'^j_{\mathbf{v}}, T'^j_{\mathbf{v}})$. Then, $\widetilde{\Pi}$ is defined recursively as follows:

- At the root node ϵ, let $\widetilde{\tau}_\epsilon = 0$, $\widetilde{C}_\epsilon = C$ and $\mathbf{v}_j(\epsilon) := \epsilon$ for $1 \leqslant j \leqslant d$.
- At node $\mathbf{v}$, given $(\tau_{\mathbf{v}'}, C_{\mathbf{v}'}, \mathbf{v}_j(\mathbf{v}'))_{\mathbf{v}' \sqsubseteq \mathbf{v}}$ (*i.e.*, given $(\widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}, \widetilde{T}_{\mathbf{v}'})_{\mathbf{v}' \sqsubset \mathbf{v}}$) define:

$$\widetilde{T}_{\mathbf{v}} = \min_{1\leqslant j\leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}, \qquad \widetilde{J}_{\mathbf{v}} := \operatorname*{arg\,min}_{1\leqslant j\leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}, \qquad \widetilde{S}_{\mathbf{v}} = S'^j_{\mathbf{v}_j(\mathbf{v})}, \tag{2.49}$$

$$\mathbf{v}_j(\mathbf{v}a) = \begin{cases} \mathbf{v}_j(\mathbf{v})a & \text{if } j = \widetilde{J}_{\mathbf{v}} \\ \mathbf{v}_j(\mathbf{v}) & \text{else.} \end{cases} \tag{2.50}$$

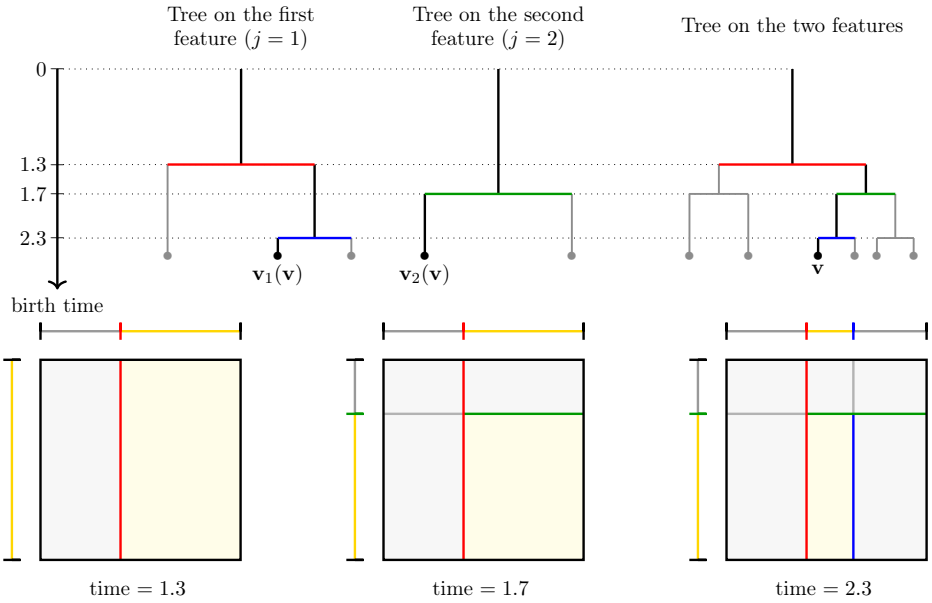


Figure 2.3: Modified construction in dimension two. At the top, from left to right: trees associated to partitions Π'^1, Π'^2 and $\widetilde{\Pi}$ respectively. At the bottom, from left to right: successive splits in $\widetilde{\Pi}$ leading to the leaf $\mathbf{v}$ (depicted in yellow).

Finally, for every $\lambda \in \mathbf{R}^+$, define $\widetilde{\Pi}_\lambda$ and $\widetilde{K}_\lambda$ as before from $\widetilde{\Pi}$. This construction is illustrated in Figure 2.3.

Computation of $\mathbb{E}[\widetilde{K}_\lambda]$. Now, it can be seen that the partition $\widetilde{\Pi}_\lambda$ is a rectangular grid which is the "product" of the partitions Π'^j of the intervals C^j, $1 \leqslant j \leqslant d$. Indeed, let $x \in [0,1]^d$, and let $\widetilde{C}_\lambda(x)$ be the cell in $\widetilde{\Pi}_\lambda$ that contains x; we need to show that $\widetilde{C}_\lambda(x) = \prod_{j=1}^d C'^j_\lambda(x)$, where $C'^j_\lambda(x)$ is the subinterval of C^j in the partition Π'^j that contains x_j. The proof proceeds in several steps:

- First, Equation (2.49) shows that, for every node $\mathbf{v}$, we have $\widetilde{C}_\mathbf{v} = \prod_{1 \leqslant j \leqslant d} C'^j_{\mathbf{v}_j(\mathbf{v})}$, since the successive splits on the j-th coordinate of $\widetilde{C}_\mathbf{v}$ are precisely the ones of $C'^j_{\mathbf{v}_j(\mathbf{v})}$.
- Second, it follows from (2.49) that $\widetilde{T}_\mathbf{v} = \min_{1 \leqslant j \leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}$; also, since the cell $C_\mathbf{v}$ is formed when its last split is performed, $\widetilde{\tau}_\mathbf{v} = \max_{1 \leqslant j \leqslant d} \tau'^j_{\mathbf{v}_j(\mathbf{v})}$.
- Let $\widetilde{\mathbf{v}}$ be the node such that $\widetilde{C}_{\widetilde{\mathbf{v}}} = \widetilde{C}_\lambda(x)$, and $\mathbf{v}'^j$ be such that $C'^j_{\mathbf{v}'^j} = C'^j_\lambda(x_j)$. By the first point, it suffices to show that $\mathbf{v}_j(\widetilde{\mathbf{v}}) = \mathbf{v}'_j$ for $1 \leqslant j \leqslant d$.
- Observe that $\widetilde{\mathbf{v}}$ (resp. $\mathbf{v}'_j$) is characterized by the fact that $x \in \widetilde{C}_{\widetilde{\mathbf{v}}}$ and $\widetilde{\tau}_{\widetilde{\mathbf{v}}} \leqslant \lambda < \widetilde{T}_{\widetilde{\mathbf{v}}}$ (resp. $x_j \in C'^j_{\mathbf{v}'^j}$ and $\tau'^j_{\mathbf{v}'^j} \leqslant \lambda < T'^j_{\mathbf{v}'^j}$). But since $\widetilde{C}_{\widetilde{\mathbf{v}}} = \prod_{1 \leqslant j \leqslant d} C'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$ (first point), $x \in \widetilde{C}_{\widetilde{\mathbf{v}}}$ implies $x_j \in C'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$. Likewise, since $\widetilde{\tau}_{\widetilde{\mathbf{v}}} = \max_{1 \leqslant j \leqslant d} \tau'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$ and $\widetilde{T}_{\widetilde{\mathbf{v}}} = \min_{1 \leqslant j \leqslant d} T'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$ (second point), $\widetilde{\tau}_{\widetilde{\mathbf{v}}} \leqslant \lambda < \widetilde{T}_{\widetilde{\mathbf{v}}}$ implies $\tau'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})} \leqslant \lambda < T'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$. Since these properties characterize $\mathbf{v}'^j$, we have $\mathbf{v}_j(\widetilde{\mathbf{v}}) = \mathbf{v}'_j$, which concludes the proof.

Hence, the partition $\widetilde{\Pi}_\lambda$ is the product of the partitions $\Pi'^j = \Phi_{C^j}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}})_\lambda$ of the intervals C^j, $1 \leqslant j \leqslant d$, which are independent Mondrians distributed as $\mathsf{MP}(\lambda, C^j)$. By Fact 2.1, the splits of the Mondrian partition $\mathsf{MP}(\lambda, C^j)$ are distributed as a Poisson point process on C^j of intensity λ, so that the expected number of cells in such a partition is $1 + \lambda|C^j|$. Since $\widetilde{\Pi}_\lambda$ is a "product" of such independent partitions, we have:

$$\mathbb{E}[\widetilde{K}_\lambda] = \prod_{j=1}^{d} (1 + \lambda|C^j|)\,. \tag{2.51}$$

Equality of $\mathbb{E}[K_\lambda]$ and $\mathbb{E}[\widetilde{K}_\lambda]$. In order to establish Proposition 2.2, it is thus sufficient to prove that $\mathbb{E}[K_\lambda] = \mathbb{E}[\widetilde{K}_\lambda]$. First, note that, since the number of cells in a partition is one plus the number of splits (each split increases the number of cells by one)

$$K_\lambda = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbf{1}(T_{\mathbf{v}} \leqslant \lambda)$$

so that we have, respectively,

$$\mathbb{E}[K_\lambda] = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbb{P}(T_{\mathbf{v}} \leqslant \lambda) \tag{2.52}$$

$$\mathbb{E}[\widetilde{K}_\lambda] = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbb{P}(\widetilde{T}_{\mathbf{v}} \leqslant \lambda)\,. \tag{2.53}$$

Hence, it suffices to show that $\mathbb{P}(T_{\mathbf{v}} \leqslant \lambda) = \mathbb{P}(\widetilde{T}_{\mathbf{v}} \leqslant \lambda)$ for every $\mathbf{v} \in \{0,1\}^*$ and $\lambda \geqslant 0$, *i.e.* that $T_{\mathbf{v}}$ and $\widetilde{T}_{\mathbf{v}}$ have the same distribution for every $\mathbf{v}$.

In order to establish this, we show that, for every $\mathbf{v} \in \{0,1\}^*$, the conditional distribution of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathcal{F}}_{\mathbf{v}} = \sigma((\widetilde{T}_{\mathbf{v}'}, \widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$ has the same form as the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathcal{F}_{\mathbf{v}} = \sigma((T_{\mathbf{v}'}, J_{\mathbf{v}'}, S_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$, in the sense that there exits a family of conditional distributions $(\Psi_{\mathbf{v}})_{\mathbf{v}}$ such that, for every $\mathbf{v}$, the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathcal{F}_{\mathbf{v}}$ is $\Psi_{\mathbf{v}}(\cdot|(T_{\mathbf{v}'}, J_{\mathbf{v}'}, S_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$ and the conditional distribution of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathcal{F}}_{\mathbf{v}}$ is $\Psi_{\mathbf{v}}(\cdot|(\widetilde{T}_{\mathbf{v}'}, \widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$.

First, recall that the variables $(E^j_{\mathbf{v}'}, U^j_{\mathbf{v}'})_{\mathbf{v}' \in \{0,1\}^*, 1 \leqslant j \leqslant d}$ are independent, so $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{1 \leqslant j \leqslant d}$ is independent from $\mathcal{F}_{\mathbf{v}}$. Hence, conditionally on $\mathcal{F}_{\mathbf{v}}$, $E^j_{\mathbf{v}}, U^j_{\mathbf{v}}$, $1 \leqslant j \leqslant d$ are independent with $E^j_{\mathbf{v}} \sim \mathsf{Exp}(1)$ and $U^j_{\mathbf{v}} \sim \mathcal{U}([0,1])$. Also, recall that if $T_1, \ldots, T_d$ are independent exponential random variables of intensities $\lambda_1, \ldots, \lambda_d$, and if $T = \min_{1 \leqslant j \leqslant d} T_j$ and $J = \arg\min_{1 \leqslant j \leqslant d} T_j$, then $\mathbb{P}(J = j) = \lambda_j / \sum_{j'=1}^d \lambda_{j'}$, $T \sim \mathsf{Exp}(\sum_{j=1}^d \lambda_j)$ and J and T are independent. Hence, conditionally on $\mathcal{F}_{\mathbf{v}}$, $T_{\mathbf{v}} - \tau_{\mathbf{v}} = \min_{1 \leqslant j \leqslant d} E^j_{\mathbf{v}}/|C^j_{\mathbf{v}}| \sim \mathsf{Exp}(\sum_{j=1}^d |C^j_{\mathbf{v}}|) = \mathsf{Exp}(|C_{\mathbf{v}}|)$, $J_{\mathbf{v}} := \arg\min_{1 \leqslant j \leqslant d} E^j_{\mathbf{v}}/|C^j_{\mathbf{v}}|$ equals j with probability $|C^j_{\mathbf{v}}|/|C_{\mathbf{v}}|$, $T_{\mathbf{v}}, J_{\mathbf{v}}$ are independent and $(S_{\mathbf{v}}|T_{\mathbf{v}}, J_{\mathbf{v}}) \sim \mathcal{U}(C^{J_{\mathbf{v}}}_{\mathbf{v}})$.

Now consider the conditional distribution of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathcal{F}}_{\mathbf{v}}$. Let $(\mathbf{v}_v)_{v \in \mathbf{N}}$ be a path in $\{0,1\}^*$ from the root: $\mathbf{v}_0 := \epsilon$, $\mathbf{v}_{v+1}$ is a child of $\mathbf{v}_v$ for $v \in \mathbf{N}$, and $\mathbf{v}_v \sqsubseteq \mathbf{v}$ for $0 \leqslant v \leqslant \mathrm{depth}(\mathbf{v})$. Define for $v \in \mathbf{N}$, $E^j_v = E^j_{\mathbf{v}_v}$ and $U^j_v = U^j_{\mathbf{v}_v}$ if $\mathbf{v}_{v+1}$ is the left child of $\mathbf{v}_v$, and $1 - U^j_{\mathbf{v}_v}$ otherwise. Then, the variables $(E^j_v, U^j_v)_{v \in \mathbf{N}, 1 \leqslant j \leqslant d}$ are independent, with $E^j_v \sim \mathsf{Exp}(1)$, $U^j_v \sim \mathcal{U}([0,1])$, so that the following Lemma applies.

Lemma 2.2. *Let $(E^j_v, U^j_v)_{v \in \mathbf{N}^\star, 1 \leqslant j \leqslant d}$ be a family of independent random variables, with $U^j_v \sim \mathcal{U}([0,1])$ and $E^j_v \sim \mathsf{Exp}(1)$. Let $a_1, \ldots, a_d > 0$. For $1 \leqslant j \leqslant d$, define the sequence $(T^j_v, L^j_v)_{v \in \mathbf{N}}$ as follows*:

- $L_0^j = a_j$, $T_0^j = \frac{E_0^j}{a_j}$;
- *for* $v \in \mathbf{N}$, $L_{v+1}^j = U_v^j L_v^j$, $T_{v+1}^j = T_v^j + \frac{E_{v+1}^j}{L_{v+1}^j}$.

Define recursively the variables $\widetilde{V}_v^j$ $(v \in \mathbf{N}, 1 \leqslant j \leqslant d)$ *as well as* $\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v$ $(v \in \mathbf{N})$ *as follows*:

- $\widetilde{V}_0^j = 0$ *for* $j = 1, \dots, d$.
- *for* $v \in \mathbf{N}$, *given* $\widetilde{V}_v^j$ $(1 \leqslant j \leqslant d)$, *denoting* $\widetilde{T}_v^j = T_{\widetilde{V}_v^j}^j$ *and* $\widetilde{U}_v^j = U_{\widetilde{V}_v^j}^j$, *set*

$$\widetilde{J}_v = \operatorname*{arg\,min}_{1 \leqslant j \leqslant d} \widetilde{T}_v^j, \quad \widetilde{T}_v = \min_{1 \leqslant j \leqslant d} \widetilde{T}_v^j = \widetilde{T}_v^{\widetilde{J}_v}, \quad \widetilde{U}_v = \widetilde{U}_v^{\widetilde{J}_v}, \text{ and } \quad \widetilde{V}_{v+1}^j = \widetilde{V}_v^j + \mathbf{1}(\widetilde{J}_v = j).$$

Then, the conditional distribution of $(\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v)$ *given* $\mathcal{F}_v = \sigma((\widetilde{J}_{v'}, \widetilde{T}_{v'}, \widetilde{U}_{v'}),\ 0 \leqslant v' < v)$ *is the following* (*denoting* $\widetilde{L}_v^j = L_{\widetilde{V}_v^j}^j$):

- $\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v$ *are independent*,
- $\mathbb{P}(\widetilde{J}_v = j | \mathcal{F}_v) = \widetilde{L}_v^j / (\sum_{j'=1}^d \widetilde{L}_v^{j'})$,
- $\widetilde{T}_v - \widetilde{T}_{v-1} \sim \mathsf{Exp}(\sum_{j=1}^d \widetilde{L}_v^j)$ (*with the convention* $\widetilde{T}_{-1} = 0$) *and* $\widetilde{U}_v \sim \mathcal{U}([0,1])$.

In addition, note that, with the notations of Lemma 2.2, a simple induction shows that $\widetilde{J}_v = \widetilde{J}_{\mathbf{v}_v}$, $\widetilde{T}_v = \widetilde{T}_{\mathbf{v}_v}$, $\widetilde{U}_v = \widetilde{U}_{\mathbf{v}_v}$ and $L_v^j = |\widetilde{C}_{\mathbf{v}_v}^j|$, so that $\mathcal{F}_v = \mathcal{F}_{\mathbf{v}_v}$. Applying Lemma 2.2 for $v = \mathrm{depth}(\mathbf{v})$ (so that $\mathbf{v}_v = \mathbf{v}$) therefore gives the following: conditionally on $\mathcal{F}_{\mathbf{v}}$, the variables $\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{U}_{\mathbf{v}}$ are independent, $\widetilde{T}_{\mathbf{v}} - \widetilde{\tau}_{\mathbf{v}} \sim \mathsf{Exp}(|\widetilde{C}_{\mathbf{v}}^j|)$, $\mathbb{P}(\widetilde{J}_{\mathbf{v}} = j | \mathcal{F}_{\mathbf{v}}) = |\widetilde{C}_{\mathbf{v}}^j| / \big(\sum_{j'=1}^d |\widetilde{C}_{\mathbf{v}}^j|\big)$ and $\widetilde{U}_{\mathbf{v}} \sim \mathcal{U}([0,1])$, so that $(\widetilde{S}_{\mathbf{v}} | \mathcal{F}_{\mathbf{v}}, \widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}) \sim \mathcal{U}(\widetilde{C}_{\mathbf{v}}^{\widetilde{J}_{\mathbf{v}}})$. Hence, we have proven that, for every $\mathbf{v}$, the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathcal{F}_{\mathbf{v}}$ is the same as that of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathcal{F}}_{\mathbf{v}}$. By induction on $\mathbf{v}$, since $\mathcal{F}_\epsilon = \widetilde{\mathcal{F}}_\epsilon$ is the trivial σ-algebra, this shows that $T_{\mathbf{v}}$ and $\widetilde{T}_{\mathbf{v}}$ have the same distribution for every $\mathbf{v}$. Plugging this into (2.52) and (2.53) and combining it with (2.51) completes the proof of Proposition 2.2. □

Proof of Lemma 2.2. We show by induction on $v \in \mathbf{N}$ the following property: conditionally on $\mathcal{F}_{\mathbf{v}}$, $(\widetilde{T}_v^j, \widetilde{U}_v^j)_{1 \leqslant j \leqslant d}$ are independent, $\widetilde{T}_v^j - \widetilde{T}_{v-1} \sim \mathsf{Exp}(L_v^j)$ and $\widetilde{U}_v^j \sim \mathcal{U}([0,1])$.

Initialization For $v = 0$ (with $\mathcal{F}_0$ the trivial σ-algebra), since $\widetilde{V}_0^j = 0$ we have $\widetilde{T}_0^j = E_0^j / a_j \sim \mathsf{Exp}(a_j) = \mathsf{Exp}(L_0^j)$, $\widetilde{U}_0^j = U_0^j \sim \mathcal{U}([0,1])$ and these random variables are independent.

Inductive step Let $v \in \mathbf{N}$, and assume the property is true up to step v. Conditionally on $\mathcal{F}_{v+1}$, *i.e.* on $\mathcal{F}_v, \widetilde{T}_v, \widetilde{J}_v, \widetilde{U}_v$, we have:

- for $j \neq \widetilde{J}_v$, the variables $\widetilde{T}_{v+1}^j - \widetilde{T}_{v-1} = \widetilde{T}_v^j - \widetilde{T}_{v-1}$ are independent $\mathsf{Exp}(\widetilde{L}_v^j) = \mathsf{Exp}(\widetilde{L}_{v+1}^j)$ random variables (when conditioned only on $\mathcal{F}_v$, by the induction hypothesis), conditioned on $\widetilde{T}_{v+1}^j - \widetilde{T}_{v-1} \geqslant \widetilde{T}_v - \widetilde{T}_{v-1}$, so by the memory-less property of exponential random variables $\widetilde{T}_{v+1}^j - \widetilde{T}_v = (\widetilde{T}_{v+1}^j - \widetilde{T}_{v-1}) - (\widetilde{T}_v - \widetilde{T}_{v-1}) \sim \mathsf{Exp}(\widetilde{L}_{v+1}^j)$ (and those variables are independent).

- for $j \neq \widetilde{J}_v$, the variables $\widetilde{U}^j_{v+1} = \widetilde{U}^j_v$ are independent $\mathcal{U}([0,1])$ random variables (conditionally on $\mathcal{F}_v$), conditioned on the independent variables $\widetilde{T}_v, \widetilde{J}_v, \widetilde{U}_v$, so they remain independent $\mathcal{U}([0,1])$ random variables.
- $(\widetilde{T}^{\widetilde{J}_v}_{v+1} - \widetilde{T}_v, \widetilde{U}^{\widetilde{J}_v}_{v+1}) = (E^{\widetilde{J}_v}_{\widetilde{V}^{\widetilde{J}_v}_{v+1}} / \widetilde{L}^{\widetilde{J}_v}_{v+1}, U^{\widetilde{J}_v}_{\widetilde{V}^{\widetilde{J}_v}_{v+1}})$ is distributed, conditionally on $\mathcal{F}_{v+1}$, *i.e.* on $\widetilde{J}_v, \widetilde{T}_v, \widetilde{V}^{\widetilde{J}_v}_{v+1}, \widetilde{L}^{\widetilde{J}_v}_{v+1}$, as $\mathsf{Exp}(\widetilde{L}^{\widetilde{J}_v}_{v+1}) \otimes \mathcal{U}([0,1])$, and independent of $(\widetilde{T}^j_{v+1}, \widetilde{U}^j_{v+1})_{j \neq \widetilde{J}_v}$.

This completes the proof by induction.

Let $v \in \mathbf{N}$. We have established that, conditionally on $\mathcal{F}_v$, the variables $(\widetilde{T}^j_v, \widetilde{U}^j_v)_{1 \leqslant j \leqslant d}$ are independent, with $\widetilde{T}^j_v - \widetilde{T}_{v-1} \sim \mathsf{Exp}(\widetilde{L}^j_v)$ and $\widetilde{U}^j_v \sim \mathcal{U}([0,1])$. In particular, conditionally on $\mathcal{F}_v$, $\widetilde{U}_v$ is independent from $(\widetilde{J}_v, \widetilde{T}_v)$, $\widetilde{U}_v \sim \mathcal{U}([0,1])$, and (by the property of the minimum of independent exponential random variables) $\widetilde{J}_v$ is independent of $\widetilde{T}_v$, $\widetilde{T}_v \sim \mathsf{Exp}(\sum_{j=1}^d \widetilde{L}^j_v)$ and $\mathbb{P}(\widetilde{J}_v = j | \mathcal{F}_v) = \widetilde{L}^j_v / (\sum_{j'=1}^d \widetilde{L}^{j'}_v)$. This concludes the proof of Lemma 2.2. □

2.8.3 Proof of Theorem 2.1

Recall that a Mondrian Forest estimate with lifetime parameter λ is defined, for all $x \in [0,1]^d$, by

$$\widehat{f}_{\lambda,n,M}(x) = \widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}^{(m)}_{\lambda,n}(x, \Pi^{(m)}_\lambda),$$

where $\widehat{f}^{(m)}_{\lambda,n}(x, \Pi^{(m)}_\lambda)$ denotes the Mondrian Tree based on the random partition $\Pi^{(m)}_\lambda$ and $\Pi_{\lambda,M} = (\Pi^{(1)}_\lambda, \ldots, \Pi^{(M)}_\lambda)$. To ease notation, we will write $\widehat{f}^{(m)}_{\lambda,n}(x)$ instead of $\widehat{f}^{(m)}_{\lambda,n}(x, \Pi^{(m)}_\lambda)$. First, note that, by Jensen's inequality,

$$\begin{aligned} R(\widehat{f}_{\lambda,n,M}) &= \mathbb{E}_{(X,\Pi_{\lambda,M})}[(\widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) - f(X))^2] \\ &\leqslant \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{(X,\Pi^{(m)}_\lambda)}[(\widehat{f}^{(m)}_{\lambda,n}(X) - f(X))^2] \\ &\leqslant \mathbb{E}_{(X,\Pi^{(1)}_\lambda)}[(\widehat{f}^{(1)}_{\lambda,n}(X) - f(X))^2], \end{aligned}$$

since each Mondrian tree has the same distribution. Therefore, it is sufficient to prove that a single Mondrian tree is consistent. Now, since Mondrian partitions are independent of the dataset $\mathscr{D}_n$, we can apply Theorem 4.2 from Györfi et al. (2002), which states that a Mondrian tree estimate is consistent if, as $n \to \infty$,

(*i*) $D_\lambda(X) \to 0$ in probability, and

(*ii*) $K_\lambda / n \to 0$ in probability,

where $D_\lambda(X)$ is the diameter of the cell of the Mondrian tree that contains X, and K_λ is the number of cells in the Mondrian tree. Note that the initial assumptions in Theorem 4.2 in Györfi et al. (2002) contains deterministic convergence, but can be relaxed to convergences in probability by a close inspection of the proof. Hence, in order to conclude the proof, it suffices to establish (*i*) and (*ii*). The first condition follows from the fact that, by Corollary 2.1,

$$\mathbb{E}[D_\lambda(X)^2] = \mathbb{E}[\mathbb{E}[D_\lambda(X)^2 | X]] \leqslant \frac{4d}{\lambda^2}$$

as well as the assumption that $\lambda_n \to \infty$. Condition (ii) follows from Proposition 2.2 and the assumption $\lambda_n^d/n \to 0$. This concludes the proof. □

2.8.4 Proof of Proposition 2.3

Let $\Pi_\lambda^{(1)}$ be the Mondrian partition of $[0,1]$ used to construct the randomized estimator $\widehat{f}_{\lambda,n}^{(1)}$. Denote by $\bar{f}_\lambda^{(1)}$ the random function $\bar{f}_\lambda^{(1)}(x) = \mathbb{E}_X\left[f(X)|X \in C_\lambda(x)\right]$, and define $\widetilde{f}_\lambda(x) = \mathbb{E}[\bar{f}_\lambda^{(1)}(x)]$ (which is deterministic). For the seek of clarity, we will drop the exponent "(1)" in all notations, keeping in mind that we consider only one particular Mondrian partition, whose associated Mondrian Tree estimate is denoted by $\widehat{f}_{\lambda,n}$. Recall the bias-variance decomposition (2.20) for Mondrian trees:

$$R(\widehat{f}_{\lambda,n}^{(1)}) = \mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_{\lambda,n}^{(1)}(X))^2\big]\,. \tag{2.54}$$

We will provide lower bounds for the first term (the bias, depending on λ) and the second (the variance, depending on both λ and n), which will lead to the stated lower bound on the risk, valid for every value of λ.

Lower bound on the bias. As we will see, the point-wise bias $\mathbb{E}[(\bar{f}_\lambda(x) - f(x))^2]$ can be computed explicitly given our assumptions. Let $x \in [0,1]$. Since $\widetilde{f}_\lambda(x) = \mathbb{E}[\bar{f}_\lambda(x)]$, we have

$$\mathbb{E}\left[(\bar{f}_\lambda(x) - f(x))^2\right] = \mathrm{Var}(\bar{f}_\lambda(x)) + (\widetilde{f}_\lambda(x) - f(x))^2\,. \tag{2.55}$$

By Proposition 2.1, the cell of x in Π_λ can be written as $C_\lambda(x) = [L_\lambda(x), R_\lambda(x)]$, with $L_\lambda(x) = (x - \lambda^{-1}E_L) \vee 0$ and $R_\lambda(x) = (x + \lambda^{-1}E_R) \wedge 1$, where E_L, E_R are two independent $\mathsf{Exp}(1)$ random variables. Now, since $X \sim \mathcal{U}([0,1])$ and $f(u) = 1 + u$,

$$\bar{f}_\lambda(x) = \frac{1}{R_\lambda(x) - L_\lambda(x)} \int_{L_\lambda(x)}^{R_\lambda(x)} (1+u)\mathrm{d}u = 1 + \frac{L_\lambda(x) + R_\lambda(x)}{2}\,.$$

Since $L_\lambda(x)$ and $R_\lambda(x)$ are independent, we have

$$\mathrm{Var}(\bar{f}_\lambda(x)) = \frac{\mathrm{Var}(L_\lambda(x)) + \mathrm{Var}(R_\lambda(x))}{4}.$$

In addition,

$$\mathrm{Var}(R_\lambda(x)) = \mathrm{Var}\big(x + \lambda^{-1}[E_R \wedge \lambda(1-x)]\big) = \lambda^{-2}\mathrm{Var}(E_R \wedge [\lambda(1-x)])$$

Now, if $E \sim \mathsf{Exp}(1)$ and $a \geqslant 0$, we have

$$\begin{aligned}
\mathbb{E}[E \wedge a] &= \int_0^a u e^{-u}\mathrm{d}u + a\mathbb{P}(E \geqslant a) = 1 - e^{-a} \\
\mathbb{E}[(E \wedge a)^2] &= \int_0^a u^2 e^{-u}\mathrm{d}u + a^2\mathbb{P}(E \geqslant a) = 2\left(1 - (a+1)e^{-a}\right),
\end{aligned} \tag{2.56}$$

so that

$$\mathrm{Var}(E \wedge a) = \mathbb{E}[(E \wedge a)^2] - \mathbb{E}[E \wedge a]^2 = 1 - 2ae^{-a} - e^{-2a}.$$

The formula above gives the variances of $R_\lambda(x)$ and $L_\lambda(x)$ respectively:

$$\begin{aligned}\mathrm{Var}(R_\lambda(x)) &= \lambda^{-2}\big(1 - 2\lambda(1-x)e^{-\lambda(1-x)} - e^{-2\lambda(1-x)}\big)\\ \mathrm{Var}(L_\lambda(x)) &= \lambda^{-2}\big(1 - 2\lambda x e^{-\lambda x} - e^{-2\lambda x}\big)\,,\end{aligned}$$

and thus

$$\mathrm{Var}(\bar{f}_\lambda(x)) = \frac{1}{4\lambda^2}\big(2 - 2\lambda x e^{-\lambda x} - 2\lambda(1-x)e^{-\lambda(1-x)} - e^{-2\lambda x} - e^{-2\lambda(1-x)}\big)\,. \tag{2.57}$$

In addition, the formula (2.56) yields

$$\begin{aligned}\mathbb{E}[R_\lambda(x)] &= x + \lambda^{-1}\big(1 - e^{-\lambda(1-x)}\big)\\ \mathbb{E}[L_\lambda(x)] &= x - \lambda^{-1}\big(1 - e^{-\lambda x}\big)\,,\end{aligned}$$

and thus

$$\widetilde{f}_\lambda(x) = 1 + \frac{\mathbb{E}[L_\lambda(x)] + \mathbb{E}[R_\lambda(x)]}{2} = 1 + x + \frac{1}{2\lambda}\big(e^{-\lambda x} - e^{-\lambda(1-x)}\big)\,. \tag{2.58}$$

Combining (2.57) and (2.58) with the decomposition (2.55) gives

$$\mathbb{E}\big[\big(\bar{f}_\lambda(x) - f(x)\big)^2\big] = \frac{1}{2\lambda^2}\Big(1 - \lambda x e^{-\lambda x} - \lambda(1-x)e^{-\lambda(1-x)} - e^{-\lambda}\Big)\,. \tag{2.59}$$

Integrating over X, we obtain

$$\begin{aligned}\mathbb{E}\left[(\bar{f}_\lambda(X) - f(X))^2\right] &= \frac{1}{2\lambda^2}\left(1 - \int_0^1 \lambda x e^{-\lambda x}\mathrm{d}x - \int_0^1 \lambda(1-x)e^{-\lambda(1-x)}\mathrm{d}x - e^{-\lambda}\right)\\ &= \frac{1}{2\lambda^2}\left(1 - 2\times\frac{1}{\lambda}\big(1 - (\lambda+1)e^{-\lambda}\big) - e^{-\lambda}\right)\\ &= \frac{1}{2\lambda^2}\left(1 - \frac{2}{\lambda} + e^{-\lambda} + \frac{2}{\lambda}e^{-\lambda}\right)\,.\end{aligned} \tag{2.60}$$

Now, note that the bias $\mathbb{E}[(\bar{f}_\lambda(X) - f(X))^2]$ is positive for $\lambda \in \mathbf{R}_+^*$ (indeed, it is nonnegative, and non-zero since f is not piecewise constant). In addition, the expression (2.60) shows that it is continuous in λ on $\mathbf{R}_+^*$, and that it admits a limit $\frac{1}{12}$ as $\lambda \to 0$ (using the fact that $e^{-\lambda} = 1 - \lambda + \frac{\lambda^2}{2} - \frac{\lambda^3}{6} + o(\lambda^3)$). Hence, the function $\lambda \mapsto \mathbb{E}[(\bar{f}_\lambda(X) - f(X))^2]$ is positive and continuous on $\mathbf{R}_+$, so that it admits a minimum $C_1 > 0$ on the compact interval $[0, 6]$. In addition, the expression (2.60) shows that for $\lambda \geqslant 6$, we have

$$\mathbb{E}\left[(\bar{f}_\lambda(X) - f(X))^2\right] \geqslant \frac{1}{2\lambda^2}\left(1 - \frac{2}{6}\right) = \frac{1}{3\lambda^2}\,. \tag{2.61}$$

First lower bound on the variance. We now turn to the task of bounding the variance from below. In order to avoid restrictive conditions on λ, we will provide two separate lower bounds, valid in two different regimes.

Our first lower bound on the variance, valid for $\lambda \leqslant n/3$, controls the error of estimation of the optimal labels in nonempty cells. It depends on σ^2, and is of order $\Theta\big(\sigma^2\frac{\lambda}{n}\big)$. We use a general bound on the variance of regressograms (Arlot and Genuer, 2014, Proposition 2)

(note that while this result is stated for a fixed number of cells, it can be adapted to a random number of cells by conditioning on $K_\lambda = k$ and then by averaging):

$$\mathbb{E}\left[\left(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\right)^2\right] \geqslant \frac{\sigma^2}{n}\left(\mathbb{E}\left[K_\lambda\right] - 2\mathbb{E}_{\Pi_\lambda}\left[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \exp(-n\mathbb{P}(X \in C_\mathbf{v}))\right]\right). \tag{2.62}$$

Now, recall that the splits defining Π_λ form a Poisson point process on $[0,1]$ of intensity $\lambda \mathrm{d}x$ (Fact 2.1). In particular, the splits can be described as follows. Let $(E_k)_{k\geqslant 1}$ be an i.i.d. sequence of $\mathsf{Exp}(1)$ random variables, and $S_p := \sum_{k=1}^p E_k$ for $p \geqslant 0$. Then, the (ordered) splits in Π_λ have the same distribution as $(\lambda^{-1}S_1, \dots, \lambda^{-1}S_{K_\lambda - 1})$, where $K_\lambda := 1 + \sup\{p \geqslant 0 : S_p \leqslant \lambda\}$. In addition, the probability that $X \sim \mathcal{U}([0,1])$ falls in the cell $[\lambda^{-1}S_{k-1}, \lambda^{-1}S_k \wedge 1)$ $(1 \leqslant k \leqslant K_\lambda)$ is $\lambda^{-1}(S_k \wedge 1 - S_{k-1})$, so that

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \exp(-n\mathbb{P}\left(X \in C_\mathbf{v}\right))\Bigg] &= \mathbb{E}\Bigg[\sum_{k=1}^{K_\lambda - 1} e^{-n\lambda^{-1}(S_k - S_{k-1})} + e^{-n(1-\lambda^{-1}S_{K_\lambda - 1})}\Bigg] \\
&\leqslant \mathbb{E}\Bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_k \leqslant \lambda) e^{-n\lambda^{-1}E_k}\Bigg] + 1 = \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbf{1}(S_k \leqslant \lambda)\big]\mathbb{E}\big[e^{-n\lambda^{-1}E_k}\big] + 1 \qquad (2.63) \\
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbf{1}(S_k \leqslant \lambda)\big] \cdot \int_0^\infty e^{-n\lambda^{-1}u} e^{-u}\mathrm{d}u + 1 = \frac{\lambda}{n+\lambda}\mathbb{E}\Bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_k \leqslant \lambda)\Bigg] + 1 \\
&= \frac{\lambda}{n+\lambda}\mathbb{E}\left[K_\lambda\right] + 1 = \frac{\lambda}{n+\lambda}(1+\lambda) + 1 \qquad (2.64)
\end{aligned}$$

where (2.63) comes from the fact that E_k and S_{k-1} are independent. Plugging Equation (2.64) in the lower bound (2.62) yields

$$\mathbb{E}\left[\left(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\right)^2\right] \geqslant \frac{\sigma^2}{n}\left((1+\lambda) - 2(1+\lambda)\frac{\lambda}{n+\lambda} - 2\right) = \frac{\sigma^2}{n}\left((1+\lambda)\frac{n-\lambda}{n+\lambda} - 2\right).$$

Now, assume that $6 \leqslant \lambda \leqslant \frac{n}{3}$. Since

$$(1+\lambda)\frac{n-\lambda}{n+\lambda} - 2 \underset{(\lambda\leqslant n/3)}{\geqslant} (1+\lambda)\frac{n-n/3}{n+n/3} - 2 = (1+\lambda)\frac{1}{2} - 2 \underset{(\lambda\geqslant 6)}{\geqslant} \frac{\lambda}{4},$$

the above lower bound implies, for $6 \leqslant \lambda \leqslant \frac{n}{3}$,

$$\mathbb{E}\left[\left(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\right)^2\right] \geqslant \frac{\sigma^2\lambda}{4n}. \tag{2.65}$$

Second lower bound on the variance. The lower bound (2.65) is only valid for $\lambda \leqslant n/3$; as λ becomes of order n or larger, the previous bound becomes vacuous. We now provide another lower bound on the variance, valid when $\lambda \geqslant n/3$, by considering the contribution of empty cells to the variance.

Let $\mathbf{v} \in \mathcal{L}(\Pi_\lambda)$. If $C_\mathbf{v}$ contains no sample point from $\mathscr{D}_n$, then for $x \in C_\mathbf{v}$: $\widehat{f}_{\lambda,n}(x) = 0$ and thus $(\widehat{f}_{\lambda,n}(x) - \bar{f}_\lambda(x))^2 = \bar{f}_\lambda(x)^2 \geqslant 1$. Hence, the variance term is lower bounded as follows,

denoting $N_n(C)$ the number of $1 \leqslant i \leqslant n$ such that $X_i \in C$ and $N_{\lambda,n}(x) = N_n(C_\lambda(x))$:

$$\begin{aligned}
&\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant \mathbb{P}\big(N_{\lambda,n}(X) = 0\big) \\
&= \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\,\mathbb{P}(N_n(C_\mathbf{v}) = 0)\bigg] = \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)^n\bigg] \\
&\geqslant \mathbb{E}\bigg[\Big(\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)\Big)^n\bigg] \qquad (2.66) \\
&\geqslant \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)\bigg]^n \qquad (2.67) \\
&= \bigg(1 - \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})^2\bigg]\bigg)^n \qquad (2.68)
\end{aligned}$$

where (2.66) and (2.67) come from Jensen's inequality applied to the convex function $x \mapsto x^n$. Now, using the notations defined above, we have

$$\begin{aligned}
&\mathbb{E}\bigg[\sum_{\mathbf{v}\in\Pi_\lambda} \mathbb{P}(X \in C_\mathbf{v})^2\bigg] \leqslant \mathbb{E}\bigg[\sum_{k=1}^{K_\lambda} (\lambda^{-1}E_k)^2\bigg] \\
&= \lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda) E_k^2\bigg] = \lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda)\mathbb{E}\big[E_k^2 | S_{k-1}\big]\bigg] \\
&= 2\lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda)\bigg] \qquad (2.69) \\
&= 2\lambda^{-2}\,\mathbb{E}\left[K_\lambda\right] = \frac{2(\lambda+1)}{\lambda^2}\,, \qquad (2.70)
\end{aligned}$$

where the equality $\mathbb{E}[E_k^2|S_{k-1}] = 2$ (used in Equation (2.69)) comes from the fact that $E_k \sim \mathsf{Exp}(1)$ is independent of S_{k-1}.

The bounds (2.68) and (2.70) imply that, if $2(\lambda+1)/\lambda^2 \leqslant 1$, then

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant \left(1 - \frac{2(\lambda+1)}{\lambda^2}\right)^n. \qquad (2.71)$$

Now, assume that $n \geqslant 18$ and $\lambda \geqslant \frac{n}{3} \geqslant 6$. Then

$$\frac{2(\lambda+1)}{\lambda^2} \leqslant 2 \cdot \frac{3}{n}\left(1 + \frac{3}{n}\right) \leqslant 2 \cdot \frac{3}{n}\left(1 + \frac{3}{18}\right) = \frac{7}{n} \underset{(n\geqslant 18)}{\leqslant} 1\,,$$

so that, using the inequality $(1-x)^m \geqslant 1 - mx$ for $m \geqslant 0$ and $x \in \mathbf{R}$,

$$\left(1 - \frac{2(\lambda+1)}{\lambda^2}\right)^{n/8} \geqslant \left(1 - \frac{7}{n}\right)^{n/8} \geqslant 1 - \frac{n}{8}\cdot\frac{7}{n} = \frac{1}{8}\,.$$

Combining the above inequality with (2.71) gives, letting $C_2 := 1/8^8$,

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant C_2\,. \qquad (2.72)$$

Summing up. Assume that $n \geqslant 18$. Recall the bias-variance decomposition (2.54) of the risk $R(\widehat{f}_{\lambda,n})$ of the Mondrian tree.

- If $\lambda \leqslant 6$, we saw that the bias (and hence the risk) is larger than C_1;
- If $\lambda \geqslant \frac{n}{3}$, Equation (2.72) implies that the variance (and hence the risk) is larger than C_2;
- If $6 \leqslant \lambda \leqslant \frac{n}{3}$, Equations (2.61) (bias term) and (2.65) (variance term) imply that

$$R(\widehat{f}_{\lambda,n}) \geqslant \frac{1}{3\lambda^2} + \frac{\sigma^2 \lambda}{4n}\,.$$

In particular, letting $C_0 = C_1 \wedge C_2$, we conclude that

$$\inf_{\lambda \in \mathbf{R}^+} R(\widehat{f}_{\lambda,n}) \geqslant C_1 \wedge C_2 \wedge \inf_{\lambda \in \mathbf{R}^+} \left(\frac{1}{3\lambda^2} + \frac{\sigma^2 \lambda}{4n} \right) = C_0 \wedge \frac{1}{4} \left(\frac{3\sigma^2}{n} \right)^{2/3}. \tag{2.73}$$

2.8.5 Proof of Proposition 2.4

First, note that in all cases, since $|Y| \leqslant B$ almost surely, we also have $|\widehat{g}_n(X)| \leqslant B$ almost surely, so that $(Y - \widehat{g}_n(X))^2 \leqslant 4B^2$. Let $N_\varepsilon = |I_\varepsilon|$. Note that N_ε is a binomial variable with parameters $n - n_0 \geqslant n/2$ and $\mathbb{P}(X \in B_\varepsilon) \geqslant p_0(1-2\varepsilon)^d$ (since $p \geqslant p_0$). Now, recall Chernoff's bound: if $N \sim \mathsf{Bin}(m,p)$ and $\delta \in (0,1)$, then $\mathbb{P}(N \leqslant (1-\delta)mq) \leqslant e^{-mq\delta^2/2}$; in particular, $\mathbb{P}(N \leqslant mq/2) \leqslant e^{-mq/8}$. Hence, letting $c_1 = p_0(1-2\varepsilon)^d/4$,

$$\mathbb{P}\big(N_\varepsilon \leqslant c_1 n\big) \leqslant \exp(-c_1 n/4)\,. \tag{2.74}$$

Conditionally on I_ε, the sample $\mathscr{D}' = \{(X_i, Y_i) : i \in I_\varepsilon\}$ is an i.i.d. sample of size N_ε of the conditional distribution of (X,Y) given $X \in B_\varepsilon$; it is also independent of $\mathscr{D}_{n_0}$, and thus of the estimators $\widehat{f}_\alpha$, $\alpha = 0, \dots, A$. It follows from Theorem 1 in the supplementary material "Proof of the optimality of the empirical star algorithm" of Audibert (2008) that the estimator $\widehat{g}_n$ defined by (2.12) satisfies, with probability $1-\delta$ over the random sample $\mathscr{D}'$ conditionally on N_ε,

$$\begin{aligned} \mathbb{E}_{(X,Y)}\big[(\widehat{g}_n(X) - Y)^2 | X \in B_\varepsilon\big] - \min_{0 \leqslant \alpha \leqslant A} \mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X) - Y)^2 | X \in B_\varepsilon\big] \\ \leqslant \frac{CB^2 \log[(A+1)\delta^{-1}]}{N_\varepsilon} \end{aligned} \tag{2.75}$$

for every $\delta \in (0,1)$, where $C = 600$ and the expectation is taken with respect to an independent sample (X,Y) (the bound (2.75) is deduced from the aforementioned theorem by replacing Y by Y/B, which lies in $[-1,1]$). Since $Y = f(X) + \varepsilon$ with $\mathbb{E}[\varepsilon | X] = 0$, we have $\mathbb{E}[(g(X) - Y)^2 | X] = \mathbb{E}[(g(X) - f(X))^2 | X] + \mathbb{E}[\varepsilon^2 | X]$. Hence, inequality (2.75) writes

$$\begin{aligned} &\mathbb{E}_{(X,Y)}\big[(\widehat{g}_n(X) - f(X))^2 | X \in B_\varepsilon\big] \\ &\leqslant \min_{0 \leqslant \alpha \leqslant A} \mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X) - f(X))^2 | X \in B_\varepsilon\big] + \frac{CB^2 \log[(A+1)\delta^{-1}]}{N_\varepsilon}. \end{aligned}$$

By integrating the above inequality over the confidence level δ, we obtain

$$\begin{aligned}&\mathbb{E}_{(X,Y),\mathscr{D}'}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon,N_\varepsilon\big]\\&\leqslant\min_{0\leqslant\alpha\leqslant A}\mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]+\frac{CB^2[\log(A+1)+1]}{N_\varepsilon};\end{aligned}$$

by taking the expectation over $\mathscr{D}_{n_0}$, conditioning on $N_\varepsilon > c_1 n$, and recalling that $A \leqslant \log_2(n)$, we get

$$\begin{aligned}&\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon,N_\varepsilon>c_1n\big]\\&\leqslant\min_{0\leqslant\alpha\leqslant A}\mathbb{E}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]+\frac{CB^2[\log(1+\log_2 n)+1]}{c_1n}.\end{aligned}\tag{2.76}$$

Finally, combining the bounds (2.74) and (2.76) yields

$$\begin{aligned}&\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon\big]\\&\leqslant\mathbb{P}\left(N_\varepsilon\leqslant c_1n\right)\cdot4B^2+\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon,N_\varepsilon>c_1n\big]\\&\leqslant4B^2e^{-c_1n/4}+\min_{0\leqslant\alpha\leqslant A}\mathbb{E}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]+\frac{CB^2[\log(1+\log_2 n)+1]}{c_1n},\end{aligned}\tag{2.77}$$

which is precisely inequality (2.13).

Assume that f belongs to the class $\mathscr{C}^{p,\beta}(L)$, with $p \in \{0,1\}$, $\beta \in (0,1]$ and $L > 0$; we now proceed to show that $\widehat{g}_n$ achieves the minimax rate of estimation for this class. Let $s = p + \beta \in (0,2]$. If $p = 0$ (namely, $s \leqslant 1$), it follows from Theorem 2.2 (with the same adaptation as in the proof of Theorem 2.3 to bound the variance term conditionally on $X \in B_\varepsilon$) that, for every $\lambda > 0$,

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n_0,M}(X)-f(X))^2|X\in B_\varepsilon\big]\leqslant\frac{(4d)^sL^2}{\lambda^{2s}}+\frac{11B^2(1+\lambda)^d}{p_0(1-2\varepsilon)^dn_0}$$

(note that $\sigma, \|f\|_\infty \leqslant B$ since $|Y| \leqslant B$). It follows that, for some constants C_1, C_2 independent of λ, L, n,

$$\begin{aligned}\min_{0\leqslant\alpha\leqslant A}\mathbb{E}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]&\leqslant\min_{0\leqslant\alpha\leqslant A}\left[\frac{C_1L^2}{(2^\alpha)^{2s}}+\frac{C_2(1+2^\alpha)^d}{n}\right]\\&\leqslant4\min_{\lambda\in[1,n^{1/d}]}\left[\frac{C_1L^2}{\lambda^{2s}}+\frac{C_2(1+\lambda)^d}{n}\right],\end{aligned}\tag{2.78}$$

where we used the fact that, for every $\lambda \in [1, n^{1/d}]$, there exists some α, $0 \leqslant \alpha \leqslant A$, such that $\lambda/2 \leqslant 2^\alpha \leqslant \lambda$. It follows from (2.77) and (2.78) that

$$\begin{aligned}\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon\big]&=O\Big(\min_{0\leqslant\lambda\leqslant n^{1/d}}\left[\frac{C_1L^2}{\lambda^{2s}}+\frac{C_2(1+\lambda)^d}{n}\right]+\frac{\log\log n}{n}\Big)\\&=O\left(L^{2d/(d+2s)}n^{-2s/(d+2s)}\right)\end{aligned}$$

where the last bound follows from the fact that $\lambda_* = (L^2 n)^{1/(d+2s)}$ belongs to $[1, n^{1/d}]$ for n large enough (and $\log\log n/n = o(n^{2s/(d+2s)})$).

Now, consider the case $p = 1$, *i.e.*, $1 < s \leqslant 2$. It follows from Theorem 2.3 that for some constants C_3, C_4 independent of λ, L, n, we have for every $\lambda \in [1, n^{1/d}]$ (using the fact that $M \geqslant n^{2/d} \geqslant \lambda^2$, so that $1/(M\lambda^2) \leqslant 1/\lambda^4 \leqslant 1/\lambda^{2s}$, and $e^{-\lambda\varepsilon}/\lambda^3 = O(1/\lambda^{2s})$)

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \frac{C_3 L^2}{\lambda^{2s}} + \frac{C_4(1+\lambda)^d}{n}\,. \tag{2.79}$$

From the same argument as in the case $0 < s \leqslant 1$, combining inequalities (2.79) and (2.77) yields

$$\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 | X \in B_\varepsilon\big] = O\big(L^{2d/(d+2s)} n^{-2s/(d+2s)}\big)$$

which concludes the proof of Proposition 2.4. □

2.8.6 Proof of Lemma 2.1

According to Equation (2.31) from the main text, we have

$$F_\lambda(x,z) = \lambda^d \exp(-\lambda\|x-z\|_1) \prod_{1\leqslant j\leqslant d} G_\lambda(x_j, z_j) \tag{2.80}$$

where we defined, for $u, v \in [0,1]$,

$$\begin{aligned} G_\lambda(u,v) &= \mathbb{E}\left[(\lambda|u-v| + E_1 \wedge \lambda(u\wedge v) + E_2 \wedge \lambda(1-u\vee v))^{-1}\right] \\ &= H(\lambda|u-v|, \lambda u\wedge v, \lambda(1-u\vee v)) \end{aligned}$$

with E_1, E_2 two independent $\mathsf{Exp}(1)$ random variables, and $H : (\mathbf{R}_+^*)^3 \to \mathbf{R}$ the function defined by

$$H(a,b_1,b_2) = \mathbb{E}\left[(a + E_1 \wedge b_1 + E_2 \wedge b_2)^{-1}\right].$$

Also, let

$$H(a) = \mathbb{E}\left[(a + E_1 + E_2)^{-1}\right].$$

Denote

$$\begin{aligned} A &= \int_{[0,1]^d} (z-x) F_\lambda(x,z)\mathrm{d}z \\ B &= \int_{[0,1]^d} \frac{1}{2}\|z-x\|^2 F_\lambda(x,z)\mathrm{d}z. \end{aligned}$$

Since $1 = \int F_\lambda^{(1)}(u,v)\mathrm{d}v = \int \lambda\exp(-\lambda|u-v|)G_\lambda(u,v)\mathrm{d}v$, applying Fubini's theorem we obtain

$$A_j = \Phi_\lambda^1(x_j) \qquad \text{and} \qquad B = \sum_{j=1}^d \Phi_\lambda^2(x_j) \tag{2.81}$$

where we define for $u \in [0,1]$ and $k \in \mathbf{N}$

$$\Phi_\lambda^k(u) = \int_0^1 \lambda\exp(-\lambda|u-v|)G_\lambda(u,v)\frac{(v-u)^k}{k!}\mathrm{d}v\,. \tag{2.82}$$

Observe that

$$\Phi_\lambda^k(u) = \lambda^{-k} \int_{-\lambda u}^{\lambda(1-u)} \frac{v^k}{k!} \exp(-|v|) H(|v|, \lambda u + v \wedge 0, \lambda(1-u) - v \vee 0) \mathrm{d}v \,.$$

We will control $\Phi_\lambda^k(u)$ for $k = 1, 2$. First, write

$$\lambda \Phi_\lambda^1(u) = -\int_0^{\lambda u} v e^{-v} H(v, \lambda u - v, \lambda(1-u)) \mathrm{d}v + \int_0^{\lambda(1-u)} v e^{-v} H(v, \lambda u, \lambda(1-u) - v) \mathrm{d}v.$$

Now, let $\beta := \lambda \frac{u \wedge (1-u)}{2}$. We have

$$\lambda \Phi_\lambda^1(u) - \int_0^\beta v e^{-v} \left[H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \right] \mathrm{d}v =$$
$$- \underbrace{\int_\beta^{\lambda u} v e^{-v} H(v, \lambda u - v, \lambda(1-u)) \mathrm{d}v}_{:= I_1 \geqslant 0} + \underbrace{\int_\beta^{\lambda(1-u)} v e^{-v} H(v, \lambda u, \lambda(1-u) - v) \mathrm{d}v}_{:= I_2 \geqslant 0}$$

so that the left-hand side of the above equation is between $-I_1 \leqslant 0$ and $I_2 \geqslant 0$, and thus its absolute value is bounded by $|I_1| \vee |I_2|$. Now, note that, since $H(v, \cdot, \cdot) \leqslant v^{-1}$, we have

$$|I_2| \leqslant \int_\beta^\infty v e^{-v} v^{-1} \mathrm{d}v = e^{-\beta}$$

and similarly $|I_1| \leqslant e^{-\beta}$, so that

$$\left| \lambda \Phi_\lambda^1(u) - \underbrace{\int_0^\beta v e^{-v} \left[H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \right] \mathrm{d}v}_{:= I_3} \right| \leqslant e^{-\beta} \,. \tag{2.83}$$

It now remains to bound $|I_3|$. For that purpose, note that since H is decreasing in its second and third argument, we have

$$\begin{aligned} H(v) - H(v, \lambda u - v, \lambda(1-u)) &\leqslant H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \\ &\leqslant H(v, \lambda u, \lambda(1-u) - v) - H(v) \end{aligned}$$

which implies

$$\begin{aligned} &|H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))| \\ &\leqslant \max(|H(v, \lambda u, \lambda(1-u) - v) - H(v)|, |H(v) - H(v, \lambda u - v, \lambda(1-u))|). \end{aligned}$$

Besides, since $(a + E_1 \wedge b_1 + E_2 \wedge b_2)^{-1} \leqslant (a + E_1 + E_2)^{-1} + a^{-1}(\mathbf{1}\{E_1 \geqslant b_1\} + \mathbf{1}\{E_2 \geqslant b_2\})$,

$$H(a, b_1, b_2) - H(a) \leqslant a^{-1}(e^{-b_1} + e^{-b_2}), \tag{2.84}$$

for all a, b_1, b_2. Since $\lambda u - v \geqslant \beta$ and $\lambda(1-u) - v \geqslant \beta$ for $v \in [0, \beta]$, we have

$$|H(v) - H(v, \lambda u - v, \lambda(1-u))|, |H(v) - H(v, \lambda u, \lambda(1-u) - v)| \leqslant 2 v^{-1} e^{-\beta}$$

so that for $v \in [0, \beta]$

$$|H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))| \leqslant 2v^{-1}e^{-\beta}$$

and hence

$$\begin{aligned}
|I_3| &\leqslant \int_0^\beta v e^{-v} \left|H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))\right| \mathrm{d}v \\
&\leqslant \int_0^\beta v e^{-v} 2v^{-1} e^{-\beta} \mathrm{d}v \\
&\leqslant 2e^{-\beta} \int_0^\infty e^{-v} \mathrm{d}v \\
&= 2e^{-\beta}
\end{aligned} \tag{2.85}$$

Combining Equations (2.83) and (2.85) yields:

$$|\Phi_\lambda^1(u)| \leqslant \frac{3}{\lambda} e^{-\lambda[u \wedge (1-u)]/2} \tag{2.86}$$

that is,

$$\left\| \int_{[0,1]^d} (z - x) F_\lambda(x, z) \mathrm{d}z \right\|^2 = \sum_{j=1}^d \left(\Phi_\lambda^1(x_j)\right)^2 \leqslant \frac{9}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} .$$

Furthermore,

$$\begin{aligned}
0 \leqslant \Phi_\lambda^2(u) &= \lambda^{-2} \int_{-\lambda u}^{\lambda(1-u)} \frac{v^2}{2} e^{-|v|} H(|v|, \lambda u + v \wedge 0, \lambda(1-u) - v \vee 0) \mathrm{d}v \\
&\leqslant \lambda^{-2} \int_0^\infty v^2 e^{-v} v^{-1} \mathrm{d}v \\
&= \lambda^{-2} ,
\end{aligned}$$

so that

$$0 \leqslant \Phi_\lambda^2(u) \leqslant \frac{1}{\lambda^2},$$

which proves the second inequality by summing over $j = 1, \ldots, d$. This concludes the proof of Lemma 2.1. □

Chapter 3

Aggregated Mondrian forests for online learning

Abstract. Random Forests (RF) is one of the algorithms of choice in many supervised learning applications, be it classification or regression. The appeal of such methods comes from a combination of several characteristics: a remarkable accuracy in a variety of tasks, a small number of parameters to tune, robustness with respect to features scaling, a reasonable computational cost for training and prediction, and their suitability in high-dimensional settings. The most commonly used RF variants however are "offline" algorithms, which require the availability of the whole dataset at once. In this chapter, we introduce AMF, an online random forest algorithm based on Mondrian Forests. Using a variant of the Context Tree Weighting algorithm, we show that it is possible to efficiently perform an exact aggregation over all prunings of the trees; in particular, this enables to obtain a truly online parameter-free algorithm which is competitive with the optimal pruning of the Mondrian tree, and thus adaptive to the unknown regularity of the regression function. Numerical experiments show that AMF is competitive with respect to several strong baselines on a large number of datasets for multi-class classification.

Contents

3.1 Introduction

Introduced by Breiman (2001a), Random Forests (RF) is one of the algorithms of choice in many supervised learning applications. The appeal of these methods comes from their remarkable accuracy in a variety of tasks, the small number (or even the absence) of parameters to

tune, their reasonable computational cost at training and prediction time, and their suitability in high-dimensional settings.

Most commonly used RF algorithms, such as the original random forest procedure (Breiman, 2001a), extra-trees (Geurts et al., 2006), or conditional inference forest (Hothorn et al., 2010) are batch algorithms, that require the whole dataset to be available at once. Several online random forests variants have been proposed to overcome this issue and handle data that come sequentially. Utgoff (1989) was the first to extend Quinlan's ID3 batch decision tree algorithm (see Quinlan, 1986) to an online setting. Later on, Domingos and Hulten (2000) introduce Hoeffding Trees that can be easily updated: since observations are available sequentially, a cell is split when (i) enough observations have fallen into this cell, (ii) the best split in the cell is statistically relevant (a generic Hoeffding inequality being used to assess the quality of the best split).

Since random forests are known to exhibit better empirical performances than individual decision trees, online random forests have been proposed (see, e.g., Saffari et al., 2009; Denil et al., 2013). These procedures aggregate several trees by computing the mean of the tree predictions (regression setting) or the majority vote among trees (classification setting). The tree construction differs from one forest to another but share similarities with Hoeffding trees: a cell is to be split if (i) and (ii) (defined above) are verified.

One forest of particular interest for this work is the Mondrian Forest (Lakshminarayanan et al., 2014) based on the Mondrian process (Roy and Teh, 2009). Their construction differs from the construction described above since each new observation modifies the tree structure: instead of waiting for enough observations to fall into a cell in order to split it, the properties of the Mondrian process allow to update the Mondrian tree partition each time a sample is collected. Once a Mondrian tree is built, its prediction function uses a hierarchical prior on all subtrees and the average of predictions on all subtrees is computed with respect to this hierarchical prior using an approximation algorithm.

The algorithm we propose, called AMF, and illustrated in Figure 3.1 below on a toy binary classification dataset, differs from Mondrian Forest by the smoothing procedure used on each tree. While the hierarchical Bayesian smoothing proposed in Lakshminarayanan et al. (2014) requires approximations, the prior we choose allows for exact computation of the posterior distribution. The choice of this posterior is inspired by Context Tree Weighting (see, e.g., Willems et al., 1995; Willems, 1998; Helmbold and Schapire, 1997; Catoni, 2004), commonly used in lossless compression to aggregate all subtrees of a prespecified tree, which is both computationally efficient and theoretically sound. Since we are able to compute exactly

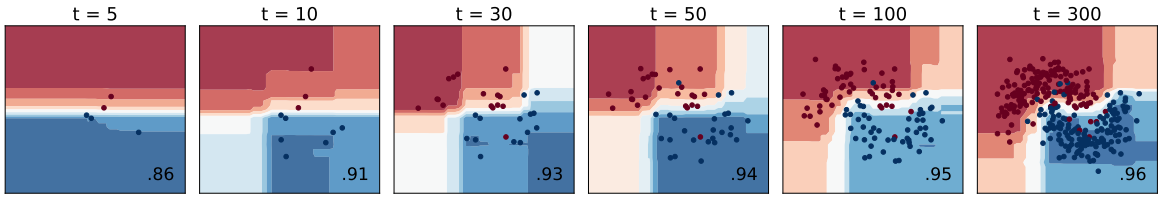


Figure 3.1: Evolution of the decision function of AMF along the online learning steps. We observe the online property of this algorithm, which produces a smooth decision function at each iteration, and leads to a correct AUC on a test set even in the early stages.

the posterior distribution, our approach is drastically different from Bayesian trees (see, for instance, Chipman et al., 1998; Denison et al., 1998; Taddy et al., 2011), and from BART

(Chipman et al., 2010) which implement MCMC methods to approximate posterior distributions on trees. The Context Tree Weighting algorithm has been applied to regression trees by Blanchard (1999) in the case of a fixed-design tree, in which splits are prespecified. This requires to split the dataset into two parts (using the first part to select the best splits and the second to compute the posterior distribution) and to have access to the whole dataset, since the tree structure needs to be fixed in advance.

As noted by Rockova and van der Pas (2017), the theoretical study of Bayesian methods on trees (Chipman et al., 1998; Denison et al., 1998) or sum of trees (Chipman et al., 2010) is less developed. Rockova and van der Pas (2017) analyzes some variant of Bayesian regression trees and sum of trees; they obtain near minimax optimal posterior concentration rates. Likewise, Linero and Yang (2018) analyze Bayesian sums of soft decision trees models, and establish minimax rates of posterior concentration for the resulting SBART procedure. While these frameworks differ from ours (herein results are posterior concentration rates as opposed to regret and excess risk bounds, and the design is fixed), their approach differs from ours primarily in the chosen trade-off between computational complexity and adaptivity of the method: these procedures involve approximate posterior sampling over large functional spaces through MCMC methods, and it is unclear whether the considered priors allow for reasonably efficient posterior computations. In particular, the prior used in Rockova and van der Pas (2017) is taken over all subsets of variables, which is exponentially large in the number of features.

The literature focusing on the original RF algorithm or its related variants is more extensive, even if the data-dependent nature of the algorithm and its numerous components (sampling procedure, split selection, aggregation) make the theoretical analysis difficult. The consistency of stylized RF algorithms was first established by Biau et al. (2008), and later obtained for more sophisticated variants in Denil et al. (2013); Scornet et al. (2015). Note that consistency results do not provide rates of convergence, and hence only offer limited guidance on how to properly tune the parameters of the algorithm. Starting with Biau (2012); Genuer (2012), some recent work has thus sought to quantify the speed of convergence of some stylized variants of RF. Minimax optimal nonparametric rates were first obtained by Arlot and Genuer (2014) in dimension 1 for the Purely Uniformly Random Forests (PURF) algorithm, in conjunction with suboptimal rates in arbitrary dimension (the number of features exceeds 1).

Several recent works (Wager and Walther, 2015; Duroux and Scornet, 2018) also established rates of convergence for variants of RF that essentially amount to some form of Median Forests, where each node contains at least a fixed fraction of observations of its parent. While valid in arbitrary dimension, the established rates are suboptimal. More recently, adaptive minimax optimal rates were obtained by Mourtada et al. (2018) (Chapter 2) in arbitrary dimension for the batch Mondrian Forests algorithm. Our proposed online algorithm, AMF, also achieves minimax rates in an adaptive fashion, namely without knowing the smoothness of the regression function.

In this chapter, we introduce AMF, a random forest algorithm which is fully online and computationally exact: unlike Bayesian trees and sum-of-trees procedures relying on approximate posterior sampling, we are able to compute exactly the prediction function of AMF in a very efficient way. Section 3.2 introduces the setting considered and general notations, and provides a precise construction of the AMF algorithm. A theoretical analysis of AMF is given in Section 3.3, where we establish regret bounds for AMF together with a minimax adaptive upper bound. Section 3.4 introduces a modification of AMF which is used in all the numerical experiments of the chapter, together with a guarantee and a discussion on its computational

complexity. Numerical experiments are provided in Section 3.5, on a large number of datasets, that include a comparison of AMF with several strong baselines. Our conclusions are provided in Section 3.6. The proofs of all the results are gathered in Section 3.7.

3.2 Forests of aggregated Mondrian trees

We define in Section 3.2.1 the setting and notations that will be used throughout the chapter, together with the definition of the Mondrian process, introduced by Roy and Teh (2009), which is a key element of our algorithm. In Section 3.2.2, we explicitly describe the prediction function that we want to compute, and prove in Proposition 3.1 that the AMF algorithm described in Section 3.2.3 computes it exactly.

3.2.1 The setting, trees, forests and the Mondrian process

We are interested in an *online supervised learning* problem in which we assume that the dataset is not fixed in advance. In this scenario, we are given an i.i.d. sequence (x_1, y_1), $(x_2, y_2), \dots$ of $[0,1]^d \times \mathcal{Y}$-valued random variables that come sequentially, such that each (x_t, y_t) has the same distribution as a generic pair (x, y).

Our aim is to design an *online* algorithm that can be updated "on the fly" given new sample points, that is, at each time step $t \geqslant 1$, a *randomized prediction function*

$$\widehat{f}_t(\cdot, \mathbf{\Pi}_t, \mathscr{D}_t) : [0,1]^d \to \widehat{\mathcal{Y}},$$

where $\mathscr{D}_t = \{(x_1, y_1), \dots, (x_t, y_t)\}$ is the dataset available at time t, where $\mathbf{\Pi}_t$ is a random variable that accounts for the randomization procedure and $\widehat{\mathcal{Y}}$ is a prediction space, see Examples 3.1 and 3.2 below for example. In the rest of the chapter, we omit the explicit dependence in $\mathscr{D}_t$.

We consider prediction rules $(\widehat{f}_t)_{t \geqslant 1}$ that are *random forests*, defined as the averaging of a set of $M \geqslant 1$ randomized decision trees. We let $\widehat{f}_t(x, \Pi_t^{(1)}), \dots, \widehat{f}_t(x, \Pi_t^{(M)})$ be randomized tree predictors at a point $x \in [0,1]^d$ at time t, associated to the same randomized mechanism, where the $(\Pi_t^{(m)})_{t \geqslant 1}$ for $m = 1, \dots, M$ are i.i.d. and correspond to a random tree partition, which is described below. Setting $\mathbf{\Pi}_t^{(M)} = (\Pi_t^{(1)}, \dots, \Pi_t^{(M)})$, the *random forest estimate* $\widehat{f}_t^{(M)}(x, \mathbf{\Pi}_t^{(M)})$ is then defined by

$$\widehat{f}_t^{(M)}(x, \mathbf{\Pi}_t^{(M)}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_t(x, \Pi_t^{(m)}), \tag{3.1}$$

namely taking the average over all tree predictions $\widehat{f}_t(x, \Pi_t^{(m)})$. The online training of each tree can be done in parallel, since they are fully independent of each other and each of them follow the exact same randomized construction. Therefore, we describe only the construction of single tree (and its associated random partition and prediction function) and omit from now on the dependence on $m = 1, \dots, M$.

The random tree partitions are given by $\Pi_t = (\mathcal{T}_t, \Sigma_t)$, where $\mathcal{T}_t$ is a binary tree and Σ_t contains information about each node in $\mathcal{T}_t$, such as splits, as explained below. Let us now introduce notations and definitions of these objects, for simplicity we first assume that t is fixed, and remove the dependence on t for a little while.

Definition 3.1 (Tree partition)**.** Let $C \subseteq [0,1]^d$ be a hyper-rectangular box of the form $\prod_{j=1}^d [a_j, b_j]$, $-\infty \leqslant a_j \leqslant b_j \leqslant +\infty$ (the interval being open at an infinite extremity). A *tree partition* (or *kd tree*, *guillotine partition*) of C is a pair $(\mathcal{T}, \Sigma)$, where

- $\mathcal{T}$ is a finite ordered binary tree, which is represented as a finite subset of the set $\{0,1\}^* = \bigcup_{n \geqslant 0} \{0,1\}^n$ of all finite words on the alphabet $\{0,1\}$. The set $\{0,1\}^*$ is endowed with a tree structure (and called the complete binary tree): the empty word ϵ is the root, and for any $\mathbf{v} \in \{0,1\}^*$, the left (resp. right) child of $\mathbf{v}$ is $\mathbf{v}0$ (resp. $\mathbf{v}1$), obtained by adding a 0 (resp. 1) at the end of $\mathbf{v}$. We denote by $\mathcal{N}^\circ(\mathcal{T}) = \{\mathbf{v} \in \mathcal{T} : \mathbf{v}0, \mathbf{v}1 \in \mathcal{T}\}$ the set of its *interior nodes* and by $\mathcal{L}(\mathcal{T}) = \{\mathbf{v} \in \mathcal{T} : \mathbf{v}0, \mathbf{v}1 \notin \mathcal{T}\}$ the set of its leaves, which are disjoint by definition.
- $\Sigma = (\sigma_\mathbf{v})_{\mathbf{v} \in \mathcal{N}^\circ(\mathcal{T})}$ is a family of *splits* at the interior nodes of $\mathcal{T}$, where each split $\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$ is characterized by its split dimension $j_\mathbf{v} \in \{1, \dots, d\}$ and its threshold $s_\mathbf{v} \in [0,1]$. In Section 3.2.3, we will actually store in $\sigma_\mathbf{v} \in \Sigma$ more information about nodes $\mathbf{v} \in \mathcal{T}$.

One can associate to $(\mathcal{T}, \Sigma)$ a partition $(C_\mathbf{v})_{\mathbf{v} \in \mathcal{L}(\mathcal{T})}$ of $[0,1]^d$ as follows. For each node $\mathbf{v} \in \mathcal{T}$, its *cell* $C_\mathbf{v}$ is a hyper-rectangular region $C_\mathbf{v} \subseteq [0,1]^d$ defined recursively: the cell associated to the root ϵ of $\mathcal{T}$ is $[0,1]^d$, and, for each $\mathbf{v} \in \mathcal{N}^\circ(\mathcal{T})$, we define

$$C_{\mathbf{v}0} := \{x \in C_\mathbf{v} : x_{j_\mathbf{v}} \leqslant s_{j_\mathbf{v}}\} \quad \text{and} \quad C_{\mathbf{v}1} := C_\mathbf{v} \setminus C_{\mathbf{v}0}.$$

Then, the leaf cells $(C_\mathbf{v})_{\mathbf{v} \in \mathcal{L}(\mathcal{T})}$ form a partition of $[0,1]^d$ by construction.

Mondrian partitions are a specific family of random tree partitions whose construction is described below. An infinite Mondrian partition Π of $[0,1]^d$ can be sampled from the infinite Mondrian process, denoted MP from now on, using the procedure $\texttt{SampleMondrian}([0,1]^d, \tau = 0)$ described below. If $C = \prod_{j=1}^d C^j$ with intervals $C^j = [a_j, b_j]$, we denote $|C^j| = b_j - a_j$ and $|C| = \sum_{j=1}^d |C^j|$. We denote by $\mathsf{Exp}(\lambda)$ the exponential distribution with intensity $\lambda > 0$ and by $\mathcal{U}([a,b])$ the uniform distribution on a finite interval $[a,b]$.

Algorithm 2 $\texttt{SampleMondrian}(C_\mathbf{v}, \tau_\mathbf{v})$: sample a Mondrian starting from a cell $C_\mathbf{v}$ and time $\tau_\mathbf{v}$

Inputs: The cell $C_\mathbf{v} = \prod_{1 \leqslant j \leqslant d} C_\mathbf{v}^j$ and creation time $\tau_\mathbf{v}$ of a node $\mathbf{v}$
Sample a random variable $E \sim \mathsf{Exp}(|C_\mathbf{v}|)$ and put $\tau_{\mathbf{v}0} = \tau_{\mathbf{v}1} = \tau_\mathbf{v} + E$
Sample a split coordinate $j_\mathbf{v} \in \{1, \dots, d\}$ with $\mathbb{P}(j_\mathbf{v} = j) = |C_\mathbf{v}^j| / |C_\mathbf{v}|$
Sample a split threshold $s_\mathbf{v}$ conditionally on $j_\mathbf{v}$ as $s_\mathbf{v} | j_\mathbf{v} \sim \mathcal{U}(C_\mathbf{v}^{j_\mathbf{v}})$
Following Definition 3.1, the split $(j_\mathbf{v}, s_\mathbf{v})$ defines children cells $C_{\mathbf{v}0}$ and $C_{\mathbf{v}1}$
return $\texttt{SampleMondrian}(C_{\mathbf{v}0}, \tau_{\mathbf{v}0}) \cup \texttt{SampleMondrian}(C_{\mathbf{v}1}, \tau_{\mathbf{v}1})$

The call to $\texttt{SampleMondrian}([0,1]^d, \tau = 0)$ corresponds to a call starting at the root node $\mathbf{v} = \epsilon$, since $C_\epsilon = [0,1]^d$ and the birth time of ϵ is $\tau_\epsilon = 0$. This random partition is built by iteratively splitting cells at some random time, which depends on the linear dimension $C_\mathbf{v}$ of the input cell $C_\mathbf{v}$. The split coordinate $j_\mathbf{v}$ is chosen at random, with a probability of sampling j which is proportional to the side length $|C_\mathbf{v}^j| / |C_\mathbf{v}|$ of the cell, and the split threshold is sampled uniformly in $C_\mathbf{v}^j$. The number of recursions in this procedure is infinite, the Mondrian process MP is a distribution on infinite tree partitions of $[0,1]^d$, see Roy and

Teh (2009) and Roy (2011) for a rigorous construction. The random partition described in Section 3.2.3 below, is, however, not infinite, and depends on the features vectors x_t seen until time t. The implementation of AMF used in all our experiments, described in Section 3.4 below, also considers finite partitions, through the concept of *restricted Mondrian partitions*, introduced in Lakshminarayanan et al. (2014). At this point, the *birth times* $\tau_{\mathbf{v}}$ computed in Algorithm 2 are not used. They will allow to define *time prunings* of a Mondrian partition in Section 3.3.1 below, a notion which is necessary to prove that AMF has adaptation capabilities to the optimal time pruning. Birth times $\tau_{\mathbf{v}}$ are also necessary for the definition of restricted Mondrian partitions in Section 3.4, which is an important ingredient in the actual implementation of AMF.

3.2.2 Aggregation with exponential weights and prediction functions

The prediction function of AMF is an aggregation of the predictions given by all finite subtrees of the infinite Mondrian partition MP. This aggregation step is performed in a purely online fashion, using an aggregation algorithm based on exponential weights, with a branching process prior over the subtrees, see Definition 3.3 below. This weighting scheme gives more importance to subtrees with a good predictive performance.

Let us assume that the realization of an infinite Mondrian partition $\Pi = (\mathcal{T}^{\Pi}, \Sigma^{\Pi}) \sim \mathsf{MP}$ is available at some fixed step t. We will argue in Section 3.2.3 that it suffices to store a finite partition Π_t, and show how to update it. The definition of the prediction function used in AMF require the notion of *node* and *subtree prediction*, defined below.

Definition 3.2. Given $\Pi = (\mathcal{T}^{\Pi}, \Sigma^{\Pi}) \sim \mathsf{MP}$, we define

$$\widehat{y}_{\mathbf{v},t} = h((y_s)_{1 \leqslant s \leqslant t-1 \,:\, x_s \in C_{\mathbf{v}}}) \quad \text{and} \quad L_{\mathbf{v},t} = \sum_{1 \leqslant s \leqslant t \,:\, x_s \in C_{\mathbf{v}}} \ell(\widehat{y}_{\mathbf{v},s}, y_s)$$

for each node $\mathbf{v} \in \mathcal{T}^{\Pi}$ (which defines a cell $C_{\mathbf{v}} \subseteq [0,1]^d$ following Definition 3.1) and each $t \geqslant 1$, where $h : \bigcup_{t \geqslant 0} \mathcal{Y}^t \to \widehat{\mathcal{Y}}$ is a prediction algorithm used in each cell, with $\widehat{\mathcal{Y}}$ its prediction space and $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ a generic loss function. The prediction at time t of a finite subtree $\mathcal{T} \subset \mathcal{T}^{\Pi}$ associated to some features vector $x \in [0,1]^d$ is defined by

$$\widehat{y}_{\mathcal{T},t}(x) = \widehat{y}_{\mathbf{v}_{\mathcal{T}}(x),t},$$

where $\mathbf{v}_{\mathcal{T}}(x)$ is the leaf of $\mathcal{T}$ that contains x. We define also the cumulative loss of $\mathcal{T}$ at time t as

$$L_t(\mathcal{T}) = \sum_{s=1}^{t} \ell(\widehat{y}_{\mathcal{T},s}(x_s), y_s)\,.$$

Before defining the prediction function of AMF, let us first make explicit the prediction function h and the loss considered in two specific cases of interest: regression and classification.

Example 3.1 (Regression)**.** In regression, we use empirical mean forecasters

$$\widehat{y}_{\mathbf{v},t+1} = \frac{1}{n_{\mathbf{v},t}} \sum_{1 \leqslant s \leqslant t \,:\, x_s \in C_{\mathbf{v}}} y_s,$$

where $n_{\mathbf{v},t} = |\{1 \leqslant s \leqslant t \,:\, x_s \in C_{\mathbf{v}}\}|$, and we simply put $\widehat{y}_{\mathbf{v},t} = 0$ if $\mathbf{v}$ is empty (namely, $C_{\mathbf{v}}$ contains no data point). The loss is the *quadratic loss* $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ for any $y \in \mathcal{Y}$ and $\widehat{y} \in \widehat{\mathcal{Y}}$ where $\widehat{\mathcal{Y}} = \mathcal{Y} = \mathbf{R}$.

Example 3.2 (Classification)**.** For multi-class classification, we have labels $y_t \in \mathcal{Y}$ where $\mathcal{Y}$ is a finite set of label modalities (such as $\mathcal{Y} = \{1, \dots, K\}$) and predictions are in $\widehat{\mathcal{Y}} = \mathcal{P}(\mathcal{Y})$, the set of probability distributions on $\mathcal{Y}$. We use the *Krichevsky-Trofimov* (KT) forecaster (see Tjalkens et al., 1993) in each node $\mathbf{v}$, which predicts

$$\widehat{y}_{\mathbf{v},t+1}(y) = \frac{n_{\mathbf{v},t}(y) + 1/2}{t + |\mathcal{Y}|/2}, \tag{3.2}$$

for any $y \in \mathcal{Y}$, where $n_{\mathbf{v},t}(y) = |\{1 \leqslant s \leqslant t : x_s \in C_{\mathbf{v}}, y_s = y\}|$. For an empty $\mathbf{v}$ we use the uniform distribution on $\mathcal{Y}$. We consider the *logarithmic loss* (also called *cross-entropy* or *self-information* loss) $\ell(\widehat{y}, y) = -\log \widehat{y}(y)$, where $\widehat{y}(y) = \widehat{y}(\{y\}) \in [0, 1]$.

Remark 3.1. The Krichevsky-Trofimov forecaster coincides with the exponential weights algorithm under the logarithmic loss (with $\eta = 1$) on $\mathcal{P}(\mathcal{Y})$ with a prior equal to the Dirichlet distribution $\mathsf{Dir}(\frac{1}{2}, \dots, \frac{1}{2})$, namely the *Jeffreys prior* on the multinomial model $(\mathcal{Y}, \mathcal{P}(\mathcal{Y}))$.

Definition 3.3. Let $t \geqslant 1$ and $x \in [0, 1]^d$. The prediction function $\widehat{f}_t$ of AMF at step t is given by

$$\widehat{f}_t(x) = \frac{\sum_{\mathcal{T}} \pi(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})} \widehat{y}_{\mathcal{T},t}(x)}{\sum_{\mathcal{T}} \pi(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})}},$$

where the sum is over all subtrees $\mathcal{T}$ of $\mathcal{T}^{\Pi}$ and where the *prior* π on subtrees is the probability distribution defined by

$$\pi(\mathcal{T}) = 2^{-|\mathcal{T}|}, \tag{3.3}$$

where $|\mathcal{T}|$ is the number of nodes in $\mathcal{T}$ and $\eta > 0$ is a parameter called *learning rate*.

Note that π is the distribution of the branching process with branching probability $1/2$ at each node of $\mathcal{T}^{\Pi}$, with exactly two children when it branches; this branching process gives finite subtrees almost surely. The learning rate η can be optimally tuned following theoretical guarantees from Section 3.3, see in particular Corollaries 3.1 and 3.2. This aggregation procedure is a *non-greedy way to prune trees*: the weights do not depend only on the quality of one single split but rather on the performance of each subsequent split.

Let us stress that computing $\widehat{f}_t$ from Definition 3.3 seems computationally infeasible in practice, since it involves a sum over all subtrees of $\mathcal{T}^{\Pi}$. Besides, it requires to keep in memory one weight $e^{-\eta L_{t-1}(\mathcal{T})}$ for all subtrees $\mathcal{T}$, which seems prohibitive as well. Indeed, the number of subtrees of the minimal tree that separates n points is exponential in the number of nodes, and hence *exponential in* n. However, the proper choice of the prior in Equation (3.3) allows us to prove that $\widehat{f}_t$ can actually be computed very efficiently, at almost no memory cost, as stated in Proposition 3.1 below, where we prove that the AMF algorithm described in Section 3.2.3 below allows to compute $\widehat{f}_t$ exactly and efficiently.

Proposition 3.1. *Let $t \geqslant 1$ and $x \in [0, 1]^d$. The value $\widehat{f}_t(x)$ from Definition 3.3 can be computed exactly via the* `AmfPredict` *procedure (see Algorithms 3 and 4 from Section 3.2.3 below).*

The proof of Proposition 3.1 is given in Section 3.7. It proves that aggregating predictions of all subtrees weighted by the prior π can be done exactly via Algorithm 4. This prior choice enables to bypass the need to maintain one weight per subtree, and leads to a "collapsed" implementation that only requires to maintain one weight per node (which is exponentially

smaller). Note that this algorithm is *exact*, in the sense that it does not require any approximation scheme. Moreover, this online algorithm corresponds to its batch counterpart, in the sense that there is no loss of information coming from the online (or streaming) setting versus the batch setting (where the whole dataset is available at once).

The proof of Proposition 3.1 relies on some standard identities that enable to efficiently compute sums of products over tree structures in a recursive fashion (from Helmbold and Schapire, 1997), recalled in Lemma 3.3 from Section 3.7. Such identities are at the core of the *Context Tree Weighting* algorithm (CTW), which our online algorithm implements (albeit over an evolving tree structure, as explained in Section 3.2.3 below), and which consists of an efficient way to perform Bayesian mixtures of contextual tree models under a branching process prior. The CTW algorithm, based on a sum-product factorization, is a state-of-the art algorithm used in lossless coding and compression. We use a variant of the *Tree Expert* algorithm (Helmbold and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), which is closely linked to CTW (Willems et al., 1995; Willems, 1998; Catoni, 2004).

3.2.3 AMF: a forest of aggregated Mondrian trees

In an online setting, the number of sample points increases over time, allowing one to capture more details on the distribution of y conditionally on x. This means that the complexity of our models (in this context, the complexity of the decision trees) should increase over time. We will therefore need to consider not just an individual, fixed tree partition Π, but a sequence $(\Pi_t)_{t\geqslant 1}$, indexed by "time" t corresponding to the number of samples available. Furthermore, AMF uses the aggregated prediction function given in Definition 3.3 (independently within each tree $\Pi_t^{(1)}, \ldots, \Pi_t^{(M)}$ from the forest, see Equation (3.1)). When a new sample point (x_t, y_t) becomes available, the algorithm does two things, in the following order:

- *Partition update.* Using x_t, update the decision tree structure from $\Pi_t = (\mathcal{T}_t, \Sigma_t)$ to $\Pi_{t+1} = (\mathcal{T}_{t+1}, \Sigma_{t+1})$, *i.e.* sample new splits in order to ensure that each leaf in the tree *contains at most one point* among $\{x_1, \ldots, x_t\}$. This update uses the recursive properties of Mondrian partitions;

- *Prediction function update.* Using x_t and y_t, update the prediction functions $\widehat{y}_{\mathbf{v},t}$ and weights $w_{\mathbf{v},t}$ and $\overline{w}_{\mathbf{v},t}$ that are necessary for the computation of $\widehat{f}_t$ from Definition 3.3. These updates are *local* and are performed only along the path of nodes leading to the leaf containing x_t. This update is efficient and enables the computation of $\widehat{f}_t$ from Definition 3.3, which aggregates the decision functions of all the prunings of the tree, thanks to a variant of CTW.

Both updates can be implemented on the fly in a *purely sequential manner*. Training over a sequence $(x_1, y_1), \ldots, (x_t, y_t)$ means using each sample once for training, and both updates are *exact* and do not rely on an approximate sampling scheme. Both steps are precisely described in Algorithm 3 below and illustrated in Figure 3.2. Also, in order to ease the reading of this technical part of the chapter, we gather in Table 3.1 notations that are used in this Section.

Partition update. Before seeing the point (x_t, y_t), the algorithm maintains a partition $\Pi_t = (\mathcal{T}_t, \Sigma_t)$, which corresponds to the minimal subtree of the infinite Mondrian partition $\Pi \sim \mathsf{MP}$ that separates all distinct sample points in $\{x_1, \ldots, x_{t-1}\}$. This corresponds to the tree obtained from the infinite tree Π by removing all splits of "empty" cells (that do not

Notation or formula	Description
$\mathbf{v} \in \{0,1\}^*$	A node
$\mathcal{T} \subset \{0,1\}^*$	A tree
$\mathbf{v}0$ (resp. $\mathbf{v}1$)	The left (resp. right) child of $\mathbf{v}$
$\mathcal{T}_\mathbf{v}$	A subtree rooted at $\mathbf{v}$
$\mathcal{L}(\mathcal{T})$	The set of leaves of $\mathcal{T}$
$\mathcal{N}^\circ(\mathcal{T})$	The set of the interior nodes of $\mathcal{T}$
$(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(\mathcal{T})}$	The cells of the partition defined by $\mathcal{T}$
$\widehat{y}_{\mathbf{v},t}$	Prediction of a node $\mathbf{v}$ at time t
$L_{\mathbf{v},t} = \sum_{1\leqslant s\leqslant t\,:\,X_s\in C_\mathbf{v}} \ell(\widehat{y}_{\mathbf{v},s}, y_s)$	Cumulative loss of the node $\mathbf{v}$ at time t
$w_{\mathbf{v},t} = \exp(-\eta L_{\mathbf{v},t-1})$	Weight stored in node $\mathbf{v}$ at time t
$\overline{w}_{\mathbf{v},t} = \sum_{\mathcal{T}_\mathbf{v}} 2^{-\lvert\mathcal{T}_\mathbf{v}\rvert} \prod_{\mathbf{v}'\in\mathcal{L}(\mathcal{T}_\mathbf{v})} w_{\mathbf{v}',t}$	Average weight stored in node $\mathbf{v}$ at time t

Table 3.1: Notations and definitions used in AMF

contain any point among $\{x_1, \ldots, x_{t-1}\}$). As x_t becomes available, this tree is updated as follows (this corresponds to Lines 2–11 in Algorithm 3 below):

- find the leaf in Π_t that contains x_t; it contains at most one point among $\{x_1, \ldots, x_{t-1}\}$;
- if the leaf contains no point $x_s \neq x_t$, then let $\Pi_{t+1} = \Pi_t$. Otherwise, let x_s be the unique point among $\{x_1, \ldots, x_{t-1}\}$ (distinct from x_t) in this cell. Splits of the cell containing $\{x_s, x_t\}$ are successively sampled (following the recursive definition of the Mondrian distribution), until a split separates x_s and x_t.

Prediction function update. The algorithm maintains weights $w_{\mathbf{v},t}$ and $\overline{w}_{\mathbf{v},t}$ and predictions $\widehat{y}_{\mathbf{v},t}$ in order to compute the aggregation over the tree structure (lines 12–18 in Algorithm 3). Namely, after round $t-1$ (after seeing sample (x_{t-1}, y_{t-1})), each node $\mathbf{v} \in \mathcal{T}_t$ has the following quantities in memory:

- the weight $w_{\mathbf{v},t} = \exp(-\eta L_{\mathbf{v},t-1})$, where $L_{\mathbf{v},t} := \sum_{1\leqslant s\leqslant t\,:\,x_s\in C_\mathbf{v}} \ell(\widehat{y}_{\mathbf{v},s}, y_s)$;
- the averaged weight $\overline{w}_{\mathbf{v},t} = \sum_{\mathcal{T}_\mathbf{v}} 2^{-|\mathcal{T}_\mathbf{v}|} \prod_{\mathbf{v}'\in\mathcal{L}(\mathcal{T}_\mathbf{v})} w_{\mathbf{v}',t}$, where the sum ranges over all subtrees $\mathcal{T}_\mathbf{v}$ rooted at $\mathbf{v}$;
- the forecast $\widehat{y}_{\mathbf{v},t}$ in node $\mathbf{v}$ at time t.

Now, given a new sample point (x_t, y_t), the update is performed as follows: we find the leaf $\mathbf{v}_t = \mathbf{v}_{\Pi_{t+1}}(x_t)$ containing x_t in Π_{t+1} (the partition has been updated with x_t already, since the partition update is performed *before* the prediction function update). Then, we update the values of $w_{\mathbf{v},t}, \overline{w}_{\mathbf{v},t}, \widehat{y}_{\mathbf{v},t}$ for each $\mathbf{v}$ along an *upwards* recursion from $\mathbf{v}_t$ to the root, while *the values of nodes outside of the path are kept unchanged*:

- $w_{\mathbf{v},t+1} = w_{\mathbf{v},t} \exp(-\eta \ell(\widehat{y}_{\mathbf{v},t}, y_t))$;
- if $\mathbf{v} = \mathbf{v}_t$ then $\overline{w}_{\mathbf{v},t+1} = w_{\mathbf{v},t+1}$, otherwise

$$\overline{w}_{\mathbf{v},t+1} = \frac{1}{2} w_{\mathbf{v},t+1} + \frac{1}{2} \overline{w}_{\mathbf{v}0,t+1} \overline{w}_{\mathbf{v}1,t+1};$$

- $\widehat{y}_{\mathbf{v},t+1} = h((y_s)_{1\leqslant s\leqslant t\,:\,x_s\in C_{\mathbf{v}}})$ using the prediction algorithm $h : \bigcup_{t\geqslant 0}\mathcal{Y}^t \to \widehat{\mathcal{Y}}$, see Definition 3.2. Note that the prediction algorithms given in Examples 3.1 and 3.2 can be updated online using y_t only and do not require to look back at the sequence $y_1, \ldots, y_{t-1}$.

The *partition update* and *prediction function update* correspond to the `AmfUpdate`(x, y) procedure described in Algorithm 3 below. Training AMF over a sequence $(x_1, y_1), \ldots, (x_t, y_t)$

Algorithm 3 `AmfUpdate`(x, y) : update AMF with a new sample $(x, y) \in [0, 1]^d \times \mathcal{Y}$

Input: a new sample $(x, y) \in [0, 1]^d \times \mathcal{Y}$
Let $\mathbf{v}(x)$ be the leaf such that $x \in C_{\mathbf{v}(x)}$ and put $\mathbf{v} = \mathbf{v}(x)$
while $C_{\mathbf{v}}$ contains some $x' \neq x$ **do**
  Use Lines 1–5 from Algorithm 2 to split $C_{\mathbf{v}}$ and obtain children cells $C_{\mathbf{v}0}$ and $C_{\mathbf{v}1}$
  if $\{x, x'\} \subset C_{\mathbf{v}a}$ for some $a \in \{0, 1\}$ **then**
    Put $\mathbf{v} = \mathbf{v}a$, $(w_{\mathbf{v}a}, \overline{w}_{\mathbf{v}a}, \widehat{y}_{\mathbf{v}a}) = (w_{\mathbf{v}}, \overline{w}_{\mathbf{v}}, \widehat{y}_{\mathbf{v}})$ and $(w_{\mathbf{v}(1-a)}, \overline{w}_{\mathbf{v}(1-a)}, \widehat{y}_{\mathbf{v}(1-a)}) = (1, 1, h(\varnothing))$ ($h(\varnothing)$ is the default initial prediction described in Examples 3.1 and 3.2)
  else
    Let $a \in \{0, 1\}$ be such that $x \in C_{\mathbf{v}a}$ and $x' \in C_{\mathbf{v}(1-a)}$. Put $\mathbf{v} = \mathbf{v}a$ and $(w_{\mathbf{v}a}, \overline{w}_{\mathbf{v}a}, \widehat{y}_{\mathbf{v}a}) = (1, 1, h(\varnothing))$ and $(w_{\mathbf{v}(1-a)}, \overline{w}_{\mathbf{v}(1-a)}, \widehat{y}_{\mathbf{v}(1-a)}) = (w_{\mathbf{v}}, \overline{w}_{\mathbf{v}}, \widehat{y}_{\mathbf{v}})$
  end if
end while
Put $x_{\mathbf{v}} = x$ (memorize the fact that $\mathbf{v}$ contains x)
Let `continueUp` $\leftarrow$ `true`
while `continueUp` **do**
  Set $w_{\mathbf{v}} = w_{\mathbf{v}} \exp(-\eta \ell(\widehat{y}_{\mathbf{v}}, y))$
  Set $\overline{w}_{\mathbf{v}} = w_{\mathbf{v}}$ if $\mathbf{v}$ is a leaf and $\overline{w}_{\mathbf{v}} = \frac{1}{2} w_{\mathbf{v}} + \frac{1}{2} \overline{w}_{\mathbf{v}0} \overline{w}_{\mathbf{v}1}$ otherwise
  Update $\widehat{y}_{\mathbf{v}}$ using y (following Definition 3.2)
  If $\mathbf{v} \neq \epsilon$ let $\mathbf{v} =$ `parent`$(\mathbf{v})$, otherwise let `continueUp` $=$ `false`
end while

means using successive calls to `AmfUpdate`$(x_1, y_1), \ldots,$ `AmfUpdate`(x_t, y_t). `AmfUpdate`(x, y) maintains in memory the current state of the Mondrian partition $\Pi = (\mathcal{T}, \Sigma)$. The tree $\mathcal{T}$ contains the parent and children relations between all nodes $\mathbf{v} \in \mathcal{T}$, while each $\sigma_{\mathbf{v}} \in \Sigma$ can contain

$$\sigma_{\mathbf{v}} = (j_{\mathbf{v}}, s_{\mathbf{v}}, \widehat{y}_{\mathbf{v}}, w_{\mathbf{v}}, \overline{w}_{\mathbf{v}}, x_{\mathbf{v}}), \tag{3.4}$$

namely the split coordinate $j_{\mathbf{v}} \in \{1, \ldots, d\}$ and split threshold $s_{\mathbf{v}} \in [0, 1]$ (only if $\mathbf{v} \in \mathcal{N}^\circ(\mathcal{T})$), the prediction function $\widehat{y}_{\mathbf{v}} \in \widehat{\mathcal{Y}}$, aggregation weights $w_{\mathbf{v}}, \overline{w}_{\mathbf{v}} \in (0, +\infty)$ and a vector $x_{\mathbf{v}} \in [0, 1]^d$ if $\mathbf{v} \in \mathcal{L}(\mathcal{T})$. An illustration of Algorithm 3 is given in Figure 3.2 below.

Remark 3.2. The complexity of `AmfUpdate`(x, y) is twice the depth of the tree at the moment it is called, since it requires to follow a downwards path to a leaf, and to go back upwards to the root. As explained in Proposition 3.2 from Section 3.4 below, the depth of the Mondrian tree used in AMF is $\Theta(\log n)$ in expectation at step n of training, which leads to a complexity $\Theta(\log n)$ both for Algorithms 3 and 4, where $\Theta(1)$ corresponds to the update complexity of a single node, while the original MF algorithm uses an update with complexity that is linear in the number of leaves in the tree (which is typically exponentially larger).

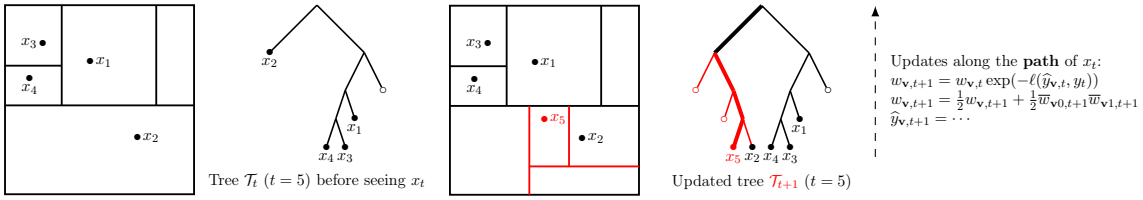


Figure 3.2: Illustration of the `AmfUpdate`(x_t, y_t) procedure from Algorithm 3: update of the partition, weights and node predictions as a new data point (x_t, y_t) for $t = 5$ becomes available. *Left*: tree partition Π_t before seeing (x_t, y_t). *Right*: update of the partition (in red) and new splits to separate x_5 from x_2. Empty circles ($\circ$) denote empty leaves, while leaves containing a point are indicated by a filled circle ($\bullet$). The path of x_t in the tree is indicated in bold. The updates of weights and predictions along the path are indicated, and are computed in an upwards recursion.

Prediction. At any point in time, one can ask AMF to perform prediction for an arbitrary features vector $x \in [0,1]^d$. Let us assume that AMF did already t training steps on the M trees it contains and let us recall that the prediction produced by AMF is the average of their predictions, see Equation (3.1), where the prediction $\widehat{f}_t(x, \Pi_t^{(m)})$ of each decision tree $m = 1, \ldots, M$ is computed in parallel following Definition 3.3.

The prediction of a decision tree is performed through a call to procedure `AmfPredict`(x) described in Algorithm 4 below. First, we perform a temporary partition update of Π using x, following Lines 2–10 of Algorithm 3, so that we find or create a new leaf node $\mathbf{v}(x)$ such that $x \in C_{\mathbf{v}(x)}$. Let us stress that this update of Π using x is discarded once the prediction for x is produced, so that the decision function of AMF does not change after producing predictions. The prediction is then computed recursively, along an upwards recursion going from $\mathbf{v}(x)$ to the root ϵ, in the following way:

- if $\mathbf{v} = \mathbf{v}(x)$ we set $\widetilde{y}_{\mathbf{v}} = \widehat{y}_{\mathbf{v}}$;
- if $\mathbf{v} \neq \mathbf{v}(x)$ (it is an interior node such that $x \in C_{\mathbf{v}}$), then assuming that $\mathbf{v}a$ ($a \in \{0,1\}$) is the child of $\mathbf{v}$ such that $x \in C_{\mathbf{v}a}$, we set

$$\widetilde{y}_{\mathbf{v}} = \frac{1}{2} \frac{w_{\mathbf{v}}}{\overline{w}_{\mathbf{v}}} \widehat{y}_{\mathbf{v}} + \frac{1}{2} \frac{\overline{w}_{\mathbf{v}0} \overline{w}_{\mathbf{v}1}}{\overline{w}_{\mathbf{v}}} \widetilde{y}_{\mathbf{v}a} \,.$$

The prediction $\widehat{f}_t(x)$ of the tree is given by $\widetilde{y}_\epsilon$, which is the last value obtained in this recursion. Let us recall that this computes the aggregation with exponential weights of all the decision functions produced by all the prunings of the current Mondrian tree, as described in Definition 3.3 and stated in Proposition 3.1 above. The prediction procedure is summarized in Algorithm 4 below.

The next Section 3.3 provides theoretical guarantees for AMF, but before that, let us provide the following numerical illustration on three toy datasets for binary classification. The aim of this illustration is to exhibit the effect of aggregation in AMF, compared to the same method with no aggregation, the original Mondrian Forest algorithm, batch Random Forest and Extra Trees (see Section 3.5 for a precise description of the implementations used). We observe that AMF with aggregation (AMF(agg)) produces a very smooth decision function in all cases, which generalizes better on this instance (AUCs displayed on the bottom right-hand side of each plot are computed on a 30% test dataset) than all other methods. All

Algorithm 4 `AmfPredict`(x) : predict the label of $x \in [0,1]^d$

Input: a features vector $x \in [0,1]^d$
Follow Lines 2–10 of Algorithm 3 to do a temporary update of the current partition Π using x and let $\mathbf{v}(x)$ be the leaf such that $x \in C_{\mathbf{v}(x)}$
Set $\widetilde{y}_{\mathbf{v}} = \widehat{y}_{\mathbf{v}(x)}$
while $\mathbf{v} \neq \epsilon$ **do**
    Let $(\mathbf{v}, \mathbf{v}a) = (\mathtt{parent}(\mathbf{v}), \mathbf{v})$ (for some $a \in \{0,1\}$)
    Let $\widetilde{y}_{\mathbf{v}} = \frac{1}{2}\frac{w_{\mathbf{v}}}{\overline{w}_{\mathbf{v}}}\widehat{y}_{\mathbf{v}} + \frac{1}{2}\frac{\overline{w}_{\mathbf{v}(1-a)}\overline{w}_{\mathbf{v}a}}{\overline{w}_{\mathbf{v}}}\widetilde{y}_{\mathbf{v}a}$
end while
Return $\widetilde{y}_{\epsilon}$

the other algorithms display rather non-smooth decision functions, which suggests that the underlying probability estimates are not well-calibrated.

3.3 Theoretical guarantees

In addition to being efficiently implementable in a streaming fashion, AMF is amenable to a thorough end-to-end theoretical analysis. This relies on two main ingredients: (i) a precise control of the geometric properties of the Mondrian partitions and (ii) a regret analysis of the aggregation procedure (exponentially weighted aggregation of all finite prunings of the infinite Mondrian) which in turn yields excess risk bounds and adaptive minimax rates. The guarantees provided below hold for a single tree in the Forest, but hold also for the average of several trees (used in by the forest) by convexity of the loss (see Examples 3.1 and 3.2).

3.3.1 Regret bounds

For now, the sequence $(x_1, y_1), \ldots, (x_n, y_n) \in [0,1]^d \times \mathcal{Y}$ is arbitrary, and is in particular not required to be i.i.d. Let us recall that at step t, we have a realization $\Pi_t = (\mathcal{T}_t, \Sigma_t)$ of a finite Mondrian tree, which is the minimal subtree of the infinite Mondrian partition $\Pi = (\mathcal{T}^{\Pi}, \Sigma^{\Pi})$ that separates all distinct sample points in $\{x_1, \ldots, x_t\}$. Let us recall also that $\widehat{y}_{\mathcal{T},t} : [0,1]^d \to \widehat{\mathcal{Y}}$ are the tree forecasters from Definition 3.2, where $\mathcal{T}$ is some subtree of $\mathcal{T}^{\Pi}$. We need the following

Definition 3.4. Let $\eta > 0$. A loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ is said to be *η-exp-concave* if the function $\exp(-\eta\, \ell(\cdot, y)) : \widehat{\mathcal{Y}} \to \mathbf{R}$ is concave for each $y \in \mathcal{Y}$.

The following loss functions are η-exp-concave:

- The *logarithmic loss* $\ell(\widehat{y}, y) = -\log \widehat{y}(y)$, with $\mathcal{Y}$ a finite set and $\widehat{\mathcal{Y}} = \mathcal{P}(\mathcal{Y})$, with $\eta = 1$ (see Example 3.2 above);
- The *quadratic loss* $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ on $\mathcal{Y} = \widehat{\mathcal{Y}} = [-B, B] \subset \mathbf{R}$, with $\eta = 1/(8B^2)$.

We start with Lemma 3.1, which states that the prediction function used in AMF (see Definition 3.3) satisfies a regret bound where the regret is computed with respect to any pruning $\mathcal{T}$ of $\mathcal{T}^{\Pi}$.

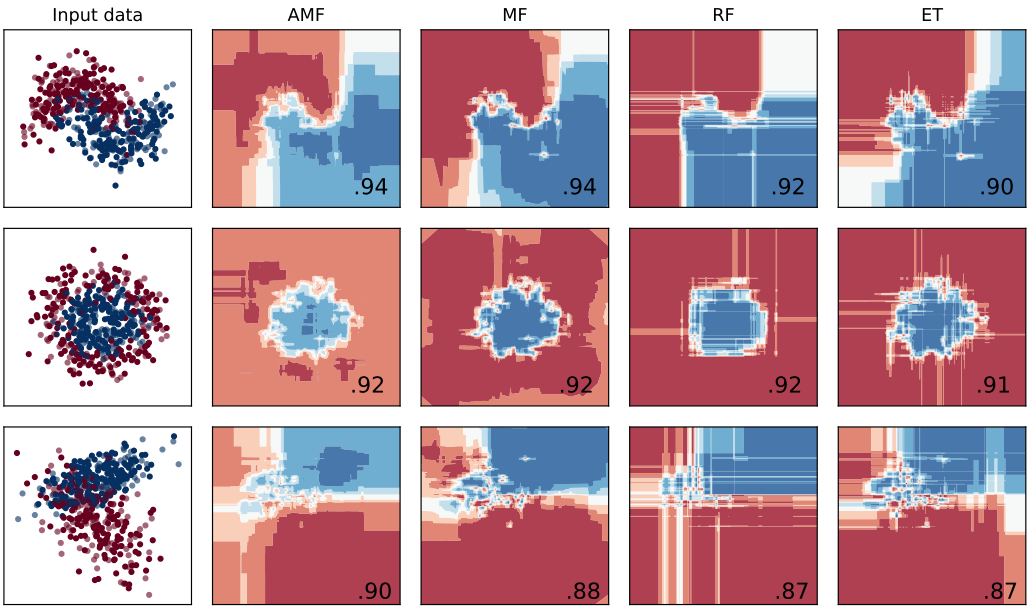


Figure 3.3: Decision functions of AMF, Mondrian Forest (MF), Breiman's batch random forest (RF) and batch Extra Trees (ET), on several toy datasets for binary classification (input data $n = 500$). We observe that AMF, thanks to aggregation, leads to a smooth decision function with a better generalization property (AUC on the test sets, displayed bottom right of each plot, is slightly better in all cases). Let us stress that both AMF and MF do a *single pass* on the data, while RF and ET require many passes. All algorithms use a forest containing 10 trees.

Lemma 3.1. *Consider a η-exp-concave loss function ℓ. Fix a realization $\Pi = (\mathcal{T}^\Pi, \Sigma^\Pi) \sim \mathsf{MP}$ and let $\mathcal{T} \subset \mathcal{T}^\Pi$ be a finite subtree. For every sequence $(x_1, y_1), \dots, (x_n, y_n)$, the prediction functions $\widehat{f}_1, \dots, \widehat{f}_n$ based on Π and computed by AMF satisfy*

$$\sum_{t=1}^{n} \ell(\widehat{f}_t(x_t), y_t) - \sum_{t=1}^{n} \ell(\widehat{y}_{\mathcal{T},t}(x_t), y_t) \leqslant \frac{1}{\eta} |\mathcal{T}| \log 2, \tag{3.5}$$

where we recall that $|\mathcal{T}|$ is the number of nodes in $\mathcal{T}$.

Lemma 3.1 is a direct consequence of a standard regret bound for the exponential weights algorithm (see Lemma 3.4 from Section 3.7), together with the fact that the Context Tree Weighting algorithm performed in Algorithms 3 and 4 computes it exactly, as stated in Proposition 3.1. By combining Lemma 3.1 with regret bounds for the online algorithms used in each node, both for the logarithmic loss (see Example 3.2) and the quadratic loss (see Example 3.1), we obtain the following regret bounds with respect to any pruning $\mathcal{T}$ of $\mathcal{T}^\Pi$.

Corollary 3.1 (Classification)**.** *Fix $\Pi = (\mathcal{T}^\Pi, \Sigma^\Pi)$ as in Lemma 3.1 and consider the classification setting described in Example 3.2 above. For any finite subtree $\mathcal{T}$ of $\mathcal{T}^\Pi$ and every sequence $(x_1, y_1), \dots, (x_n, y_n)$, the prediction functions $\widehat{f}_1, \dots, \widehat{f}_n$ based on Π computed by AMF with $\eta = 1$ satisfy*

$$\sum_{t=1}^{n} \ell(\widehat{f}_t(x_t), y_t) - \sum_{t=1}^{n} \ell(g_{\mathcal{T}}(x_t), y_t) \leqslant |\mathcal{T}| \log 2 + \frac{(|\mathcal{T}| + 1)(|\mathcal{Y}| - 1)}{4} \log(4n) \tag{3.6}$$

for any function $g_{\mathcal{T}} : [0,1]^d \to \mathcal{P}(\mathcal{Y})$ which is constant on the leaves of $\mathcal{T}$.

Corollary 3.2 (Regression)**.** *Fix $\Pi = (\mathcal{T}^\Pi, \Sigma^\Pi)$ as in Lemma 3.1 and consider the regression setting described in Example 3.1 above with $\mathcal{Y} = [-B, B]$. For every finite subtree $\mathcal{T}$ of $\mathcal{T}^\Pi$ and every sequence $(x_1, y_1), \dots, (x_n, y_n)$, the prediction functions $\widehat{f}_1, \dots, \widehat{f}_n$ based on Π computed by AMF with $\eta = 1/(8B^2)$ satisfy*

$$\sum_{t=1}^{n} \ell(\widehat{f}_t(x_t), y_t) - \sum_{t=1}^{n} \ell(g_{\mathcal{T}}(x_t), y_t) \leqslant 4B^2 (|\mathcal{T}| + 1) \log n \tag{3.7}$$

for any function $g_{\mathcal{T}} : [0,1]^d \to \mathcal{Y}$ which is constant on the leaves of $\mathcal{T}$.

The proofs of Corollaries 3.1 and 3.2 are given in Section 3.7, and rely in particular on Lemmas 3.5 and 3.6 that provide regret bounds for the online predictors $\widehat{y}_{\mathbf{v},t}$ considered in the nodes. Corollaries 3.1 and 3.2 that control the regret with respect to any pruning of $\mathcal{T}^\Pi$ imply in particular regret bounds with respect to any *time pruning* of MP.

Definition 3.5 (Time pruning)**.** For $\lambda > 0$, the *time pruning* Π_λ of Π at time λ is obtained by removing any node $\mathbf{v}$ whose creation time $\tau_{\mathbf{v}}$ satisfies $\tau_{\mathbf{v}} > \lambda$. We denote by $\mathsf{MP}(\lambda)$ the distribution of the tree partition Π_λ of $[0,1]^d$.

The parameter λ corresponds to a complexity parameter, allowing to choose a subtree of $\mathcal{T}^\Pi$ where all leaves have a creation time not larger than λ. We obtain the following regret bound for the regression setting (a similar statement holds for the classification setting), where the regret is with respect to any time pruning Π_λ of Π.

Corollary 3.3. *Consider the same regression setting as in Corollary 3.2. Then, AMF with* $\eta = 1/(8B^2)$ *satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{f}_t(x_t), y_t)\right] \leqslant \mathbb{E}\left[\inf_{g} \sum_{t=1}^{n} \ell(g(x_t), y_t)\right] + 8B^2(1+\lambda)^d \log n\,, \tag{3.8}$$

where the expectations on both sides are over the random sampling of the partition $\Pi_\lambda \sim \mathsf{MP}(\lambda)$*, and the infimum is taken over all functions* $g : [0,1]^d \to \mathbf{R}$ *that are constant on the cells of* Π_λ*.*

Corollary 3.3 controls the regret of AMF with respect to a time pruned Mondrian partition $\Pi_\lambda \sim \mathsf{MP}(\lambda)$ for *any* $\lambda > 0$. This result is one of the main ingredients allowing to prove that AMF is able to adapt to the unknown smoothness of the regression function, as stated in Theorem 3.2 below. Corollary 3.3, proven in Section 3.7, follows from the fact that $\mathbb{E}[|\mathcal{L}(\Pi_\lambda)|] = (1+\lambda)^d$ whenever $\Pi_\lambda \sim \mathsf{MP}(\lambda)$, where $|\mathcal{L}(\Pi_\lambda)|$ stands for the number of leaves in the partition Π_λ (Proposition 2.2). Let us pause for a minute and discuss the choice of the prior used in AMF, compared to what is done in literature with Bayesian approaches for instance.

Prior choice. The use of a branching process prior on prunings of large trees is common in literature on Bayesian regression trees. Indeed, Chipman et al. (1998) choose a branching process prior on subtrees, with a splitting probability of each node $\mathbf{v}$ of the form

$$\alpha(1+d_{\mathbf{v}})^{-\beta} \tag{3.9}$$

for some $\alpha \in (0,1)$ and $\beta \geqslant 0$, where $d_{\mathbf{v}}$ is the depth of note $\mathbf{v}$. Note that the locations of the splits themselves are also parameters of the Bayesian model, which enables more flexible estimation, but prevents efficient closed-form computations. The same prior on subtrees is used in the BART algorithm (Chipman et al., 2010), which considers sums of trees. We note that several values are proposed for the parameters (α, β), although there does not appear to be any definitive choice or criterion (Chipman et al. 1998 considers several examples with $\beta \in [\frac{1}{2}, 2]$, while Chipman et al. 2010 suggest $(\alpha, \beta) = (0.95, 2)$ for BART). The prior π considered here (see Definition 3.3) has a splitting probability 1/2 for each node, so that $(\alpha, \beta) = (1/2, 0)$. One appeal of the regret bounds stated above is that it offers guidance on the choice of parameters. Indeed, it follows as a by-product of our analysis that the regret of AMF (with any prior π) with respect to a subtree $\mathcal{T}$ is $O(\log \pi(\mathcal{T})^{-1})$. This suggests to choose π in AMF as flat as possible, namely $(\alpha, \beta) = (1/2, 0)$.

3.3.2 Adaptive minimax rates through online to batch conversion

In this Section, we show how to turn the algorithm described in Section 3.2 into a supervised learning algorithm with generalization guarantees that entail as a by-product adaptive minimax rates for nonparametric estimation. Namely, this section is concerned with bounds on the risk (expected prediction error on unseen data) rather than the regret of the sequence of prediction functions that was studied in Section 3.3.1. Therefore, we assume in this Section that the sequence $(x_1, y_1), (x_2, y_2), \ldots$ consists of i.i.d. random variables in $[0,1]^d \times \mathcal{Y}$, such that each (x_t, y_t) that comes sequentially is distributed as some generic pair (x, y). The quality of a *prediction function* $g : [0,1]^d \to \mathcal{Y}$ is measured by its risk defined as

$$R(g) = \mathbb{E}[\ell(g(x), y)]\,. \tag{3.10}$$

Online to batch conversion. Our supervised learning algorithm remains *online* (it does not require the knowledge of a fixed number of points n in advance). It is also virtually parameter-free, the only parameter being the learning rate η (set to 1 for the log-loss). In order to obtain a supervised learning algorithm with provable guarantees, we use *online to batch* conversion from Cesa-Bianchi et al. (2004), which turns any regret bound for an online algorithm into an excess risk bound for the average or a randomization of the past values of the online algorithm. As explained below, it enables to obtain fast rates for the excess risk, provided that the online procedure admits appropriate regret guarantees.

Lemma 3.2 (Online to batch conversion)**.** *Assume that the loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}^+$ is measurable, with $\widehat{\mathcal{Y}}$ a measurable space, and let $\mathcal{G}$ be a class of measurable functions $[0,1]^d \to \widehat{\mathcal{Y}}$. Given $f_1, \dots, f_n$ where $f_t : ([0,1]^d \times \mathcal{Y})^{t-1} \to \widehat{\mathcal{Y}}^{[0,1]^d}$, we denote $\widehat{f}_t = f_t((x_1,y_1),\dots,(x_{t-1},y_{t-1}))$. Let $\widetilde{f}_n = \widehat{f}_{I_n}$ with I_n a random variable uniformly distributed on $\{1,\dots,n\}$. Then, we have*

$$\mathbb{E}[R(\widetilde{f}_n)] - R(g) = \frac{1}{n}\mathbb{E}\Big[\sum_{t=1}^{n}\big(\ell(\widehat{f}_t(x_t), y_t) - \ell(g(x_t), y_t)\big)\Big], \tag{3.11}$$

which entails that the expected excess risk of $\widetilde{f}_n$ with respect to any $g \in \mathcal{G}$ is equal to the expected per-round regret of $\widehat{f}_1, \dots, \widehat{f}_n$ with respect to g.

Although this result is well-known (Cesa-Bianchi et al., 2004), we provide for completeness a proof of this specific formulation in Section 3.7. In our case, $\mathcal{G}$ will be the (random) family of functions that are constant on the leaves of some pruning of an infinite Mondrian partition Π, and $\widehat{f}_1, \dots, \widehat{f}_n$ will be the sequence of prediction functions of AMF. Note that, when conditioning on Π which is used to define both the class $\mathcal{G}$ and the algorithm, both $\mathcal{G}$ and the maps $f_1, \dots, f_n$ become deterministic, so that we can apply Lemma 3.2 conditionally on Π. In what follows, we denote by $\widetilde{f}_n$ the outcome of the online to batch conversion applied to our online procedure.

Oracle inequality and minimax rates. Let us show now that $\widetilde{f}_n$ achieves adaptive minimax rates under nonparametric assumptions, which complements and improves previous results (Mourtada et al., 2017, 2018). Indeed, AMF addresses the practical issue of optimally tuning the complexity parameter λ of Mondrian trees, while remaining a very efficient online procedure. As the next result shows, the procedure $\widetilde{f}_n$, which is virtually parameter-free, performs at least almost as well as the Mondrian tree with the best λ chosen with hindsight. For the sake of conciseness, Theorems 3.1 and 3.2 are stated only in the regression setting, although a similar result holds for the log-loss.

Theorem 3.1. *Consider the same setting as in Corollary 3.2, the only difference being the fact that the sequence $(x_1,y_1),\dots,(x_n,y_n)$ is i.i.d. and consider the online to batch conversion $\widetilde{f}_n$ from Lemma 3.2 applied to AMF. For every $\lambda > 0$ and every function g_λ which is constant on the cells of a random partition $\Pi_\lambda \sim \mathsf{MP}(\lambda)$, we have*

$$\mathbb{E}[R(\widetilde{f}_n)] - \mathbb{E}[R(g_\lambda)] \leqslant 8B^2(1+\lambda)^d \frac{\log n}{n}\,. \tag{3.12}$$

The proof of Theorem 3.1 is given in Section 3.7. It provides an oracle bound which is distribution-free, since it requires no assumption on the joint distribution of (x,y) apart from

$\mathcal{Y} = [-B, B]$. Combined with previous results on Mondrian partitions (Mourtada et al., 2018), which enable to control the approximation properties of Mondrian trees, Theorem 3.1 implies that $\widetilde{f}_n$ is adaptive with respect to the smoothness of the regression function, as shown in Theorem 3.2.

Theorem 3.2. *Consider the same setting as in Theorem 3.1 and assume that the regression function $f^*(\cdot) = \mathbb{E}[y|x = \cdot]$ is β-Hölder with $\beta \in (0, 1]$ unknown. Then, we have*

$$\mathbb{E}[(\widetilde{f}_n(x) - f^*(x))^2] = O\Big(\Big(\frac{\log n}{n}\Big)^{2\beta/(d+2\beta)}\Big), \tag{3.13}$$

which is (up to the $\log n$ term) the minimax optimal rate of estimation over the class of β-Hölder functions.

The proof of Theorem 3.2 is given in Section 3.7. Theorem 3.2 states that the online to batch conversion $\widetilde{f}_n$ of AMF is adaptive to the unknown Hölder smoothness $\beta \in (0, 1]$ of the regression function since it achieves, up to the $\log n$ term, the minimax rate $n^{-2\beta/(d+2\beta)}$, see Stone (1982). It would be theoretically possible to modify the procedure in order to ensure adaptivity to higher regularities (say, up to some order $\bar{\beta} \in \mathbf{N} \setminus \{0\}$), by replacing the constant estimates inside each node by polynomials (of order $\bar{\beta} - 1$). However, this would lead to a numerically involved procedure, that is beyond the scope of the chapter. In addition, it is known that averaging can reduce the bias of individual randomized tree estimators for twice differentiable functions, see Arlot and Genuer (2014) and Mourtada et al. (2018) for Mondrian Forests. Such results cannot be applied to AMF, since its decision function involves a more complicated process of aggregation over all subtrees.

3.4 Practical implementation of AMF

This section describes a modification of AMF that we use in practice, in particular for all the numerical experiments performed in the present chapter. Because of extra technicalities involved with the modified version described below, we are not able to provide theoretical guarantees similar to what is done in Section 3.3. Indeed, the procedure described in this section exhibits a more intricate behaviour: new splits may be inserted above previous splits, which affects the underlying tree structure as well as the underlying prior over subtrees. This section mainly modifies the procedures described in Algorithms 3 and 4 so that splits are sampled only within the range of the features seen in each node, see Section 3.4.1, with motivations to do so described below. Moreover, we provide in Section 3.4.2 a guarantee on the average computational complexity of AMF through a control of the expected depth of the Mondrian partition.

3.4.1 Restriction to splits within the range of sample points

Algorithm 3 from Section 3.2.3 samples successive splits on the whole domain $[0, 1]^d$. In particular, when a new features vector x_t is available, it samples splits of the leaf $\mathbf{v}_t$ containing x_t until a split successfully separates x_t from the other point $x_s \neq x_t$ contained in $\mathbf{v}_t$ (unless $\mathbf{v}_t$ was empty). In the process, several splits outside of the box containing x_s and x_t can be performed. These splits are somewhat superfluous, since they induce empty leaves and delay the split that separates these two points. Removing those splits is critical to the performance

of the method, in particular when the ambient dimension of the features is not small. In such cases, many splits may be needed to separate the feature points. On the other hand, only keeping those splits that are necessary to separate the sample points may yield a more adaptive partition, which can better adapt to a possible low-dimensional structure of the distribution of x.

We describe below a modified algorithm that samples splits in the range of the features vectors seen in each cell, exactly as in the original Mondrian Forest algorithm (Lakshminarayanan et al., 2014). In particular, each leaf will contain exactly one sample point by construction (possibly with repetition if $x_s = x_t$ for some $s \neq t$) and no empty leaves. Formally, this procedure amounts to considering the *restriction* of the Mondrian partition to the finite set of points $\{x_1, \ldots, x_t\}$ (Lakshminarayanan et al., 2014), where it is shown that such a restricted Mondrian partition can be updated efficiently in an online fashion, thanks to properties of the Mondrian process. This update exploits the creation time $\tau_{\mathbf{v}}$ of each node, as well as the range of the features vectors $R_{\mathbf{v}}$ seen inside each node (as opposed to only leaves). Moreover, this procedure can possibly split an interior node and not only a leaf. The algorithm considered here is a modification of the procedure `ExtendMondrianTree`$(\mathcal{T}, \lambda, (x_t, y_t))$ described in Lakshminarayanan et al. (2014), where we use $\lambda = +\infty$ and where we perform the exponentially weighted aggregation of subtrees described in Section 3.2.

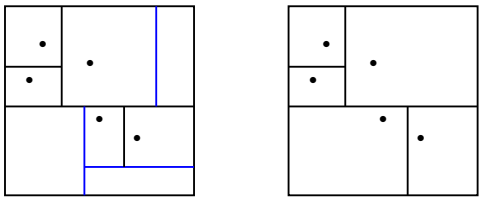


Figure 3.4: Unrestricted (left) *vs.* restricted (right) Mondrian partitions. Dots (•) represent sample points. In both cases, cells containing one sample point are no longer split. In addition, the restricted Mondrian partition is obtained by removing from the unrestricted partition all splits (in blue) that create empty leaves.

We call the former partition (from Section 3.2.3) an *unrestricted* Mondrian partition, while the one described here will be referred to as a *restricted* Mondrian partition. The difference between the two is illustrated in Figure 3.4. The tree $\mathcal{T}$ contains, as before, parent and children relations between all nodes $\mathbf{v} \in \mathcal{T}$, while each $\sigma_{\mathbf{v}} \in \Sigma$ contains

$$\sigma_{\mathbf{v}} = (j_{\mathbf{v}}, s_{\mathbf{v}}, \tau_{\mathbf{v}}, \widehat{y}_{\mathbf{v}}, w_{\mathbf{v}}, \overline{w}_{\mathbf{v}}, R_{\mathbf{v}}), \tag{3.14}$$

which differs from Equation (3.4) since we keep in memory the creation time $\tau_{\mathbf{v}}$ of $\mathbf{v}$, and the range

$$R_{\mathbf{v}} = \prod_{j=1}^{d} [a_{\mathbf{v}}^j, b_{\mathbf{v}}^j]$$

of features vectors in $C_{\mathbf{v}}$ instead of $x_{\mathbf{v}}$ (a past sample point). Another advantage of the restricted Mondrian partition is that the algorithm is *range-free*, since it does not require to assume that all features vectors are in $[0,1]^d$ (we simply use as initial root cell $C_\epsilon = \mathbf{R}^d$).

Algorithms 5 and 6 below implement AMF with a restricted Mondrian partition, and are used instead of the previous Algorithm 3 in our numerical experiments. These algorithms, together with Algorithm 7 below for prediction, maintain in memory, as in Section 3.2, the

current state of the Mondrian partition $\Pi = (\mathcal{T}, \Sigma)$, which contains the tree structure $\mathcal{T}$ (containing parent/child relationships between nodes) and data $\sigma_{\mathbf{v}} \in \Sigma$ for all nodes, see Equation (3.14). An illustration of Algorithms 5 and 6 is provided in Figure 3.5. We use the notation $x_+ = \max(x, 0)$ for any $x \in \mathbf{R}$.

Algorithm 5 `AmfUpdate`(x, y) : update AMF with a new sample $(x, y) \in \mathbf{R}^d \times \mathcal{Y}$

1: **Input:** a new sample $(x, y) \in \mathbf{R}^d \times \mathcal{Y}$
2: **if** $\mathcal{T} = \varnothing$ **then**
3: Put $\mathcal{T} = \{\epsilon\}$ and $(\tau_\epsilon, \widehat{y}_\epsilon, w_\epsilon, \overline{w}_\epsilon, R_\epsilon) = (0, h(\varnothing), 1, 1, \{x\})$
4: **else**
5: Call `NodeUpdate`(ϵ, x) from Algorithm 6
6: **end if**
7: Let $\mathbf{v}(x)$ be the leaf such that $x \in C_{\mathbf{v}(x)}$ and put $\mathbf{v} = \mathbf{v}(x)$
8: Let `continueUp` = `true`.
9: **while** `continueUp` **do**
10: Set $w_{\mathbf{v}} = w_{\mathbf{v}} \exp(-\eta \ell(\widehat{y}_{\mathbf{v}}, y))$
11: Set $\overline{w}_{\mathbf{v}} = w_{\mathbf{v}}$ if $\mathbf{v}$ is a leaf and $\overline{w}_{\mathbf{v}} = \frac{1}{2} w_{\mathbf{v}} + \frac{1}{2} \overline{w}_{\mathbf{v}0} \overline{w}_{\mathbf{v}1}$ otherwise
12: Update $\widehat{y}_{\mathbf{v}}$ using y (following Definition 3.2)
13: If $\mathbf{v} \neq \epsilon$ let $\mathbf{v} =$ `parent`$(\mathbf{v})$, otherwise let `continueUp` = `false`
14: **end while**

In algorithm 5, Line 3 initializes the tree the first time `AmfUpdate` is called, otherwise the recursive procedure `NodeUpdate` is used to update the restricted Mondrian partition, starting at the root ϵ. Lines 7–14 perform the update of the aggregation weights in the same way as what we did in Section 3.2.3.

In Algorithm 6, Line 2 computes the range extension of x with respect to $R_{\mathbf{v}}$. In particular, if $x \in R_{\mathbf{v}}$, then no split will be performed and we go directly to Line 15. Otherwise, if x is outside of $R_{\mathbf{v}}$, a split of $\mathbf{v}$ is performed whenever $\tau_{\mathbf{v}} + E < \tau_{\mathbf{v}0}$ (a new node created at time $\tau_{\mathbf{v}} + E$ can be inserted before the creation time $\tau_{\mathbf{v}0}$ of the current child $\mathbf{v}0$ of $\mathbf{v}$). In this case, we sample the split coordinate j proportionally to Δ_j (coordinates with the largest extension are more likely to be used to split $\mathbf{v}$) and we sample the split threshold uniformly at random within the corresponding extension (Line 7 or Line 9). Now, at Line 11, we move downwards the whole tree rooted at $\mathbf{v}$: any node at index $\mathbf{v}\mathbf{v}'$ for any $\mathbf{v}' \in \mathcal{T}_{\mathbf{v}}$ is renamed as $\mathbf{v}(1-a)\mathbf{v}'$. For instance, if $a = 0$ (Line 7, the extension is on the left of the current range), the node $\mathbf{v}0$ is renamed as $\mathbf{v}10$, the node $\mathbf{v}1$ as $\mathbf{v}11$, etc. Then, at Line 12, new nodes $\mathbf{v}0$ and $\mathbf{v}1$ are created, where $\mathbf{v}a$ is a new leaf containing x and $\mathbf{v}(1-a)$ is a new node which is the root of the subtree we moved downwards at Line 11. Line 12 also initializes $\sigma_{\mathbf{v}a}$ and copies $\sigma_{\mathbf{v}}$ into $\sigma_{\mathbf{v}(1-a)}$. The process performed in Lines 11–12 therefore simply inserts two new nodes below $\mathbf{v}$ (since we just split node $\mathbf{v}$): a leaf containing x, and another node rooting the tree that was rooted at $\mathbf{v}$ before the split. Line 13 updates the range of $\mathbf{v}$ using x and exits the procedure. If no split is performed, Line 15 updates the range of $\mathbf{v}$ using x and calls `NodeUpdate` on the child of $\mathbf{v}$ containing x.

The prediction algorithm described in Algorithm 7 below is a modification of Algorithm 4, where we use `NodeUpdate` instead of Algorithm 3. Finally, the algorithm used in our experiments do not use the online to batch conversion from Section 3.3.2: it simply uses the current tree, namely the most recent updated Mondrian tree partition $\Pi_{t+1} = (\mathcal{T}_{t+1}, \Sigma_{t+1})$

Algorithm 6 `NodeUpdate`$(\mathbf{v}, x)$: update node $\mathbf{v}$ using x

Input: a node $\mathbf{v} \in \mathcal{T}$ from the current tree and a features vector $x \in \mathbf{R}^d$
Let $\Delta_j = (x_j - b_{\mathbf{v}}^j)_+ + (a_{\mathbf{v}}^j - x_j)_+$ and $\Delta = \sum_{j=1}^d \Delta_j$
Sample $E \sim \mathsf{Exp}(\Delta)$ and put $E = +\infty$ if $\Delta = 0$ (namely $x \in R_{\mathbf{v}}$)
if $\mathbf{v}$ is a leaf **or** $\tau_{\mathbf{v}} + E < \tau_{\mathbf{v}0}$ **then**
    Sample a split coordinate $J \in \{1, \ldots, d\}$ with $\mathbb{P}(J = j) = \Delta_j / \Delta$
    if $x_J < a_{\mathbf{v}}^J$ **then**
        Put $a = 0$ and sample the split threshold $S | J \sim \mathcal{U}([x_J, a_{\mathbf{v}}^J])$
    else
        Put $a = 1$ and sample the split threshold $S | J \sim \mathcal{U}([b_{\mathbf{v}}^J, x_J])$
    end if
    Set $\mathcal{T}_{\mathbf{v}(1-a)} = \mathcal{T}_{\mathbf{v}}$, namely nodes $\mathbf{v}\mathbf{v}'$ are renamed as $\mathbf{v}(1-a)\mathbf{v}'$ for any $\mathbf{v}' \in \mathcal{T}_{\mathbf{v}}$
    Create nodes $\mathbf{v}0$ and $\mathbf{v}1$ and put $(\tau_{\mathbf{v}a}, \widehat{y}_{\mathbf{v}a}, w_{\mathbf{v}a}, \overline{w}_{\mathbf{v}a}, R_{\mathbf{v}a}) = (\tau_{\mathbf{v}} + E, h(\varnothing), 1, 1, \{x\})$, put $\sigma_{\mathbf{v}(1-a)} = \sigma_{\mathbf{v}}$ (see Equation 3.14) but set $\tau_{\mathbf{v}(1-a)} = \tau_{\mathbf{v}} + E$
    Put $a_{\mathbf{v}}^j = \min(a_{\mathbf{v}}^j, x_j)$ and $b_{\mathbf{v}}^j = \max(b_{\mathbf{v}}^j, x_j)$
else
    Put $a_{\mathbf{v}}^j = \min(a_{\mathbf{v}}^j, x_j)$ and $b_{\mathbf{v}}^j = \max(b_{\mathbf{v}}^j, x_j)$
    Let $a \in \{0, 1\}$ be such that $x \in C_{\mathbf{v}a}$ and call `NodeUpdate`$(\mathbf{v}a, x)$
end if

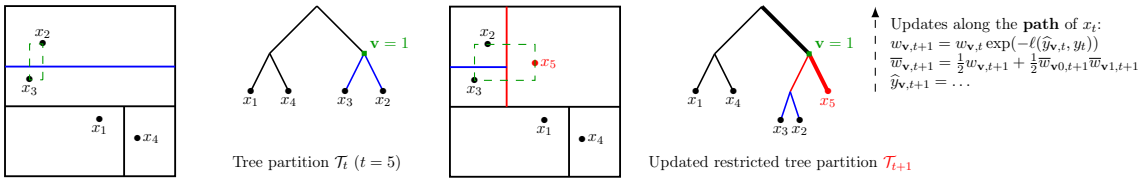


Figure 3.5: Illustration of the `AmfUpdate`(x_t, y_t) procedure from Algorithms 5 and 6: update of the partition, weights and node predictions as a new data point (x_t, y_t) for $t = 5$ becomes available. *Left*: tree partition Π_t before seeing (x_t, y_t). *Right*: update of the partition using (x_t, y_t). The path of x_t in the tree is indicated in bold. In green is the node $\mathbf{v} = 1$ and dashed lines indicates its range R_1. Since x_5 is outside of R_1 at $t = 5$, the range is extended. A new split (in red) is sampled in the extended range, since its creation time $\tau_1 + E$ is smaller than the one of the next split τ_{10} and two new nodes named 10 and 11 are inserted below 1, while the previous nodes 10 and 11 are moved as 100 and 101. The weights and node predictions are then updated using an upwards path from the new leaf containing x to the root (in bold). All leaves contain exactly one single point.

after calling `AmfUpdate`$(x_1, y_1), \ldots,$ `AmfUpdate`(x_t, y_t), where (x_t, y_t) is the last sample seen.

3.4.2 Computational complexity

The next Proposition provides a bound on the average depth of a Mondrian tree. This is of importance, since the computational complexities of `AmfUpdate` and `AmfPredict` are linear with respect to this depth, see below for a discussion.

Proposition 3.2. *Assume that x has a density p satisfying the following property: there exists*

Algorithm 7 `AmfPredict`(x): predict the label of $x \in [0,1]^d$

Input: a features vector $x \in [0,1]^d$
Call `NodeUpdate`(ϵ, x) in order to obtain a temporary update of the current partition Π using x and let $\mathbf{v}(x)$ be the leaf such that $x \in C_{\mathbf{v}(x)}$
Set $\widetilde{y}_{\mathbf{v}} = \widehat{y}_{\mathbf{v}(x)}$
while $\mathbf{v} \neq \epsilon$ **do**
    Let $(\mathbf{v}, \mathbf{v}a) = (\texttt{parent}(\mathbf{v}), \mathbf{v})$ (for some $a \in \{0,1\}$)
    Let $\widetilde{y}_{\mathbf{v}} = \frac{1}{2}\frac{w_{\mathbf{v}}}{\overline{w}_{\mathbf{v}}}\widehat{y}_{\mathbf{v}} + \frac{1}{2}\frac{\overline{w}_{\mathbf{v}(1-a)}\overline{w}_{\mathbf{v}a}}{\overline{w}_{\mathbf{v}}}\widetilde{y}_{\mathbf{v}a}$
end while
Return $\widetilde{y}_{\epsilon}$

a constant $M > 0$ such that, for every $x', x'' \in [0,1]^d$ which only differ by one coordinate,

$$\frac{p(x')}{p(x'')} \leqslant M\,. \tag{3.15}$$

Then, the depth $D_n^{\Pi}(x)$ of the leaf containing a random point x in the Mondrian tree restricted to the observations $x_1, \ldots, x_n, x$ satisfies

$$\mathbb{E}[D_n^{\Pi}(x)] \leqslant \frac{\log n}{\log[(2M)/(2M-1)]} + 2M\,.$$

Assumption (3.15) is satisfied when p is upper and lower bounded: $c \leqslant p \leqslant C$ with $M = C/c$, but this assumption is weaker: for instance, it only implies that $M^{-d} \leqslant p \leqslant M^d$, which is a milder property when the dimension d is large. The proof of Proposition 3.2 is given in Section 3.7. Since the lower bound is also trivially $\Omega(\log n)$ (a binary tree with n nodes has at least a depth $\log_2 n$), Proposition 3.2 entails that $\mathbb{E}[D_n^{\Pi}] = \Theta(\log n)$. If the number of features is d, then the update complexity of a single tree is $\Theta(d \log n)$, which makes full online training time $\Theta(nd \log n)$ over a dataset of size n. Prediction is $\Theta(\log n)$ since it requires a downwards and upwards path on the tree (see Algorithms 4 and 7).

3.5 Numerical experiments

This Section proposes a thorough comparison of AMF with several baselines on several datasets for multi-class classification. We describe all the considered algorithms in Section 3.5.1, including both online methods and batch methods. A comparison of the average losses (assessing the online performance of algorithms) on several datasets is given in Section 3.5.3 for online methods only. Batch and online methods are compared in Section 3.5.4 and an experiment comparing the sensitivity of all methods with respect to the number of trees used is given in Section 3.5.5. All these experiments are performed for multi-class classification problems, using datasets described in Section 3.5.2.

3.5.1 Algorithms

In this Section, we describe precisely the procedures considered in our experiments for online and batch classification problems.

AMF. This is the AMF algorithm with restricted Mondrian partitions described in Algorithms 5 and 7. AMF is implemented in Python and C++ in our open-source `tick` library, available at https://github.com/X-DataInitiative/tick, and is documented here https://x-datainitiative.github.io/tick/. We use `OnlineForestClassifier` from the `tick.online` module, with default parameters: we use 10 trees; we use aggregation with exponential weights with learning rate $\eta = 1$; we don't split nodes containing only a single data class.

Dummy. We consider a dummy baseline that only estimates the distribution of the labels (without taking into account the features) in an online manner. At step $t+1$, it simply computes the Krichevsky-Trofimov forecaster (see Example 3.2) $\widehat{y}_{t+1}(k) = (n_t(k) + 1/2)/(t + K/2)$ of the classes $k = 1, \dots, K$, where $n_t(k) = \sum_{s=1}^{t} \mathbf{1}(y_s = k)$.

MF (Mondrian Forest). This is the Mondrian Forest Lakshminarayanan et al. (2014, 2016) proposed in the `scikit-garden` library, available at https://github.com/scikit-garden/scikit-garden. We use `MondrianForestClassifier` in our experiments, with the default settings proposed with the method: 10 trees are used; no depth restriction is used on the trees; we stop growing the trees if all nodes have less than 2 samples; all trees are trained using the entire dataset (no bootstrap).

SGD (Stochastic Gradient Descent). This is logistic regression trained with a single pass of stochastic gradient descent. We use `SGDClassifier` from the `scikit-learn` library, see Pedregosa et al. (2011) and https://scikit-learn.org. We use a constant learning rate 0.1 and the default choice of ridge penalization with strength 0.0001, since it provides good results on all the datasets.

RF (Random Forests). This is Random Forest (Breiman, 2001a) for classification. We use the implementation available in the `scikit-learn` library, namely `RandomForestClassifer` from the `sklearn.ensemble` module. This is a reference implementation, which is highly optimized and among the fastest implementations available in the open-source community. Details on this implementation are available in Louppe (2014). Note that this is a batch algorithm, that cannot be trained sequentially, which requires a large number of passes through the data to optimize some impurity criterion (default is the Gini index). We use the default parameters of the procedure (with 10 trees).

ET (Extra Trees). This is the Extra Trees algorithm (Geurts et al., 2006). Once-again, we use the implementation available in the `scikit-learn` library, namely `ExtraTreesClassifier` from the `sklearn.ensemble` module. As for RF, it is a reference implementation from the open source community. We use the default parameters of the procedure (with 10 trees).

3.5.2 Considered datasets

The datasets we use are from the UCI Machine Learning repository, see Dua and Graff (2019) and are described in Table 3.2 below.

dataset	#samples	#features	#classes
adult	32561	107	2
bank	45211	51	2
car	1728	21	4
cardio	2126	24	3
churn	3333	71	2
default_cb	30000	23	2
letter	20000	16	26
satimage	5104	36	6
sensorless	58509	48	11
spambase	4601	57	2

Table 3.2: List of datasets from the UCI Machine Learning repository considered in our experiments.

3.5.3 Online learning: comparison of averaged losses

We compare the curves of averaged losses over time of all the considered online algorithms. At each round t, we reveal a new sample (x_t, y_t) and update all algorithms using this new sample. Then, we ask all algorithms to give a prediction $\widehat{y}_{t+1}$ of the label y_{t+1} associated to x_{t+1}, and compute the log-loss $\ell(\widehat{y}_{t+1}, y_{t+1})$ incurred by all algorithms. Along the rounds $t = 1, \ldots, n-1$ when the considered data has sample size n, we compute the average loss $\frac{1}{t-1}\sum_{s=1}^{t-1} \ell(\widehat{y}_{s+1}, y_{s+1})$. This is what is displayed in Figure 3.6 below, on 10 datasets for the online procedures AMF, Dummy, SGD and MF.

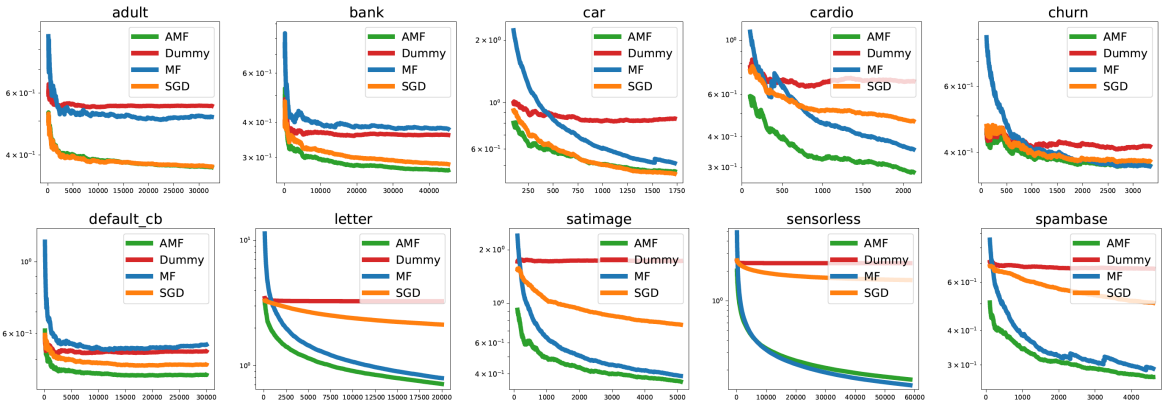


Figure 3.6: Average losses of all the online algorithms considered on 10 datasets for multi-class classification. The x-axis corresponds to the step t (number of samples revealed) and the y-axis is the value of average log loss obtained until this step (the lower the better). AMF almost always exhibits the smallest average loss on all the considered datasets.

On most datasets, AMF exhibits the smallest average loss, and is always competitive with respect to the considered baselines. As a comparison, the performance of SGD and MF strongly varies depending on the dataset: the "robustness" of AMF comes from the aggregation algorithm it uses, which always produces a non-overfitting and smooth decision function, as

illustrated in Figure 3.3 above, even in the early iterations. This is confirmed by the early values of the average losses observed in all displays in Figure 3.6, where we see that it is always the smallest compared to all the baselines.

3.5.4 Online versus Batch learning

In this Section, we consider a "batch" setting, where we hold out a test dataset (containing 30% of the whole data), and we consider only binary classification problems, in which all methods are assessed using the area under the ROC curve (AUC) on the test dataset. We consider the datasets `adult`, `bank`, `default`, `spambase` and all the methods (online and batch) described in Section 3.5.1. The performances of batch methods (RF and ET) are assessed only once using the test set, since these methods are not trained in an online fashion, but rather at once. Therefore, the test AUCs of these batch methods are displayed in Figure 3.7 as a constant horizontal line along the iterations. Online methods (AMF, Dummy, SGD and MF) are tested every 100 iterations: each time 100 samples are revealed, we produce predictions on the full test dataset, and report the corresponding test AUCs in Figure 3.7. We observe that, as more samples are revealed, the online methods improve their test AUCs.

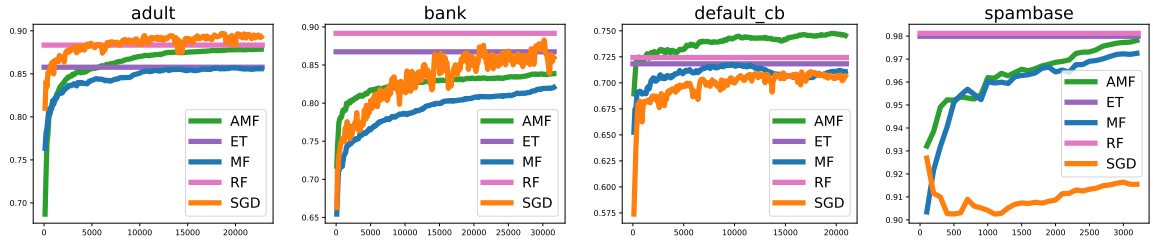


Figure 3.7: Area under the ROC curve (AUC) obtained on a held-out testing dataset (30% of the whole data) obtained by batch methods (RF and ET) and online methods (SGD, MF and AMF) on four binary classification datasets. The x-axis corresponds to the online steps (number of samples seen) over the train dataset. AMF is very competitive and always achieve a good AUC after a few steps. In the `defaultcb` dataset, AMF even improves the test AUC of RF and ET.

We observe that the batch methods RF and ET generally perform best; this ought to be expected, since their splits are optimized using training data, while those of AMF and MF are chosen on-the-fly. However, the performance of AMF is very competitive against all baselines. In particular, it performs better than MF and even improves upon ET and RF on the `default_cv` dataset.

3.5.5 Sensitivity to the number of trees

The aim of this Section is to exhibit another positive effect of the aggregation algorithm used in AMF. Indeed, we illustrate in Figure 3.8 below the fact that AMF can achieve good performances using less trees than MF, RF and ET. This comes from the fact that even a single tree in AMF can be a good classifier, since the aggregation algorithm used in it (see Section 3.2.2) aggregates all the prunings of the Mondrian tree. This allows to avoid overfitting, even when a single tree is used, as opposed to the other tree-based methods considered here. We consider in Figure 3.8 the same experimental setting as in Section 3.5.4, and compare the

test AUCs obtained on four binary classification problems for an increasing number of trees in all methods. The test AUCs obtained by all algorithms with $1, 2, 5, 10, 20$ and 50 trees are displayed in Figure 3.8, where the x-axis corresponds to the number of trees and the y-axis corresponds to the test AUC.

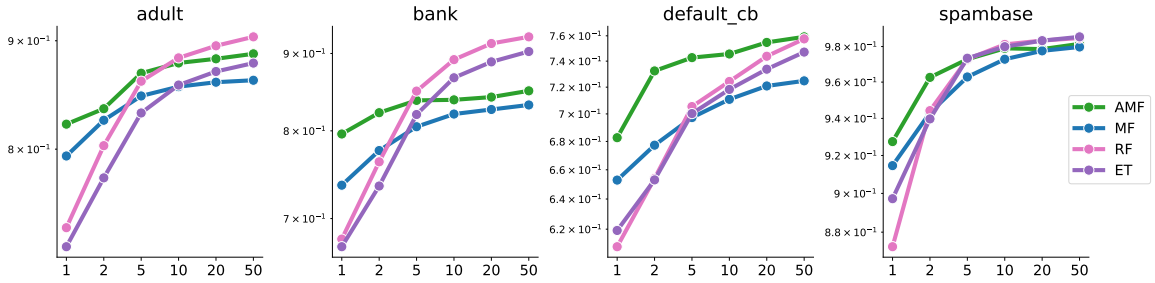


Figure 3.8: Area under the ROC curve (AUC) obtained on a held-out testing dataset (30% of the whole data) obtained by AMF, MF, RF and ET as a function of the number of trees used. We observe that AMF is less sensitive to the number of trees used in the forest than all the baselines, and that it has good performances even when using one or two trees.

We observe that when using one or two trees, AMF performs better than all the baselines. The performance of RF strongly increases with an increasing number of trees, and ends up with the best performances with 50 trees. The performance of AMF also improves when more trees are used (averaging over several realizations of the Mondrian partition certainly helps prediction), but the aggregation algorithm makes AMF somehow less sensitive to the number of trees in the forest.

3.6 Conclusion

In this chapter we introduced AMF, an online random forest algorithm based on a combination of Mondrian forests and an efficient implementation of an aggregation algorithm over all the prunings of the Mondrian partition. This algorithm is almost parameter-free, and has strong theoretical guarantees expressed in terms of regret with respect to the optimal time-pruning of the Mondrian partition, and in terms of adaptation with respect to the smoothness of the regression function. We illustrated on a large number of datasets the performances of AMF compared to strong baselines, where AMF appears as an interesting procedure for online learning.

A limitation of AMF, however, is that it does not perform feature selection. It would be interesting to develop an online feature selection procedure that could indicate along which coordinates the splits should be sampled in Mondrian trees, and prove that such a procedure performs dimension reduction in some sense. This is a challenging question in the context of online learning which deserves future investigations.

3.7 Proofs

This Section gathers the proofs of all the results of the Chapter, following their order of appearance, namely the proofs of Proposition 3.1, Lemma 3.1, Corollaries 3.1, 3.2 and 3.3, Lemma 3.2 and Theorems 3.1 and 3.2.

3.7.1 Proof of Proposition 3.1

Consider a realization $\Pi = (\mathcal{T}^\Pi, \Sigma^\Pi) \sim \mathsf{MP}$ of the infinite Mondrian partition, and assume that we are at step $t \geqslant 1$, namely we observed $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$ and performed the updates described in Algorithm 3 on each sample. Given $x \in [0,1]^d$, we want to predict the label (or its distribution) using

$$\widehat{f}_t(x) = \frac{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \pi(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})} \widehat{y}_{\mathcal{T},t}(x)}{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \pi(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})}}, \tag{3.16}$$

see Definition 3.3, where we recall that $\pi(\mathcal{T}) = 2^{-|\mathcal{T}|}$ with $|\mathcal{T}|$ the number of nodes in $\mathcal{T}$ and where we recall that the sum in (3.16) is an infinite sum over all subtrees $\mathcal{T}$ of $\mathcal{T}^\Pi$. See also Definition 3.2 for the tree prediction $\widehat{y}_{\mathcal{T},t}(x)$.

Reduction to a finite sum. Let $\mathbf{T}$ denote the minimal subtree of $\mathcal{T}^\Pi$ that separates the elements of $\{x_1, \dots, x_{t-1}, x\}$ (if $x = x_t$ then $\mathbf{T} = \mathcal{T}_{t+1}$). Also, for every finite tree $\mathcal{T}$, denote $\mathcal{T}|_{\mathbf{T}} := \mathcal{T} \cap \mathbf{T}$. For any subtree $\mathcal{T}$ of $\mathbf{T}$, we have

$$\sum_{\mathcal{T}' : \mathcal{T}'|_{\mathbf{T}} = \mathcal{T}} \pi(\mathcal{T}') = 2^{-\|\mathcal{T}\|} =: \pi_{\mathbf{T}}(\mathcal{T}), \tag{3.17}$$

where $\|\mathcal{T}\|$ denotes the number of nodes of $\mathcal{T}$ which are not leaves of $\mathbf{T}$; note that $\pi_{\mathbf{T}}$ is a probability distribution on the subtrees of $\mathbf{T}$, since π is a probability distribution on finite subtrees of $\{0,1\}^*$. To see why Equation (3.17) is true, consider the following representation of π: let $(B_{\mathbf{v}})_{\mathbf{v} \in \{0,1\}^*}$ be an i.i.d. family of Bernoulli random variables with parameter $1/2$; a node $\mathbf{v}$ is said to be *open* if $B_{\mathbf{v}} = 1$, and *closed* otherwise. Then, denote $\mathcal{T}'$ the subtree of $\{0,1\}^*$ all of whose interior nodes are open, and all of whose leaves are closed; clearly, $\mathcal{T}' \sim \pi$. Now, $\mathcal{T}'|_{\mathbf{T}} = \mathcal{T}$ if and only if all interior nodes of $\mathcal{T}$ are open and all leaves of $\mathcal{T}$ except leaves of $\mathbf{T}$ are closed. By independence of the $B_{\mathbf{v}}$, this happens with probability $2^{-\|\mathcal{T}\|}$.

In addition, note that if $\mathcal{T}'$ is a finite subtree of $\{0,1\}^*$ and $\mathcal{T} = \mathcal{T}'|_{\mathbf{T}}$, then $\widehat{y}_{\mathcal{T}',t}(x) = \widehat{y}_{\mathcal{T},t}(x)$. Indeed, let $\mathbf{v}_{\mathcal{T}'}(x)$ be the leaf of $\mathcal{T}'$ that contains x; if $\mathbf{v}_{\mathcal{T}'}(x) \in \mathbf{T}$, then $\mathbf{v}_{\mathcal{T}'}(x) = \mathbf{v}_{\mathcal{T}}(x)$ and hence $\widehat{y}_{\mathcal{T}',t}(x) = \widehat{y}_{\mathbf{v}_{\mathcal{T}'}(x),t} = \widehat{y}_{\mathbf{v}_{\mathcal{T}}(x),t} = \widehat{y}_{\mathcal{T},t}(x)$; otherwise, by definition of $\mathbf{T}$, both $\mathbf{v}_{\mathcal{T}'}(x)$ and $\mathbf{v}_{\mathcal{T}}(x)$ only contain the x_s ($s \leqslant t-1$) such that $x_s = x$, so that again $\widehat{y}_{\mathbf{v}_{\mathcal{T}'}(x),t} = \widehat{y}_{\mathbf{v}_{\mathcal{T}}(x),t}$. Similarly, this result for $x = x_t$ also holds for x_s, $s \leqslant t-1$, so that $L_{t-1}(\mathcal{T}') = L_{t-1}(\mathcal{T})$. From the points above, it follows that

$$\begin{aligned}
\widehat{f}_t(x) &= \frac{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \pi(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})} \widehat{y}_{\mathcal{T},t}(x)}{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \pi(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})}} \\
&= \frac{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \sum_{\mathcal{T}' : \mathcal{T}'|_{\mathbf{T}} = \mathcal{T}} \pi(\mathcal{T}') e^{-\eta L_{t-1}(\mathcal{T}')} \widehat{y}_{\mathcal{T}',t}(x)}{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \sum_{\mathcal{T}' : \mathcal{T}'|_{\mathbf{T}} = \mathcal{T}} \pi(\mathcal{T}') e^{-\eta L_{t-1}(\mathcal{T}')}} \\
&= \frac{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \sum_{\mathcal{T}' : \mathcal{T}'|_{\mathbf{T}} = \mathcal{T}} \pi(\mathcal{T}') e^{-\eta L_{t-1}(\mathcal{T})} \widehat{y}_{\mathcal{T},t}(x)}{\sum_{\mathcal{T} \subset \mathcal{T}^\Pi} \sum_{\mathcal{T}' : \mathcal{T}'|_{\mathbf{T}} = \mathcal{T}} \pi(\mathcal{T}') e^{-\eta L_{t-1}(\mathcal{T})}} \\
&= \frac{\sum_{\mathcal{T} \subset \mathbf{T}} \pi_{\mathbf{T}}(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})} \widehat{y}_{\mathcal{T},t}(x)}{\sum_{\mathcal{T} \subset \mathbf{T}} \pi_{\mathbf{T}}(\mathcal{T}) e^{-\eta L_{t-1}(\mathcal{T})}}.
\end{aligned} \tag{3.18}$$

Computation for the finite tree $\mathbf{T}$. The expression in Equation (3.18) involves finite sums, over all subtrees of $\mathbf{T}$ (involving an exponential in the number of leaves of $\mathbf{T}$, namely t, terms). However, it can be computed efficiently because of the specific choice of the prior π. More precisely, we will use the following lemma (Helmbold and Schapire, 1997, Lemma 1) several times to efficiently compute sums of products. Let us recall that $\mathcal{N}(\mathbf{T})$ stands for the set of nodes of $\mathbf{T}$.

Lemma 3.3. *Let $g : \mathcal{N}(\mathbf{T}) \to \mathbf{R}$ be an arbitrary function and define $G : \mathcal{N}(\mathbf{T}) \to \mathbf{R}$ as*

$$G(\mathbf{v}) = \sum_{\mathcal{T}_\mathbf{v}} 2^{-\|\mathcal{T}_\mathbf{v}\|} \prod_{\mathbf{w} \in \mathcal{L}(\mathcal{T}_\mathbf{v})} g(\mathbf{w}), \tag{3.19}$$

where the sum is over all subtrees $\mathcal{T}_\mathbf{v}$ of $\mathbf{T}$ rooted at $\mathbf{v}$. Then, $G(\mathbf{v})$ can be computed recursively as follows:

$$G(\mathbf{v}) = \begin{cases} g(\mathbf{v}) & \text{if } \mathbf{v} \in \mathcal{L}(\mathbf{T}) \\ \frac{1}{2} g(\mathbf{v}) + \frac{1}{2} G(\mathbf{v}0) G(\mathbf{v}1) & \text{otherwise,} \end{cases}$$

for each node $\mathbf{v} \in \mathcal{N}(\mathbf{T})$.

Let us introduce

$$w_t(\mathcal{T}) = \pi_\mathbf{T}(\mathcal{T}) \exp(-\eta L_{t-1}(\mathcal{T})),$$

so that Equation (3.16) writes

$$\widehat{f}_t(x) = \frac{\sum_{\mathcal{T} \subset \mathbf{T}} w_t(\mathcal{T}) \widehat{y}_{\mathcal{T},t}(x)}{\sum_{\mathcal{T} \subset \mathbf{T}} w_t(\mathcal{T})}, \tag{3.20}$$

where the sums hold over all subtrees $\mathcal{T}$ of $\mathbf{T}$. We will show how to efficiently compute and update the numerator and denominator in Equation (3.20). Note that $w_t(\mathcal{T})$ may be written as

$$w_t(\mathcal{T}) = 2^{-\|\mathcal{T}\|} \prod_{\mathbf{v} \in \mathcal{L}(\mathcal{T})} w_{\mathbf{v},t} \tag{3.21}$$

with $w_{\mathbf{v},t} = \exp(-\eta L_{\mathbf{v},t-1})$, where $L_{\mathbf{v},t} := \sum_{s \leqslant t \,:\, X_t \in C_\mathbf{v}} \ell(\widehat{y}_{\mathbf{v},s}, y_s)$.

Denominator of Equation (3.20). For each node $\mathbf{v} \in \mathcal{N}(\mathbf{T})$ and every $t \geqslant 1$, denote

$$\overline{w}_{\mathbf{v},t} = \sum_{\mathcal{T}_\mathbf{v}} 2^{-\|\mathcal{T}_\mathbf{v}\|} \prod_{\mathbf{v}' \in \mathcal{L}(\mathcal{T}_\mathbf{v})} w_{\mathbf{v}',t} \tag{3.22}$$

so that (3.21) entails

$$\overline{w}_{\epsilon,t} = \sum_{\mathcal{T}} w_t(\mathcal{T}) \,. \tag{3.23}$$

Using Equation (3.22), the weights $\overline{w}_{\mathbf{v},t}$ can be computed recursively using Lemma 3.3. We denote by $\texttt{path}(x_t)$ the path from ϵ to $\mathbf{v}_\mathbf{T}(x_t)$ (from the root to the leaf containing x_t). Note that, by definition of $w_{\mathbf{v},t}$, if $\mathbf{v} \notin \texttt{path}(x_t)$ (namely $x_t \notin C_\mathbf{v}$), we have $w_{\mathbf{v},t+1} = w_{\mathbf{v},t}$. In addition, if $\mathbf{v} \notin \texttt{path}(x_t)$, so are all its descendants, so that (by induction, and using the above recursive formula) $\overline{w}_{\mathbf{v},t+1} = \overline{w}_{\mathbf{v},t}$. In other words, *only the nodes of* $\texttt{path}(x_t)$ *have updated weights.*

As a result, at each round $t \geqslant 1$, after seeing $(x_t, y_t) \in [0,1]^d \times \mathcal{Y}$, the weights $w_{\mathbf{v},t}$ and $\overline{w}_{\mathbf{v},t}$ are updated for $\mathbf{v} \in \texttt{path}(x_t)$ as follows (note that they are all initialized at $w_{\mathbf{v},1} = \overline{w}_{\mathbf{v},1} = 1$):

- for every $\mathbf{v} \notin \texttt{path}(x_t)$, $w_{\mathbf{v},t+1} = w_{\mathbf{v},t}$ and $\overline{w}_{\mathbf{v},t+1} = \overline{w}_{\mathbf{v},t}$;
- for every $\mathbf{v} \in \texttt{path}(x_t)$, $w_{\mathbf{v},t+1} = w_{\mathbf{v},t} \exp(-\eta \ell(\widehat{y}_{\mathbf{v},t}, y_t))$;
- for every $\mathbf{v} \in \texttt{path}(x_t)$, we have

$$\overline{w}_{\mathbf{v},t+1} = \begin{cases} w_{\mathbf{v},t+1} & \text{if } \mathbf{v} \in \mathcal{L}(\mathbf{T}) \quad (\text{namely } \mathbf{v} = \mathbf{v}_{\mathbf{T}}(x_t)), \\ \frac{1}{2} w_{\mathbf{v},t+1} + \frac{1}{2} \overline{w}_{\mathbf{v}0,t+1} \overline{w}_{\mathbf{v}1,t+1} & \text{otherwise.} \end{cases}$$

The weights $w_{\mathbf{v},t}, \overline{w}_{\mathbf{v},t}$ as well as the predictions $\widehat{y}_{\mathbf{v},t}$ are updated recursively in an "upwards" traversal of $\texttt{path}(x_t)$ in $\mathbf{T}$ (from $\mathbf{v}_{\mathbf{T}}(x_t)$ to ϵ), as indicated in Algorithm 3.

Note that when updating the structure of the tree, the weights $w_{\mathbf{v},t+1}, \overline{w}_{\mathbf{v},t+1}$ and predictions $\widehat{y}_{\mathbf{v},t+1}$ for the newly created nodes in $\mathcal{T}_{t+1} \setminus \mathcal{T}_t$ (which are offsprings of $\mathbf{v}_{\mathcal{T}_t}(x_t)$ created from the splits necessary to separate x_t from the other point $x_s \in C_{\mathbf{v}_{\mathcal{T}_t}(x_t)}$) can be set depending on whether these nodes contain x_s or x_t. This does not affect the values of $w_{\mathbf{v},t}$ and $\widehat{y}_{\mathbf{v},t}$ at other nodes, but only the values of $\overline{w}_{\mathbf{v},t}$ for $\mathbf{v} \in \texttt{path}(x_t)$ that are computed in the upwards recursion.

Numerator of Equation (3.20). The numerator of Equation (3.20) can be computed in the same fashion as the denominator. Let $w'_{\mathbf{v},t} = w_{\mathbf{v},t} \widehat{y}_{\mathbf{v},t}$ if $\mathbf{v} \in \texttt{path}(x)$, and $w'_{\mathbf{v},t} = w_{\mathbf{v},t}$ otherwise. Additionally, let

$$\widehat{w}_{\mathbf{v},t} = \sum_{\mathcal{T}_{\mathbf{v}}} 2^{-\|\mathcal{T}_{\mathbf{v}}\|} \prod_{\mathbf{v}' \in \mathcal{L}(\mathcal{T}_{\mathbf{v}})} w'_{\mathbf{v}',t} \,.$$

Note that we have

$$\widehat{w}_{\epsilon,t} = \sum_{\mathcal{T}} 2^{-\|\mathcal{T}\|} \prod_{\mathbf{v}' \in \mathcal{L}(\mathcal{T})} w'_{\mathbf{v}',t} = \sum_{\mathcal{T}} w_t(\mathcal{T}) \widehat{y}_{\mathbf{v}_{\mathcal{T}}(x),t} = \sum_{\mathcal{T}} w_t(\mathcal{T}) \widehat{y}_t(\mathcal{T}) \,. \tag{3.24}$$

Lemma 3.3 with $g(\mathbf{v}) = w'_{\mathbf{v},t}$ (so that $G(\mathbf{v}) = \widehat{w}_{\mathbf{v},t}$) enables to recursively compute $\widehat{w}_{\mathbf{v},t}$ from $w'_{\mathbf{v},t}$. First, note that $w'_{\mathbf{v},t} = w_{\mathbf{v},t}$ for every $\mathbf{v} \notin \texttt{path}(x)$. Since every descendant $\mathbf{v}'$ of $\mathbf{v}$ is also outside of $\texttt{path}(x)$, it follows by induction that $\widehat{w}_{\mathbf{v},t} = \overline{w}_{\mathbf{v},t}$ for every $\mathbf{v} \notin \texttt{path}(x)$. It then remains to show how to compute $\widehat{w}_{\mathbf{v},t}$ for $\mathbf{v} \in \texttt{path}(x)$. This is done again recursively, starting from the leaf $\mathbf{v}_{\mathbf{T}}(x)$ up to the root ϵ:

$$\widehat{w}_{\mathbf{v},t} = \begin{cases} w_{\mathbf{v},t} \widehat{y}_{\mathbf{v},t} & \text{if } \mathbf{v} = \mathbf{v}_{\mathbf{T}}(x) \\ \frac{1}{2} w_{\mathbf{v},t} \widehat{y}_{\mathbf{v},t} + \frac{1}{2} \overline{w}_{\mathbf{v}(1-a),t} \widehat{w}_{\mathbf{v}a,t} & \text{otherwise, where } a \in \{0,1\} \text{ is such that } \mathbf{v}a \in \texttt{path}(x) \end{cases}$$

Finally, we define

$$\widetilde{y}_{\mathbf{v},t}(x) = \frac{\widehat{w}_{\mathbf{v},t}}{\overline{w}_{\mathbf{v},t}}$$

for each node $\mathbf{v} \in \mathbf{T}$. It follows from Equations (3.20), (3.23) and (3.24) that $\widehat{f}_t(x) = \widetilde{y}_{\epsilon,t}(x)$. Additionally, the recursive expression for $\overline{w}_{\mathbf{v},t}$ and $\widehat{w}_{\mathbf{v},t}$ imply that $\widetilde{y}_{\mathbf{v},t}$ can be computed recursively as well, in the upwards traversal from $\mathbf{v}_{\mathbf{T}}(x)$ to ϵ: we set

$$\widetilde{y}_{\mathbf{v},t}(x) = \widehat{y}_{\mathbf{v},t}$$

for $\mathbf{v} = \mathbf{v}_{\mathbf{T}}(x)$, otherwise we set

$$\widetilde{y}_{\mathbf{v},t}(x) = \frac{1}{2}\frac{w_{\mathbf{v},t}}{\overline{w}_{\mathbf{v},t}}\widehat{y}_{\mathbf{v},t} + \frac{1}{2}\frac{\overline{w}_{\mathbf{v}a,t}\overline{w}_{\mathbf{v}(1-a),t}}{\overline{w}_{\mathbf{v},t}}\widetilde{y}_{\mathbf{v}a,t}(x) = \frac{1}{2}\frac{w_{\mathbf{v},t}}{\overline{w}_{\mathbf{v},t}}\widehat{y}_{\mathbf{v},t} + \left(1 - \frac{1}{2}\frac{w_{\mathbf{v},t}}{\overline{w}_{\mathbf{v},t}}\right)\widetilde{y}_{\mathbf{v}a,t}(x),$$

where $a \in \{0,1\}$ is such that $\mathbf{v}a \in \texttt{path}(x)$. The recursions constructed above are precisely the ones describing AMF in Algorithms 3 and 4 from Section 3.2.3, so that this concludes the proof of Proposition 3.1. □

3.7.2 Proofs of Lemma 3.1, Corollaries 3.1, 3.2, 3.3, Lemma 3.2, Theorems 3.1, 3.2 and Proposition 3.2

We start with some well-known lemmas that are used to bound the regret: Lemma 3.4 controls the regret with respect to each tree forecaster, while Lemmas 3.5 and 3.6 bound the regret of each tree forecaster with respect to the optimal labeling of its leaves.

Lemma 3.4 (Vovk, 1998)**.** *Let $\mathcal{E}$ be a countable set of experts and $\pi = (\pi_i)_{i\in\mathcal{E}}$ be a probability measure on $\mathcal{E}$. Assume that ℓ is η-*exp-concave*. For every $t \geqslant 1$, let $y_t \in \mathcal{Y}$, $\widehat{y}_{i,t} \in \widehat{\mathcal{Y}}$ be the prediction of expert $i \in \mathcal{E}$ and $L_{i,t} = \sum_{s=1}^{t} \ell(\widehat{y}_{i,s}, y_s)$ be its cumulative loss. Consider the predictions defined as*

$$\widehat{y}_t = \frac{\sum_{i\in\mathcal{E}} \pi_i \, e^{-\eta L_{i,t-1}} \widehat{y}_{i,t}}{\sum_{i\in\mathcal{E}} \pi_i \, e^{-\eta L_{i,t-1}}}. \tag{3.25}$$

Then, irrespective of the values of $y_t \in \mathcal{Y}$ and $\widehat{y}_{i,t} \in \widehat{\mathcal{Y}}$, we have the following regret bound

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \sum_{t=1}^{n} \ell(\widehat{y}_{i,t}, y_t) \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i} \tag{3.26}$$

for each $i \in \mathcal{E}$ and $n \geqslant 1$.

Lemma 3.5 (Tjalkens et al., 1993)**.** *Let ℓ be the logarithmic loss on the finite set $\mathcal{Y}$, and let $y_t \in \mathcal{Y}$ for every $t \geqslant 1$. The* Krichevsky-Trofimov (KT) *forecaster, which predicts*

$$\widehat{y}_t(y) = \frac{n_{t-1}(y) + 1/2}{(t-1) + |\mathcal{Y}|/2}, \tag{3.27}$$

with $n_{t-1}(y) = |\{1 \leqslant s \leqslant t-1 : y_s = y\}|$, satisfies the following regret bound with respect to the class $\mathcal{P}(\mathcal{Y})$ of constant experts (which always predict the same probability distribution on $\mathcal{Y}$):

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{p\in\mathcal{P}(\mathcal{Y})} \sum_{t=1}^{n} \ell(p, y_t) \leqslant \frac{|\mathcal{Y}| - 1}{2} \log(4n) \tag{3.28}$$

for each $n \geqslant 1$.

Lemma 3.6 (Cesa-Bianchi and Lugosi, 2006, p. 43)**.** *Consider the square loss $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ on $\mathcal{Y} = \widehat{\mathcal{Y}} = [-B, B]$, with $B > 0$. For every $t \geqslant 1$, let $y_t \in [-B, B]$. Consider the strategy defined by $\widehat{y}_1 = 0$, and for each $t \geqslant 2$,*

$$\widehat{y}_t = \frac{1}{t-1} \sum_{s=1}^{t-1} y_s \,. \tag{3.29}$$

The regret of this strategy with respect to the class of constant experts (which always predict some $b \in [-B, B]$) is upper bounded as follows:

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{b \in [-B,B]} \sum_{t=1}^{n} \ell(b, y_t) \leqslant 8B^2(1 + \log n) \tag{3.30}$$

for each $n \geqslant 1$.

Proof of Lemma 3.1. This follows from Proposition 3.1 and Lemma 3.4. □

Proof of Corollary 3.1. Since the logarithmic loss is 1-exp-concave, Lemma 3.1 implies

$$\sum_{t=1}^{n} \ell(\widehat{f}_t(x_t), y_t) - \sum_{t=1}^{n} \ell(\widehat{y}_{\mathcal{T},t}(x_t), y_t) \leqslant |\mathcal{T}| \log 2 \tag{3.31}$$

for every subtree $\mathcal{T}$. It now remains to bound the regret of the tree forecaster $\mathcal{T}$ with respect to the optimal labeling of its leaves. By Lemma 3.5, for every leaf $\mathbf{v}$ of $\mathcal{T}$,

$$\sum_{1 \leqslant t \leqslant n \,:\, x_t \in C_{\mathbf{v}}} \ell(\widehat{y}_{\mathcal{T},t}(x_t), y_t) - \inf_{p_{\mathbf{v}} \in \mathcal{P}(\mathcal{Y})} \sum_{1 \leqslant t \leqslant n \,:\, x_t \in C_{\mathbf{v}}} \ell(p_{\mathbf{v}}, y_t) \leqslant \frac{|\mathcal{Y}| - 1}{2} \log(4N_{\mathbf{v},n})$$

where $N_{\mathbf{v},n} = |\{1 \leqslant t \leqslant n : x_t \in C_{\mathbf{v}}\}|$ (assuming that $N_{\mathbf{v},n} \geqslant 1$). Summing the above inequality over the leaves $\mathbf{v}$ of $\mathcal{T}$ such that $N_{\mathbf{v},n} \geqslant 1$ yields

$$\sum_{t=1}^{n} \ell(\widehat{y}_{\mathcal{T},t}(x_t), y_t) - \inf_{g_{\mathcal{T}}} \sum_{t=1}^{n} \ell(g_{\mathcal{T}}(x_t), y_t) \leqslant \frac{|\mathcal{Y}| - 1}{2} \sum_{\mathbf{v} \in \mathcal{L}(\mathcal{T}) \,:\, N_{\mathbf{v},n} \geqslant 1} \log(4N_{\mathbf{v},n}) \tag{3.32}$$

where $g_{\mathcal{T}}$ is any function constant on the leaves of $\mathcal{T}$. Now, letting $L = |\{\mathbf{v} \in \mathcal{L}(\mathcal{T}) : N_{\mathbf{v},n} \geqslant 1\}| \leqslant |\mathcal{L}(\mathcal{T})| = \frac{|\mathcal{T}|+1}{2}$, we have by concavity of the log

$$\begin{aligned} \sum_{\mathbf{v} \in \mathcal{L}(\mathcal{T}) \,:\, N_{\mathbf{v},n} \geqslant 1} \log(4N_{\mathbf{v},n}) &\leqslant L \log\left(\frac{\sum_{\mathbf{v} \in \mathcal{L}(\mathcal{T}) \,:\, N_{\mathbf{v},n} \geqslant 1} 4N_{\mathbf{v},n}}{L} \right) \\ &= L \log\left(\frac{4n}{L}\right) \leqslant \frac{|\mathcal{T}| + 1}{2} \log(4n)\,. \end{aligned}$$

Plugging this in (3.32) and combining with Equation (3.31) leads to the desired bound (3.6). □

Proof of Corollary 3.2. The proof proceeds similarly to that of Corollary 3.1, by combining Lemmas 3.1 and 3.6 and using the fact that the square loss is $\eta = 1/(8B^2)$-exp-concave on $[-B, B]$. □

Proof of Corollary 3.3. First, we reason conditionally on the Mondrian process Π. By applying Corollary 3.2 to $\mathcal{T} = \Pi_\lambda$, we obtain, since the number of nodes of Π_λ is $2|\mathcal{L}(\Pi_\lambda)| - 1$:

$$\sum_{t=1}^{n} \ell(\widehat{f}_t(x_t), y_t) - \inf_{g} \sum_{t=1}^{n} \ell(g(x_t), y_t) \leqslant 8B^2 |\mathcal{L}(\Pi_\lambda)| \log n\,, \tag{3.33}$$

where the infimum spans over all functions $g : [0,1]^d \to \widehat{\mathcal{Y}}$ which are constant on the cells of Π_λ. Corollary 3.3 follows by taking the expectation over Π and using the fact that $\Pi_\lambda \sim \mathsf{MP}(\lambda)$ implies $\mathbb{E}[|\mathcal{L}(\Pi_\lambda)|] = (1 + \lambda)^d$ (by Proposition 2.2 in Chapter 2). □

Proof of Lemma 3.2. For every $t = 1, \dots, n$, $\widehat{f}_t$ is $\mathcal{F}_{t-1} := \sigma(x_1, y_1, \dots, x_{t-1}, y_{t-1})$-measurable and since (x_t, y_t) is independent of $\mathcal{F}_t$:

$$\mathbb{E}[\ell(\widehat{f}_t(x_t), y_t)] = \mathbb{E}[\mathbb{E}[\ell(\widehat{f}_t(x_t), y_t)|\mathcal{F}_{t-1}]] = \mathbb{E}[R(\widehat{f}_t)]\,,$$

so that, for every $g \in \mathcal{G}$,

$$\frac{1}{n}\mathbb{E}\Big[\sum_{t=1}^{n}\big(\ell(\widehat{f}_t(x_t), y_t) - \ell(g(x_t), y_t)\big)\Big] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[R(\widehat{f}_t)] - R(g) = \mathbb{E}[R(\widetilde{f}_n)] - R(g)\,. \qquad \square$$

Proof of Theorem 3.1. This is a direct consequence of Lemma 3.2 and Corollary 3.2. $\square$

Proof of Theorem 3.2. Recall that the sequence $(x_1, y_1), \dots, (x_n, y_n)$ is i.i.d and distributed as a generic pair $(x, y) \in [0,1]^d \times \mathcal{Y}$. Since $f^*(\cdot) = \mathbb{E}[y|x = \cdot]$, we have

$$R(f) = \mathbb{E}[(f(x) - f^*(x))^2] + R(f^*) \tag{3.34}$$

for every function $f : [0,1]^d \to \mathbf{R}$. Now, let $\lambda > 0$ be arbitrary. Consider the estimator $\widetilde{f}_n$ defined in Lemma 3.2, and the function h^*_λ constant on the cells of a random partition $\Pi_\lambda \sim \mathsf{MP}(\lambda)$, with optimal predictions on the leaves given by $h^*_\lambda(u) = \mathbb{E}[y|x \in C_{\mathbf{v}}]$ for $u \in C_{\mathbf{v}}$, for every leaf $\mathbf{v}$ of Π_λ. Since $R(\widetilde{f}_n) - R(f^*) = R(\widetilde{f}_n) - R(h^*_\lambda) + R(h^*_\lambda) - R(f^*)$, Equation (3.34) gives, after taking the expectation over the random sampling of the Mondrian process Π_λ,

$$\mathbb{E}[(\widetilde{f}_n(x) - f^*(x))^2] = \mathbb{E}[R(\widetilde{f}_n)] - \mathbb{E}[R(h^*_\lambda)] + \mathbb{E}[(h^*_\lambda(x) - f^*(x))^2]\,. \tag{3.35}$$

Let $D_\lambda(u)$ denote the diameter of the cell $C_\lambda(u)$ of $u \in [0,1]^d$ in the Mondrian partition Π_λ used to define h^*_λ. Assume that f^* is β-Holder with constant $L > 0$, namely $|f^*(u) - f^*(v)| \leqslant L|u - v|^\beta$ for any $u, v \in [0,1]^d$. Since $h^*_\lambda(u) = \mathbb{E}[f^*(x)|x \in C_\lambda(u)]$, we have $|h^*_\lambda(u) - f^*(u)| \leqslant LD_\lambda(u)^\beta$, so that

$$\mathbb{E}[(h^*_\lambda(x) - f^*(x))^2] \leqslant L^2\mathbb{E}[D_\lambda(x)^{2\beta}]\,. \tag{3.36}$$

Now, since $u \mapsto u^\beta$ is concave,

$$\mathbb{E}[D_\lambda(x)^{2\beta}] \leqslant \mathbb{E}[D_\lambda(x)^2]^\beta \leqslant \Big(\frac{4d}{\lambda^2}\Big)^\beta \tag{3.37}$$

where the last inequality comes from Corollary 2.1 in Chapter 2. Integrating with the distribution of x and using (3.36) gives $\mathbb{E}[(h^*_\lambda(x) - f^*(x))^2] \leqslant (4d)^\beta L^2/\lambda^{2\beta}$. In addition, Theorem 3.1 gives $\mathbb{E}[R(\widetilde{f}_n)] - \mathbb{E}[R(h^*_\lambda)] \leqslant 8B^2(1+\lambda)^d(\log n)/n$. Combining these inequalities with (3.35) leads to

$$\mathbb{E}[(\widetilde{f}_n(x) - f^*(x))^2] \leqslant \frac{(4d)^\beta L^2}{\lambda^{2\beta}} + \frac{8B^2(1+\lambda)^d \log n}{n}\,. \tag{3.38}$$

Note that the bound (3.38) holds for every value of $\lambda > 0$. In particular, for $\lambda \asymp (n/\log n)^{1/(d+2\beta)}$, it yields the $O\big((\log(n)/n)^{2\beta/(d+2\beta)}\big)$ bound on the estimation risk of Theorem 3.2. $\square$

Proof of Proposition 3.2. First, we reason conditionally on the realization of an infinite Mondrian partition Π, by considering the randomness with respect to the sampling of the feature points $x_1, \dots, x_n, x$. For every depth $j \geqslant 0$, denote by N_j the number of points among $x_1, \dots, x_n$ that belong to the cell of depth j of the unrestricted Mondrian partition containing x, and $\mathbf{v}_j \in \{0,1\}^j$ the corresponding node. In addition, for every $\mathbf{v} \in \{0,1\}^*$, denote

$p_{\mathbf{v}} = \mathbb{P}(x \in C_{\mathbf{v}0} | x \in C_{\mathbf{v}})$. In addition, for $j \geqslant 0$, since conditionally on $C_{\mathbf{v}_j}, N_j$, the points x and $\{x_i : x_i \in C_{\mathbf{v}_j}\}$ are distributed i.i.d. following the conditional distribution of x given $\{x \in C_{\mathbf{v}_j}\}$:

$$\begin{aligned}\mathbb{E}[N_{j+1} | C_{\mathbf{v}_j}, N_j, \Pi] &= \mathbb{P}(\mathbf{v}_{j+1} = \mathbf{v}_j 0 | \mathbf{v}_j) \times N_j \mathbb{P}(x_1 \in C_{\mathbf{v}_j 0} | x_1 \in C_{\mathbf{v}_j}) \\ &\quad + \mathbb{P}(\mathbf{v}_{j+1} = \mathbf{v}_j 1 | \mathbf{v}_j) \times N_j \mathbb{P}(x_1 \in C_{\mathbf{v}_j 1} | x_1 \in C_{\mathbf{v}_j}) \\ &= N_j \big(p_{\mathbf{v}_j}^2 + (1 - p_{\mathbf{v}_j})^2\big) \\ &= N_j \big(1 - 2p_{\mathbf{v}_j}(1 - p_{\mathbf{v}_j})\big) \,. \end{aligned} \tag{3.39}$$

Now, note that $p_{\mathbf{v}_j}(1 - p_{\mathbf{v}_j})$ is determined by $C_{\mathbf{v}_j}$ and its split in Π, while N_j is determined by $C_{\mathbf{v}_j}$ and $x_1, \ldots, x_n$. Now, let U_j be the ratio of the volume of $C_{\mathbf{v}_j 0}$ by that of $C_{\mathbf{v}_j}$; by construction of the Mondrian process, $U_j \sim \mathcal{U}([0,1])$ conditionally on $C_{\mathbf{v}_j}$. In addition, the assumption (3.15) implies (by integrating over the coordinate of the split, fixing the other coordinates) that $p_{\mathbf{v}_j} \geqslant M^{-1} U_j$, $1 - p_{\mathbf{v}_j} \geqslant M^{-1}(1 - U_j)$. It follows that

$$p_{\mathbf{v}_j}(1 - p_{\mathbf{v}_j}) \geqslant \frac{1}{2}\{p_{\mathbf{v}_j} \wedge (1 - p_{\mathbf{v}_j})\} \geqslant \frac{1}{2M}\{U_j \wedge (1 - U_j)\}$$

so that

$$\mathbb{E}[p_{\mathbf{v}_j}(1 - p_{\mathbf{v}_j}) | C_{\mathbf{v}_j}] \geqslant \frac{1}{2M}\mathbb{E}[U_j \wedge (1 - U_j) | C_{\mathbf{v}_j}] = \frac{1}{4M} \,.$$

Using the fact that N_j and $p_{\mathbf{v}_j}$ are independent conditionally on $C_{\mathbf{v}_j}$, it follows from (3.39) that

$$\mathbb{E}[N_{j+1} | C_{\mathbf{v}_j}] = \mathbb{E}[N_j | C_{\mathbf{v}_j}]\big(1 - 2\mathbb{E}[p_{\mathbf{v}_j}(1 - p_{\mathbf{v}_j})]\big) \leqslant \Big(1 - \frac{1}{2M}\Big)\mathbb{E}[N_j | C_{\mathbf{v}_j}].$$

By induction on $k \geqslant 0$, using the fact that by definition $N_0 = n$,

$$\mathbb{E}[N_k] \leqslant n\Big(1 - \frac{1}{2M}\Big)^k. \tag{3.40}$$

Now, note that if $N_k = 0$, then the depth $D_n^\Pi(x)$ of x in the Mondrian partition Π restricted to $x_1, \ldots, x_n, x$ is at most k. Thus, inequality (3.40) implies:

$$\begin{aligned}\mathbb{E}[D_n^\Pi(x)] &= \sum_{k \geqslant 1} \mathbb{P}(D_n^\Pi(x) \geqslant k) \leqslant \sum_{k \geqslant 1} \mathbb{P}(N_k \geqslant 1) \\ &\leqslant \sum_{k \geqslant 1} \mathbb{E}[N_k] \wedge 1 \leqslant \sum_{k \geqslant 1} \left\{ n\Big(1 - \frac{1}{2M}\Big)^k \right\} \wedge 1 \end{aligned} \tag{3.41}$$

Now, let k_0 be the smallest $k \geqslant 1$ such that $n(1 - 1/(2M))^{k_0} \leqslant 1$. We have

$$k_0 = \Big\lceil \frac{\log n}{\log\{(2M)/(2M-1)\}} \Big\rceil,$$

so that $k_0 - 1 \leqslant \log(n)/\log\{(2M)/(2M-1)\}$. Hence, inequality (3.41) becomes:

$$\begin{aligned}\mathbb{E}[D_n^\Pi(x)] &\leqslant (k_0 - 1) + \sum_{k \geqslant 0} \underbrace{n\Big(1 - \frac{1}{2M}\Big)^{k_0}}_{\leqslant 1} \Big(1 - \frac{1}{2M}\Big)^k \\ &\leqslant \frac{\log n}{\log[(2M)/(2M-1)]} + 2M \end{aligned}$$

which establishes Proposition 3.2. □

Part II

Prediction with expert advice

Chapter 4

On the optimality of the Hedge algorithm in the stochastic regime

Abstract. In this chapter, we study the behavior of the Hedge algorithm in the online stochastic setting. We prove that anytime Hedge with decreasing learning rate, which is one of the simplest algorithm for the problem of prediction with expert advice, is remarkably both worst-case optimal and adaptive to the easier stochastic and adversarial with a gap problems. This shows that, in spite of its small, non-adaptive learning rate, Hedge possesses the same optimal regret guarantee in the stochastic case as recently introduced adaptive algorithms. Moreover, our analysis exhibits qualitative differences with other versions of the Hedge algorithm, such as the fixed-horizon variant (with constant learning rate) and the one based on the so-called "doubling trick", both of which fail to adapt to the easier stochastic setting. Finally, we determine the intrinsic limitations of anytime Hedge in the stochastic case, and discuss the improvements provided by more adaptive algorithms.

Contents

4.1 Introduction

The standard setting of *prediction with expert advice* (Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006) aims to provide sound strategies for sequential prediction that combine the forecasts from different sources. More precisely, in the so-called *Hedge problem* (Freund and Schapire, 1997), at each round the learner has to output a probability distribution on a finite set of *experts* $\{1, \dots, M\}$; the losses of the experts are then revealed, and the learner incurs the expected loss from its chosen probability distribution. The goal is then to control the *regret*, defined as the difference between the cumulative

loss of the learner and that of the best expert (with smallest loss). This online prediction problem is typically considered in the *individual sequences* framework, where the losses may be arbitrary and in fact set by an adversary that seeks to maximize the regret. This leads to regret bounds that hold under virtually no assumption (Cesa-Bianchi and Lugosi, 2006).

In this setting, arguably the simplest and most standard strategy is the *Hedge algorithm* (Freund and Schapire, 1997), also called the *exponentially weighted averaged forecaster* (Cesa-Bianchi and Lugosi, 2006). This algorithm depends on a time-varying parameter η_t called the *learning rate*, which quantifies by how much the algorithm departs from its initial probability distribution to put more weight on the currently leading experts. Given a known finite time horizon T, the standard tuning of the learning rate is fixed and given by $\eta_t = \eta \propto \sqrt{\log(M)/T}$, which guarantees an optimal worst-case regret of order $O(\sqrt{T \log M})$. Alternatively, when T is unknown, one can set $\eta_t \propto \sqrt{\log(M)/t}$ at round t, which leads to an *anytime* $O(\sqrt{T \log M})$ regret bound valid for all $T \geqslant 1$.

While worst-case regret bounds are robust and always valid, they turn out to be overly pessimistic in some situations. A recent line of research (Cesa-Bianchi et al., 2007; de Rooij et al., 2014; Gaillard et al., 2014; Koolen et al., 2014; Sani et al., 2014; Koolen and van Erven, 2015; Luo and Schapire, 2015) designs algorithms that combine $O(\sqrt{T \log M})$ worst-case regret guarantees with an improved regret on easier instances of the problem. An interesting example of such an easier instance is the stochastic problem, where it is assumed that the losses are stochastic and that at each round the expected loss of a "best" expert is smaller than those of the other experts by some gap Δ. Such algorithms rely either on a more careful, data-dependent tuning of the learning rate η_t (Cesa-Bianchi et al., 2007; de Rooij et al., 2014; Koolen et al., 2014; Gaillard et al., 2014), or on more sophisticated strategies (Koolen and van Erven, 2015; Luo and Schapire, 2015). As shown by Gaillard et al. (2014) (see also Koolen et al. 2016), one particular type of adaptive regret bounds (so-called *second-order bounds*) implies at the same time a $O(\sqrt{T \log M})$ worst-case bound and a better *constant* $O(\log(M)/\Delta)$ bound in the stochastic problem with gap Δ. Arguably starting with the early work on second-order bounds (Cesa-Bianchi et al., 2007), the design of online learning algorithms that combine robust worst-case guarantees with improved performance on easier instances has been an active research goal in recent years (de Rooij et al., 2014; Gaillard et al., 2014; Koolen et al., 2014; Sani et al., 2014). However, to the best of our knowledge, existing work on the Hedge problem has focused on developing new adaptive algorithms rather than on analyzing the behavior of "conservative" algorithms in favorable scenarios. Owing to the fact that the standard Hedge algorithm is designed for — and analyzed in — the adversarial setting (Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), and that its parameters are not tuned adaptively to obtain better bounds in easier instances, it may be considered as overly conservative and not adapted to stochastic environments.

Our contribution. This work fills a gap in the existing literature by providing an analysis of the standard Hedge algorithm in the stochastic setting. We show that the anytime Hedge algorithm with default learning rate $\eta_t \propto \sqrt{\log(M)/t}$ actually *adapts* to the stochastic setting, in which it achieves an optimal *constant* $O(\log(M)/\Delta)$ regret bound *without any dedicated tuning* for the easier instance, which might be surprising at first sight. This contrasts with previous works, which require the construction of new adaptive (and more involved) algorithms. Remarkably, this property is *not* shared by the variant of Hedge for a known fixed-horizon T with constant learning rate $\eta \propto \sqrt{\log(M)/T}$, since it suffers a $\Theta(\sqrt{T \log M})$ regret even in

easier instances. This exhibits a strong difference between the performances of the anytime and the fixed-horizon variants of the Hedge algorithm.

Given the aforementioned adaptivity of Decreasing Hedge, one may wonder whether there is in fact any benefit in using more sophisticated algorithms in the stochastic regime. We answer this question affirmatively, by considering a more refined measure of complexity of a stochastic instance than the gap Δ. Specifically, we show that Decreasing Hedge does not admit improved regret under Bernstein conditions, which are standard low-noise conditions from statistical learning (Mammen and Tsybakov, 1999; Tsybakov, 2004; Bartlett and Mendelson, 2006). By contrast, it was shown by Koolen et al. (2016) that algorithms which satisfy some adaptive adversarial regret bound achieve improved regret under Bernstein conditions. Finally, we characterize the behavior of Decreasing Hedge in the stochastic regime, by showing that its eventual regret on *any* stochastic instance is governed by the gap Δ.

Related work. In the bandit setting, where the feedback only consists of the loss of the selected action, there has also been some interest in "best-of-both-worlds" algorithms that combine optimal $O(\sqrt{MT})$ worst-case regret in the adversarial regime with improved $O(M \log T)$ regret (up to logarithmic factors) in the stochastic case (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016). In particular, Seldin and Slivkins (2014); Seldin and Lugosi (2017) showed that by augmenting the standard EXP3 algorithm for the adversarial regime (an analogue of Hedge with $\Theta(1/\sqrt{t})$ learning rate) with a special-purpose gap detection mechanism, one can achieve poly-logarithmic regret in the stochastic case. This result is strengthened in some recent follow-up work (Zimmert and Seldin, 2019; Zimmert et al., 2019), which appeared since the completion of the first version of the present work, that obtains optimal regret in the stochastic and adversarial regimes through a variant of the Follow-The-Regularized-Leader (FTRL) algorithm with $\Theta(1/\sqrt{t})$ learning rate and a proper regularizer choice. This result can be seen as an analogue in the bandit case of our upper bound for Decreasing Hedge. On the other hand, in the bandit setting, the hardness of an instance is essentially characterized by the gap Δ (Bubeck and Cesa-Bianchi, 2012); in particular, the Bernstein condition, which depends on the correlations between the losses of the experts, cannot be exploited under bandit feedback, where one only observes one arm at each round. Hence, it appears that the negative part of our results (on the limitations of Hedge) does not have an analogue in the bandit case.

A similar adaptivity result for FTRL with decreasing $\Theta(1/\sqrt{t})$ learning rate has been observed in a different context by Huang et al. (2017). Specifically, it is shown that, in the case of online linear optimization on a Euclidean ball, FTRL with squared norm regularizer and learning rate $\Theta(1/\sqrt{t})$ achieves $O(\log T)$ regret when the loss vectors are i.i.d. This result is an analogue of our upper bound for Hedge, since this algorithm corresponds to FTRL on the simplex with entropic regularizer (Cesa-Bianchi and Lugosi, 2006; Hazan, 2016). On the other hand, the simplex lacks the curvature of the Euclidean ball, which is important to achieve small regret; here, the improved regret is ensured by a condition on the distribution, namely the existence of a gap Δ. Our lower bound for Hedge shows that this condition is necessary, thereby characterizing the long-term regret of FTRL on the simplex with entropic regularizer. In the case of the Euclidean ball with squared norm regularizer, the norm of the expected loss vector appears to play a similar role, as shown by the upper bound from Huang et al. (2017).

Outline. We define the setting of prediction with expert advice and the Hedge algorithm in Section 4.2, and we recall herein its standard worst-case regret bound. In Section 4.3, we consider the behavior of the Hedge algorithm on easier instances, namely the stochastic setting with a gap Δ on the best expert. Under an i.i.d assumption on the sequence of losses, we provide in Theorem 4.1 an upper bound on the regret of order $(\log M)/\Delta$ for Decreasing Hedge. In Proposition 4.2, we prove that the rate $(\log M)/\Delta$ cannot be improved in this setting. In Theorem 4.2 and Corollary 4.1, we extend the regret guarantees to the adversarial with a gap setting, where a leading expert linearly outperforms the others. These results stand for any Hedge algorithm which is worst-case optimal and with any learning rate which is larger than the one of Decreasing Hedge, namely $O(\sqrt{\log M/t})$. In Proposition 4.3, we prove the sub-optimality of the fixed-horizon Hedge algorithm, and of another version of Hedge based on the so-called "doubling trick". In Section 4.4, we discuss the advantages of adaptive Hedge algorithms, and explain what the limitations of Decreasing Hedge are compared to such versions. We include numerical illustrations of our theoretical findings in Section 4.5, conclude in Section 4.6 and provide the proofs in Section 4.7.

4.2 The expert problem and the Hedge algorithm

In the Hedge setting, also called *decision-theoretic online learning* (Freund and Schapire, 1997), the learner and its adversary (the Environment) sequentially compete on the following game: at each round $t \geqslant 1$,

1. the Learner chooses a probability vector $\boldsymbol{v}_t = (v_{i,t})_{1 \leqslant i \leqslant M}$ on the M experts $1, \dots, M$;
2. the Environment picks a bounded loss vector $\boldsymbol{\ell}_t = (\ell_{i,t})_{1 \leqslant i \leqslant M} \in [0,1]^M$, where $\ell_{i,t}$ is the loss of expert i at round t, while the Learner suffers loss $\widehat{\ell}_t = \boldsymbol{v}_t^\top \boldsymbol{\ell}_t$.

The goal of the Learner is to control its *regret*

$$R_T = \sum_{t=1}^{T} \widehat{\ell}_t - \min_{1 \leqslant i \leqslant M} \sum_{t=1}^{T} \ell_{i,t} \tag{4.1}$$

for every $T \geqslant 1$, irrespective of the sequence of loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ chosen by the Environment. One of the most standard algorithms for this setting is the *Hedge* algorithm. The Hedge algorithm, also called the exponentially weighted averaged forecaster, uses the vector of probabilities $\boldsymbol{v}_t = (v_{i,t})_{1 \leqslant i \leqslant M}$ given by

$$v_{i,t} = \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^{M} e^{-\eta_t L_{j,t-1}}} \tag{4.2}$$

at each $t \geqslant 1$, where $L_{i,T} = \sum_{t=1}^{T} \ell_{i,t}$ denotes the cumulative loss of expert i for every $T \geqslant 1$. Let us also denote $\widehat{L}_T := \sum_{t=1}^{T} \widehat{\ell}_t$ and $R_{i,T} = \widehat{L}_T - L_{i,T}$ the regret with respect to expert i. We consider in this chapter the following variants of Hedge, where $c_0 > 0$ is a constant.

Decreasing Hedge (Auer et al., 2002). This is Hedge with the sequence of learning rates $\eta_t = c_0 \sqrt{\log(M)/t}$.

Constant Hedge (Littlestone and Warmuth, 1994). Given a finite time horizon $T \geqslant 1$, this is Hedge with constant learning rate $\eta_t = c_0 \sqrt{\log(M)/T}$.

Hedge with doubling trick (Cesa-Bianchi et al., 1997; Cesa-Bianchi and Lugosi, 2006). This variant of Hedge uses a constant learning rate on geometrically increasing intervals, restarting the algorithm at the beginning of each interval. Namely, it uses

$$v_{i,t} = \frac{\exp(-\eta_t \sum_{s=T_k}^{t-1} \ell_{i,s})}{\sum_{j=1}^{M} \exp(-\eta_t \sum_{s=T_k}^{t-1} \ell_{j,s})}, \tag{4.3}$$

with $T_l = 2^l$ for $l \geqslant 0$, $k \in \mathbf{N}$ such that $T_k \leqslant t < T_{k+1}$ and $\eta_t = c_0\sqrt{\log(M)/T_k}$.

Let us recall the following standard regret bound for the Hedge algorithm from Chernov and Zhdanov (2010).

Proposition 4.1. *Let $\eta_1, \eta_2, \dots$ be a decreasing sequence of learning rates. The Hedge algorithm (4.2) satisfies the following regret bound:*

$$R_T \leqslant \frac{1}{\eta_T} \log M + \frac{1}{8} \sum_{t=1}^{T} \eta_t \,. \tag{4.4}$$

In particular, the choice $\eta_t = 2\sqrt{\log(M)/t}$ yields a regret bound of $\sqrt{T \log M}$ for every $T \geqslant 1$.

Note that the regret bound stated in Equation (4.4) holds for every sequence of losses $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$, which makes it valid under no assumption (aside from the boundedness of the losses). The worst-case regret bound in $O(\sqrt{T \log M})$ is achieved by Decreasing Hedge, Hedge with doubling trick and Constant Hedge (whenever T is known in advance). The $O(\sqrt{T \log M})$ rate cannot be improved either by Hedge or any other algorithm: it is known to be the minimax optimal regret (Cesa-Bianchi and Lugosi, 2006). Contrary to Constant Hedge, Decreasing Hedge is anytime, in the sense that it achieves the $O(\sqrt{T \log M})$ regret bound simultaneously for each $T \geqslant 1$. We note that this worst-case regret analysis fails to exhibit any difference between these three algorithms.

In many cases, this $\sqrt{T}$ regret bound is pessimistic, and more "aggressive" strategies (such as the follow-the-leader algorithm, which plays at each round the uniform distribution on the experts with smallest loss, Cesa-Bianchi and Lugosi, 2006) may achieve constant regret in easier instances, even though they lack regret guarantees in the adversarial regime. We show in Section 4.3 below that Decreasing Hedge is actually better than both Constant Hedge and Hedge with doubling trick in some easier instance of the problem (including in the stochastic setting). This entails that Decreasing Hedge is actually able to adapt, without any modification, to the easiness of the problem considered.

4.3 Regret of Hedge variants on easy instances

In this section, we depart from the worst-case regret analysis and study the regret of the considered variants of the Hedge algorithm on easier instances of the prediction with expert advice problem.

4.3.1 Optimal regret for Decreasing Hedge in the stochastic regime

We examine the behavior of Decreasing Hedge in the stochastic regime, where the losses are the realization of some (unknown) stochastic process. More precisely, we consider the standard

i.i.d. case, where the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ are i.i.d. (independence holds over rounds, but not necessarily across experts). In this setting, the regret can be much smaller than the worst-case $\sqrt{T \log M}$ regret, since the best expert (with smallest expected loss) will dominate the rest after some time. Following Gaillard et al. (2014); Luo and Schapire (2015), the easiness parameter we consider in this case, which governs the time needed for the best expert to have the smallest cumulative loss and hence the incurred regret, is the sub-optimality gap $\Delta = \min_{i \neq i^*} \mathbb{E}[\ell_{i,t} - \ell_{i^*,t}]$, where $i^* = \arg\min_i \mathbb{E}[\ell_{i,t}]$.

We show below that, despite the fact that Decreasing Hedge is designed for the worst-case setting described in Section 4.2, it is able to adapt to the easier problem considered here, Indeed, Theorem 4.1 shows that Decreasing Hedge achieves a *constant*, and in fact *optimal* (by Proposition 4.2 below) regret bound in this setting, in spite of its "conservative" learning rate.

With the exception of the high-probability bound of Corollary 4.1, the upper and lower bounds in the stochastic case are stated for the *pseudo-regret* $\mathcal{R}_T = \mathbb{E}[R_{i^*,T}]$ (similar bounds hold for the the expected regret $\mathbb{E}[R_T]$, since $\mathcal{R}_T \leqslant \mathbb{E}[R_T]$ and by Remark 4.3 in Section 4.7.1).

Theorem 4.1. *Let $M \geqslant 3$. Assume that the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ are i.i.d. random variables, where $\boldsymbol{\ell}_t = (\ell_{i,t})_{1 \leqslant i \leqslant M}$. Also, assume that there exists $i^* \in \{1, \dots, M\}$ and $\Delta > 0$ such that*

$$\mathbb{E}[\ell_{i,t} - \ell_{i^*,t}] \geqslant \Delta \tag{4.5}$$

for every $i \neq i^*$. Then, the Decreasing Hedge algorithm with learning rate $\eta_t = 2\sqrt{(\log M)/t}$ achieves the following pseudo-regret bound: for every $T \geqslant 1$,

$$\mathcal{R}_T \leqslant \frac{4 \log M + 25}{\Delta}\,. \tag{4.6}$$

The proof of Theorem 4.1 is given in Section 4.7.1. Theorem 4.1 proves that, in the stochastic setting with a gap Δ, the Decreasing Hedge algorithm achieves a regret $O(\log(M)/\Delta)$, without any prior knowledge of Δ. This matches the guarantees of adaptive Hedge algorithms which are explicitly designed to adapt to easier instances (Gaillard et al., 2014; Luo and Schapire, 2015). This result may seem surprising at first: indeed, adaptive exponential weights algorithms that combine optimal regret in the adversarial setting and constant regret in easier scenarios, such as Hedge with a second-order tuning (Cesa-Bianchi et al., 2007) or AdaHedge (de Rooij et al., 2014), typically use a data-dependent learning rate η_t that adapts to the properties of the losses. While the learning rate η_t chosen by these algorithms may be as low as the worst-case tuning $\eta_t \propto \sqrt{\log(M)/t}$, in the stochastic case those algorithms will use larger, lower-bounded learning rates to ensure constant regret. As Theorem 4.1 above shows, it turns out that the data-independent, "safe" learning rates $\eta_t \propto \sqrt{\log(M)/t}$ used by "vanilla" Decreasing Hedge are still large enough to adapt to the stochastic case.

Idea of the proof. The idea of the proof of Theorem 4.1 is to divide time in two phases: a short initial phase $[\![1, t_1]\!]$, where $t_1 = O(\frac{\log M}{\Delta^2})$, and a second phase $[\![t_1, T]\!]$. The initial phase is dominated by noise, and regret during this period is bounded through the worst-case regret bound of Proposition 4.1, which gives a regret of $O(\sqrt{t_1 \log M}) = O(\frac{\log M}{\Delta})$. In the second phase, the best expert dominates the rest, and the weights concentrate on this best expert fast enough that the total regret incurred is small. The control of the regret in the second phase relies on the critical fact that, if η_t is at least as large as $\sqrt{(\log M)/t}$, then the following two things occur simultaneously at $t_1 \asymp \frac{\log M}{\Delta^2}$, namely at the beginning of the late phase:

1. with high probability, the best expert i^* dominates all the others linearly: for every $i \neq i^*$ and $t \geqslant t_1$, $L_{i,t} - L_{i^*,t} \geqslant \frac{\Delta t}{2}$;
2. the total weight of all suboptimal experts is controlled: $\sum_{i \neq i^*} v_{i,t_1} \leqslant \frac{1}{2}$. If $\eta_t \geqslant \sqrt{(\log M)/t}$ and the first condition holds, this amounts to $M \exp(-\frac{\Delta}{2}\sqrt{t \log M}) \leqslant \frac{1}{2}$, namely $t_1 \gtrsim \frac{\log M}{\Delta^2}$.

In other words, the learning rate $\eta_t \asymp \sqrt{(\log M)/t}$ ensures that the total weight of suboptimal experts starts vanishing at about the same time as when the best expert starts to dominate the others with a large probability (and remarkably, this property holds for every value of the sub-optimality gap Δ). Finally, the upper bound on the regret in the second phase rests on the two conditions above, together with the bound $\sum_{t \geqslant 1} e^{-c\sqrt{t}} = O(\frac{1}{c^2})$ for $c > 0$.

Remark 4.1. The fact that $\sum_{t \geqslant 1} e^{-c\sqrt{t}} = O(1/c^2)$ is also used in the analysis of the EXP3++ bandit algorithm (Seldin and Slivkins, 2014, Lemma 10). In the expert setting considered here, summing the contribution of all experts (which suffices in the bandit setting to obtain the correct order of regret) would yield a significantly suboptimal $O(M/\Delta)$ regret bound, with a linear dependence on the number of experts M. In our case, the decomposition of the regret in two phases, which is explained above, removes the linear dependence on M and allows to obtain the optimal rate $(\log M)/\Delta$.

We complement Theorem 4.1 by showing that the $O((\log M)/\Delta)$ regret under the gap condition cannot be improved, in the sense that its dependence on both M and Δ is optimal.

Proposition 4.2. *Let $\Delta \in (0, \frac{1}{4})$, $M \geqslant 4$ and $T \geqslant (\log M)/(16\Delta^2)$. For any algorithm for the Hedge setting, there exists an i.i.d. distribution over sequences of losses $(\boldsymbol{\ell}_t)_{t \geqslant 1}$ such that:*

- *there exists $i^* \in \{1, \dots, M\}$ such that, for any $i \neq i^*$, $\mathbb{E}[\ell_{i,t} - \ell_{i^*,t}] \geqslant \Delta$;*
- *the pseudo-regret of the algorithm satisfies:*

$$\mathcal{R}_T \geqslant \frac{\log M}{256\Delta}\,. \tag{4.7}$$

The proof of Proposition 4.2 is given in Section 4.7.2. Proposition 4.2 generalizes the well-known minimax lower bound of $\Theta(\sqrt{T \log M})$, which is recovered by taking $\Delta \asymp \sqrt{(\log M)/T}$.

4.3.2 Small regret for Decreasing Hedge in the adversarial with a gap problem

In this section, we extend the regret guarantee of Decreasing Hedge in the stochastic setting (Theorem 4.1), by showing that it holds for more general algorithms and under more general assumptions. Specifically, we consider an "adversarial with a gap" regime, similar to the one introduced by Seldin and Slivkins (2014) in the bandit case, where the leading expert linearly outperforms the others after some time. As Theorem 4.2 shows, essentially the same regret guarantee can be obtained in this case, up to an additional $\log(\Delta^{-1})/\Delta$ term. Theorem 4.2 also applies to any Hedge algorithm whose (possibly data-dependent) learning rate η_t is at least as large as that of Decreasing Hedge, and which satisfies a $O(\sqrt{T \log M})$ worst-case regret bound; this includes algorithms with *anytime* first and second-order tuning of the learning rate (Auer et al., 2002; Cesa-Bianchi et al., 2007; de Rooij et al., 2014). In what follows, we will assume $M \geqslant 3$ for convenience; similar results holds for $M = 2$.

Theorem 4.2. *Let $M \geqslant 3$. Assume that there exists $\tau_0 \geqslant 1$, $\Delta \in (0,1)$ and $i^* \in \{1,\dots,M\}$ such that, for every $t \geqslant \tau_0$ and $i \neq i^*$, one has*

$$L_{i,t} - L_{i^*,t} \geqslant \Delta t. \tag{4.8}$$

Consider any Hedge algorithm with (possibly data-dependent) learning rate η_t such that

- *$\eta_t \geqslant c_0\sqrt{(\log M)/t}$ for some constant $c_0 > 0$;*
- *it admits the following worst-case regret bound: $R_T \leqslant c_1\sqrt{T\log M}$ for every $T \geqslant 1$, for some $c_1 > 0$.*

Then, for every $T \geqslant 1$, the regret of this algorithm is upper bounded as

$$R_T \leqslant c_1\sqrt{\tau_0 \log M} + \frac{c_2 \log M + c_3 \log \Delta^{-1} + c_4}{\Delta} \tag{4.9}$$

where $c_2 = c_1 + \frac{\sqrt{8}}{c_0}$, $c_3 = \frac{\sqrt{8}}{c_0}$ and $c_4 = \frac{16}{c_0^2}$.

The idea of the proof of Theorem 4.2 is the same as that of Theorem 4.1, the only difference being the slightly longer initial phase to account for the adversarial nature of the losses. As a consequence of the general bound of Theorem 4.2, we can recover the guarantee of Theorem 4.1 (up to an additional $\log(\Delta^{-1})/\Delta$ term), both in expectation and with high probability, under more general stochastic assumptions than i.i.d. over time. The proofs of Theorem 4.2 and Corollary 4.1 are provided in Section 4.7.3.

Corollary 4.1. *Assume that the losses $(\ell_{i,t})_{1\leqslant i\leqslant M, t\geqslant 1}$ are random variables. Also, denoting $\mathcal{F}_t = \sigma\big((\ell_{i,s})_{1\leqslant i\leqslant M, 1\leqslant s\leqslant t}\big)$, assume that there exists i^* and $\Delta > 0$ such that*

$$\mathbb{E}\left[\ell_{i,t} - \ell_{i^*,t} | \mathcal{F}_{t-1}\right] \geqslant \Delta \tag{4.10}$$

for every $i \neq i^*$ and every $t \geqslant 1$. Then, for any Hedge algorithm satisfying the conditions of Theorem 4.2, and every $T \geqslant 1$:

$$\mathcal{R}_T \leqslant (5c_1 + 2c_2)\frac{\log M}{\Delta} + 2c_3\frac{\log \Delta^{-1}}{\Delta} + \frac{2c_4}{\Delta}, \tag{4.11}$$

with c_1, c_2, c_3, c_4 as in Theorem 4.2. In addition, for every $\varepsilon \in (0,1)$, we have

$$R_T \leqslant \left(c_1\sqrt{8} + 2c_2\right)\frac{\log M}{\Delta} + c_1\frac{\sqrt{8\log M \log \varepsilon^{-1}}}{\Delta} + 2c_3\frac{\log \Delta^{-1}}{\Delta} + \frac{2c_4}{\Delta} \tag{4.12}$$

with probability at least $1 - \varepsilon$.

4.3.3 Constant Hedge and Hedge with the doubling trick do not adapt to the stochastic case

Now, we show that the adaptivity of Decreasing Hedge to gaps in the losses, established in Sections 4.3.1 and 4.3.2, is not shared by the closely related Constant Hedge and Hedge with doubling trick, despite the fact that they both achieve the minimax optimal worst-case $O(\sqrt{T\log M})$ regret. Proposition 4.3 below shows that both algorithms fail to achieve a constant regret, and in fact to improve over their worst-case $\Theta(\sqrt{T\log M})$ regret guarantee, even in the extreme case of experts with constant losses 0 (for the leader), and 1 for the rest (*i.e.*, $\Delta = 1$).

Proposition 4.3. *Let $T \geqslant 1$, $M \geqslant 2$, and consider the experts $i = 1, \dots, M$ with losses $\ell_{1,t} = 0$, $\ell_{i,t} = 1$ $(1 \leqslant t \leqslant T, 2 \leqslant i \leqslant M)$. Then, the pseudo-regret of Constant Hedge with learning rate $\eta_t = c_0\sqrt{\log(M)/T}$ (where $c_0 > 0$ is a numerical constant) is lower bounded as follows:*

$$\mathcal{R}_T \geqslant \min\left(\frac{\sqrt{T\log M}}{3c_0}, \frac{T}{3}\right). \tag{4.13}$$

In addition, Hedge with doubling trick (4.3) also suffers a pseudo-regret satisfying

$$\mathcal{R}_T \geqslant \min\left(\frac{\sqrt{T\log M}}{6c_0}, \frac{T}{12}\right). \tag{4.14}$$

The proof of Proposition 4.3 is given in Section 4.7.4. Although Hedge with a doubling trick is recognized to be overly conservative and only suitable for worst-case scenarios Cesa-Bianchi and Lugosi, 2006 (especially due to its periodic restarts, after which it discards past observations), to the best of our knowledge Proposition 4.3 (together with Theorem 4.1) is the first to formally demonstrate the advantage of Decreasing Hedge over the doubling trick version. This implies that Decreasing Hedge should not be seen as merely a substitute for Constant Hedge to achieve anytime regret bounds. Indeed, even when the horizon T is fixed, Decreasing Hedge outperforms Constant Hedge in the stochastic setting.

4.4 Limitations of Decreasing Hedge in the stochastic case

In this section, we explore the limitations of the simple Decreasing Hedge algorithm in the stochastic regime, and exhibit situations where it performs worse than more sophisticated algorithms. The starting observation is that the sub-optimality gap Δ is a rather brittle measure of "hardness" of a stochastic instance, which does not fully reflect the achievable rates. We therefore consider the following fast-rate condition from statistical learning, which refines the sub-optimality gap as a measure of complexity of a stochastic instance.

Definition 4.1 (Bernstein condition)**.** Assume that the losses $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ are the realization of a stochastic process. Denote $\mathcal{F}_t = \sigma(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_t)$ the σ-algebra generated by $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_t$. For $\beta \in [0,1]$ and $B > 0$, the losses are said to satisfy the (β, B)-*Bernstein condition* if there exists i^* such that, for every $t \geqslant 1$ and $i \neq i^*$,

$$\mathbb{E}[(\ell_{i,t} - \ell_{i^*,t})^2 | \mathcal{F}_{t-1}] \leqslant B\mathbb{E}[\ell_{i,t} - \ell_{i^*,t} | \mathcal{F}_{t-1}]^\beta. \tag{4.15}$$

The Bernstein condition (Bartlett and Mendelson, 2006), a generalization of the Tsybakov margin condition (Tsybakov, 2004; Mammen and Tsybakov, 1999), is a geometric property on the losses which enables to obtain fast rates (e.g., faster than $O(1/\sqrt{n})$ for parametric classes) in statistical learning; we refer to van Erven et al. (2015) for a discussion of fast rates conditions. The Bernstein condition (4.15) quantifies the "easiness" of a stochastic instance, and generalizes the gap condition considered in the previous section (see Example 4.1 below). Roughly speaking, it states that good experts (with near-optimal expected loss) are highly correlated with the best expert. In the examples below, we assume that the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots$ are i.i.d.

Example 4.1 (Gap implies Bernstein)**.** If $\Delta_i = \mathbb{E}[\ell_{i,t} - \ell_{i^*,t}] \geqslant \Delta$ for $i \neq i^*$, then the $(1, \frac{1}{\Delta})$-Bernstein condition holds (Koolen et al., 2016, Lemma 4). Furthermore, letting $\alpha = \mathbb{E}[\ell_{i^*,t}]$

denote the expected loss of the best expert, the $(1, 1+\frac{2\alpha}{\Delta})$-Bernstein condition holds. Indeed, for any $i \neq i^*$, denoting $\mu_i := \mathbb{E}[\ell_{i,t}] = \alpha + \Delta_i$, we have (since $(u-v)^2 \leqslant \max(u^2, v^2) \leqslant u^2 + v^2 \leqslant u + v$ for $u, v \in [0,1]$):

$$\mathbb{E}\left[(\ell_{i,t} - \ell_{i^*,t})^2\right] \leqslant \mathbb{E}\left[\ell_{i,t} + \ell_{i^*,t}\right] = \frac{\mu_i + \alpha}{\mu_i - \alpha}\mathbb{E}\left[\ell_{i,t} - \ell_{i^*,t}\right] = \left(1 + \frac{2\alpha}{\Delta_i}\right)\mathbb{E}\left[\ell_{i,t} - \ell_{i^*,t}\right] ,$$

which establishes the claim since $\Delta_i \geqslant \Delta$. This provides an improvement when α is small.

Example 4.2 (Bernstein without a gap)**.** Let P be a distribution on $\mathcal{X} \times \{0,1\}$, where $\mathcal{X}$ is some measurable space. Assume that $(X_1, Y_1), (X_2, Y_2) \ldots$ are i.i.d. samples from P, and that the experts $i \in \{1, \ldots, M\}$ correspond to classifiers $f_i : \mathcal{X} \to \{0,1\}$: $\ell_{i,t} = \mathbf{1}(f_i(X_t) \neq Y_t)$, and that expert i^* is the Bayes classifier: $f_{i^*}(X) = \mathbf{1}(\eta(X) \geqslant 1/2)$, where $\eta(X) = \mathbb{P}(Y = 1|X)$. Tsybakov's low noise condition (Tsybakov, 2004), namely $\mathbb{P}(|2\eta(X) - 1| \leqslant t) \leqslant Ct^\kappa$ for some $C > 0$, $\kappa \geqslant 0$ and every $t > 0$, implies the $(\frac{\kappa}{\kappa+1}, B)$-Bernstein condition for some B (see, e.g., Boucheron et al., 2005). In addition, under the Massart condition (Massart and Nédélec, 2006) that $|\eta(X) - 1/2| \geqslant c > 0$, the $(1, 1/(2c))$-Bernstein condition holds. Note that these conditions may hold even with an arbitrarily small sub-optimality gap Δ, since the f_i, $i \neq i^*$, may be arbitrary.

Theorem 4.3 below shows that Decreasing Hedge fails to achieve improved rates under Bernstein conditions.

Theorem 4.3. *For every $T \geqslant 1$, there exists a $(1,1)$-Bernstein stochastic instance on which the pseudo-regret of the Decreasing Hedge algorithm with $\eta_t = c_0\sqrt{(\log M)/t}$ satisfies*

$$\mathcal{R}_T \geqslant \frac{1}{3}\min\left(\frac{1}{c_0}\sqrt{T\log M}, T\right).$$

The proof of Theorem 4.3 is given in Section 4.7.6. By contrast, it was shown by Koolen et al. (2016) (and implicitly used by Gaillard et al., 2014) that some adaptive algorithms with data-dependent regret bounds enjoy improved regret under the Bernstein condition. For the sake of completeness, we state this fact in Proposition 4.4 below, which corresponds to Koolen et al. (2016, Theorem 2), but where the dependence on B is made explicit. We also only provide a bound in expectation, which considerably simplifies the proof. The proof of Proposition 4.4, which uses the same ideas as Gaillard et al. (2014, Theorem 11), is provided in Section 4.7.5.

Proposition 4.4. *Consider an algorithm for the Hedge problem which satisfies the following regret bound: for every $i \in \{1, \ldots, M\}$,*

$$R_{i,T} \leqslant C_1\sqrt{(\log M)\sum_{t=1}^{T}(\widehat{\ell}_t - \ell_{i,t})^2} + C_2\log M \tag{4.16}$$

where $C_1, C_2 > 0$ are constants. Assume that the losses satisfy the (β, B)-Bernstein condition. Then, the pseudo-regret of the algorithm satisfies:

$$\mathcal{R}_T \leqslant C_3(B\log M)^{\frac{1}{2-\beta}}T^{\frac{1-\beta}{2-\beta}} + C_4\log M \tag{4.17}$$

where $C_3 = \max(1, 4C_1^2)$ and $C_4 = 2C_2$.

The data-dependent regret bound (4.16), a "second-order" bound, is satisfied by adaptive algorithms such as Adapt-ML-Prod (Gaillard et al., 2014) and Squint (Koolen and van Erven, 2015). A slightly different variant of second-order regret bounds, which depends on some cumulative variance of the losses across experts, has been considered by Cesa-Bianchi et al. (2007); de Rooij et al. (2014), and is achieved by Hedge algorithms with a data-dependent tuning of the learning rate. Second-order bounds refine so-called *first-order* bounds (Cesa-Bianchi et al., 1997; Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006), which are adversarial regret bounds that scale as $O(\sqrt{L_T^* \log M} + \log M)$, where L_T^* denotes the cumulative loss of the best expert. While first-order bounds may still scale as the worst-case $O(\sqrt{T \log M})$ rate in a typical stochastic instance (where the best expert has a positive expected loss), second-order algorithms are known to achieve constant $O((\log M)/\Delta)$ regret in the stochastic case with gap Δ (Gaillard et al., 2014; Koolen and van Erven, 2015).

Theorem 4.3, in light of Proposition 4.4, clarifies where the advantage of second-order algorithms compared to Decreasing Hedge lies: unlike the latter, they can exploit Bernstein conditions on the losses. The contrast is most apparent for Bernstein instances with $\beta = 1$. By Example 4.1, the existence of a gap Δ implies that the $(1, B)$-Bernstein condition holds with $B \leqslant \frac{1}{\Delta}$. However, as shown by Example 4.2, B can in fact be much smaller than Δ, in which case the regret bound (4.17) satisfied by second-order algorithms, namely $O(B \log M)$, significantly improves over the upper bound of $O((\log M)/\Delta)$ of Decreasing Hedge from Theorem 4.1. Theorem 4.3 provides an instance where the difference does occur, in the most pronounced case where $B = 1$, so that second-order algorithms enjoy small $O(\log M)$ regret, while Decreasing Hedge suffers $\Theta(\sqrt{T \log M})$ regret.

Remark 4.2. The advantage of larger learning rates on some stochastic instances may be understood intuitively as follows. Consider an instance with B small but small gap Δ. The learning rate of Decreasing Hedge is large enough that it can rule out bad experts (with large enough gap Δ_i) at the optimal rate (*i.e.*, at time $(\log M)/\Delta_i^2$). However, once these bad experts are ruled out, near-optimal experts (with small gap Δ_i) are ruled out late (after $(\log M)/\Delta_i^2$ rounds). On the other hand, the Bernstein assumption entails that those experts are highly correlated with the best expert, hence the amount of noise on the relative losses of these near-optimal experts is small, so that a larger learning rate could be safely used and would enable to dismiss near-optimal experts sooner.

Setting the Bernstein condition aside, we conclude by investigating the intrinsic limitations of Decreasing Hedge in the stochastic setting. Indeed, it is natural to ask whether Decreasing Hedge can exploit some other regularity of a stochastic instance, apart from the gap Δ. Theorem 4.4 shows that this is in fact not the case.

Theorem 4.4. *For every i.i.d. (over time) stochastic instance with a unique best expert*

$$i^* = \operatorname*{arg\,min}_{1 \leqslant i \leqslant M} \mathbb{E}[\ell_{i,t}],$$

the pseudo-regret of Decreasing Hedge (with $c_0 \geqslant 1$) satisfies

$$\mathcal{R}_T \geqslant \frac{1}{450 c_0^4 (\log M)^2 \Delta}$$

for $T \geqslant \frac{1}{4\Delta^2}$, where $\Delta := \inf_{i \neq i^*} \mathbb{E}[\ell_{i,t} - \ell_{i^*,t}]$.

Theorem 4.4 shows (together with the upper bound of Theorem 4.1) that the eventual regret of Decreasing Hedge on *any* stochastic instance is determined by the sub-optimality gap Δ, and scales (up to a $\log^3 M$ factor, depending on the number of near-optimal experts) as $\Theta(\frac{1}{\Delta})$. This characterizes the behavior of Decreasing Hedge on any stochastic instance.

4.5 Experiments

In this section, we illustrate our theoretical results by numerical experiments that compare the behavior of various Hedge algorithms in the stochastic regime.

Algorithms. We consider the following algorithms: `hedge` is Decreasing Hedge with the default learning rates $\eta_t = 2\sqrt{\log(M)/t}$, `hedge_constant` is Constant Hedge with constant learning rate $\eta_t = \sqrt{8\log(M)/T}$, `hedge_doubling` is Hedge with doubling trick with $c_0 = \sqrt{8}$, `adahedge` is the AdaHedge algorithm from de Rooij et al. (2014), which is a variant of the Hedge algorithm with a data-dependent tuning of the learning rate η_t (based on $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1}$). As shown in the note Koolen (2018), AdaHedge also benefits from Bernstein conditions. A related algorithm, namely Hedge with second-order tuning of the learning rate (Cesa-Bianchi et al., 2007), performed similarly to AdaHedge on the examples considered below, and was therefore not included. `FTL` is Follow-the-Leader (Cesa-Bianchi and Lugosi, 2006) which puts all mass on the expert with the smallest loss (breaking ties randomly). While FTL serves as a benchmark in the stochastic setting, unlike the other algorithms it lacks any guarantee in the adversarial regime, where its worst-case regret is *linear* in T.

Results. We report in Figure 4.1 the cumulative regrets of the considered algorithms in four examples. The results for the stochastic instances (a), (b) and (c) described below are averaged over 50 trials.

(*a*) *Stochastic instance with a gap.* This is the standard instance considered here. The losses are drawn independently from Bernoulli distributions (one of parameter 0.3, 2 of parameter 0.4 and 7 of parameter 0.5, so that $M = 10$ and $\Delta = 0.1$). The results of Figure 4.1a confirm our theoretical results: Decreasing Hedge achieves a small, constant regret which is close to that of AdaHedge and FTL, while Constant Hedge and Hedge with doubling trick suffer a larger regret of order $\sqrt{T}$ (note that, although the expected regret of Constant Hedge converges in this case, the value of this limit depends on its learning rate and hence on T).

(*b*) *"Hard" stochastic instance.* This example has a zero gap $\Delta = 0$ between the two leading experts and $M = 10$, which makes it "hard" from the standpoint of Theorem 4.1 (which no longer applies in this limit case). The losses are drawn from independent Bernoulli distributions, of parameters 0.5 for the 2 leading experts, and 0.7 for the 8 remaining ones. Although all algorithms suffer an unavoidable $\Theta(\sqrt{T})$ regret due to pure noise, Decreasing Hedge, AdaHedge and FTL achieve better regret than the two conservative Hedge variants (Figure 4.1b). This is due to the fact that for the former algorithms, the weights of suboptimal experts decrease quickly and only induce a constant regret.

(*c*) *Small loss for the best expert.* In this experiment, we illustrate one advantage of adaptive Hedge algorithms such as AdaHedge over Decreasing Hedge, namely the fact that they admit

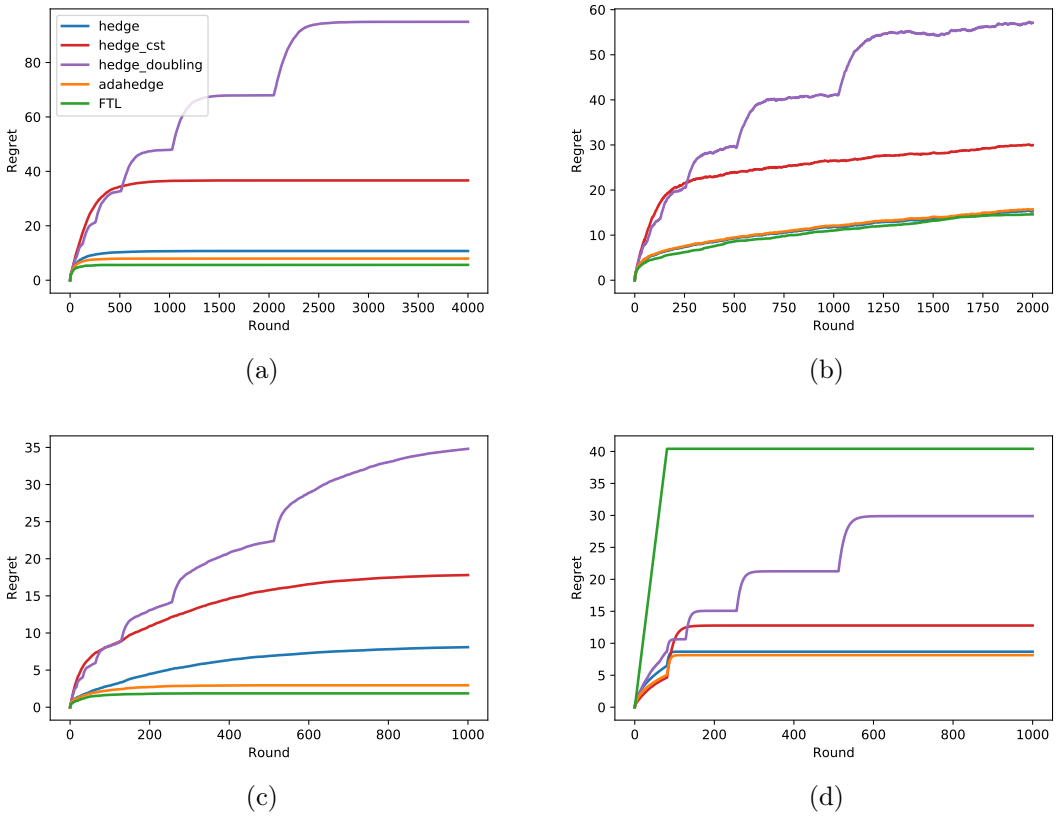


Figure 4.1: Cumulative regret of Hedge algorithms on four examples, see text for a precise description and discussion about the results. (a) Stochastic instance with a gap; (b) "Hard" stochastic instance; (c) Small loss for the best expert; (d) Adversarial with a gap instance.

improved regret bounds when the leading expert has small loss. We considered in this experiment $M = 10$, $\Delta = 0.04$ and the leading expert is $\mathsf{Beta}(0.04, 0.96)$, then 4 $\mathsf{Beta}(0.08, 0.92)$, then 5 $\mathsf{Beta}(0.5, 0.5)$.

(d) Adversarial with a gap instance. This simple instance is not random, and satisfies the assumptions of Theorem 4.2. It is defined by $M = 3$, $\Delta = 0.04$, $\ell_{3,t} = \frac{3}{4}$ for $t \geqslant 1$, $(\ell_{1,t}, \ell_{2,t}) = (\frac{1}{2}, 0)$ if $t = 1$, $(0, 1)$ if $t \geqslant 80$ or if t is even, and $(1, 0)$ otherwise. FTL suffers linear regret in the first phase, while Constant Hedge and Hedge with doubling trick suffer $\Theta(\sqrt{T})$ during the second phase.

4.6 Conclusion

In this chapter, we carried the regret analysis of the standard exponential weights (Hedge) algorithm in the stochastic expert setting, closing a gap in the existing literature. Our analysis reveals that, despite being tuned for the worst-case adversarial setting and lacking any adaptive tuning of the learning rate, Decreasing Hedge achieves optimal regret in the stochastic setting. This property also enables one to distinguish it qualitatively from other variants including the one with fixed (horizon-dependent) learning rate or the one with doubling trick, which both

fail to adapt to gaps in the losses. To the best of our knowledge, this is the first result that shows the superiority of the decreasing learning rate over the doubling trick. In addition, it suggests that, even for a fixed time horizon T, the decreasing learning rate tuning should be favored over the constant one.

Finally, we showed that the regret of Decreasing Hedge on any stochastic instance is essentially characterized by the sub-optimality gap Δ. This shows that adaptive algorithms, including algorithms achieving second-order regret bounds, can actually outperform Decreasing Hedge on some stochastic instances that exhibit a more refined form of "easiness".

A link with stochastic optimization. Our results have a similar flavor to a well-known result (Moulines and Bach, 2011; Bach, 2014) in stochastic optimization: stochastic gradient descent (SGD) with learning rate $\eta_t \propto 1/\sqrt{t}$ (which is tuned for the convex case but not for the non-strongly convex case) and Polyak-Ruppert averaging achieves a fast $O(1/(\mu t))$ excess risk rate for μ-strongly convex problems, without the knowledge of μ. However, this link stops here since the two results are of a significantly different nature: the $O(1/(\mu t))$ rate is satisfied only by SGD with iterate averaging, and it does not come from a regret bound. In fact, the opposite phenomenon occurs: in stochastic optimization, SGD uses a *larger* $\Theta(1/\sqrt{t})$ step-size than the $\Theta(1/(\mu t))$ step size which exploits the knowledge of strong convexity, but the effect of this larger step-size is balanced by the averaging. By contrast, in the expert setting, Hedge uses a *smaller* $\Theta(\sqrt{(\log M)/t})$ learning rate than the constant, large enough learning rate which exploits the knowledge of the stochastic nature of the problem.

Acknowledgments. We wish to thank four anonymous JMLR reviewers of the article Mourtada and Gaïffas (2019b) for their helpful feedback and suggestions on this work. The proof of Proposition 4.2 was proposed by an Anonymous Referee, which allowed to shorten our initial proof.

4.7 Proofs

We now provide the proofs of the results from the previous sections, by order of appearance in the text.

4.7.1 Proof of Theorem 4.1

Let $t_0 = \left\lceil \frac{8 \log M}{\Delta^2} \right\rceil$, so that $\sqrt{t_0} \leqslant \sqrt{1 + \frac{8 \log M}{\Delta^2}} \leqslant 1 + \frac{\sqrt{8 \log M}}{\Delta}$ (since $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for $a, b \geqslant 0$). The worst-case regret bound of Hedge (Proposition 4.1) shows that for $1 \leqslant T \leqslant t_0$:

$$R_{i^*,T} \leqslant \sqrt{T \log M} \leqslant \sqrt{t_0 \log M} \leqslant \sqrt{\log M} + \frac{2\sqrt{2} \log M}{\Delta} \leqslant \frac{4 \log M}{\Delta} \tag{4.18}$$

(since $\log M \geqslant 1$ as $M \geqslant 3$, $\Delta \leqslant 1$ and $2\sqrt{2} \leqslant 3$), which establishes (4.6) for $T \leqslant t_0$. In order to prove (4.6) for $T \geqslant t_0 + 1$, we start by decomposing the regret with respect to i^* as

$$R_{i^*,T} = \widehat{L}_T - L_{i^*,T} = \widehat{L}_{t_0} - L_{i^*,t_0} + \sum_{t=t_0+1}^{T} (\widehat{\ell}_t - \ell_{i^*,t}) \,. \tag{4.19}$$

Since $\widehat{L}_{t_0} - L_{i^*,t_0} \leqslant R_{t_0}$ is controlled by (4.18), it remains to upper bound the second term in (4.19). First, for every $t \geqslant t_0 + 1$,

$$\widehat{\ell}_t - \ell_{i^*,t} = \sum_{i \neq i^*} v_{i,t}(\ell_{i,t} - \ell_{i^*,t})\,. \tag{4.20}$$

Since $\boldsymbol{\ell}_t$ is independent of $\boldsymbol{v}_t$ (which is $\sigma(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$-measurable), taking the expectation in (4.20) yields, denoting $\Delta_i = \mathbb{E}[\ell_{i,t} - \ell_{i^*,t}]$,

$$\mathbb{E}[\widehat{\ell}_t - \ell_{i^*,t}] = \sum_{i \neq i^*} \Delta_i \mathbb{E}[v_{i,t}]\,. \tag{4.21}$$

First, for every $i \neq i^*$, applying Hoeffding's inequality to the i.i.d. centered variables $Z_{i,t} := -\ell_{i,t} + \ell_{i^*,t} + \Delta_i$, which belong to $[-1 + \Delta_i, 1 + \Delta_i]$, yields

$$\begin{aligned}
\mathbb{P}\left(L_{i,t-1} - L_{i^*,t-1} < \frac{\Delta_i(t-1)}{2}\right) &= \mathbb{P}\left(\sum_{s=1}^{t-1} Z_{i,s} > \frac{\Delta_i(t-1)}{2}\right) \\
&\leqslant e^{-\frac{t-1}{2}(\Delta_i/2)^2} \\
&= e^{-(t-1)\Delta_i^2/8}\,.
\end{aligned} \tag{4.22}$$

On the other hand, if $L_{i,t-1} - L_{i^*,t-1} \geqslant \Delta_i(t-1)/2$, then

$$\begin{aligned}
v_{i,t} &= \frac{e^{-\eta_t(L_{i,t-1} - L_{i^*,t-1})}}{1 + \sum_{j \neq i^*} e^{-\eta_t(L_{j,t-1} - L_{i^*,t-1})}} \\
&\leqslant e^{-2\sqrt{(\log M)/t} \times \Delta_i(t-1)/2} \\
&\leqslant e^{-\Delta_i\sqrt{(t-1)(\log M)/2}}
\end{aligned} \tag{4.23}$$

since $t \leqslant 2(t-1)$. It follows from (4.23) and (4.22) that, for $t \geqslant t_0 + 1 \geqslant 2$,

$$\begin{aligned}
\mathbb{E}[v_{i,t}] &\leqslant \mathbb{P}\left(L_{i,t-1} - L_{i^*,t-1} > \frac{\Delta_i(t-1)}{2}\right) + e^{-\Delta_i\sqrt{(t-1)(\log M)/2}} \\
&\leqslant e^{-(t-1)\Delta_i^2/8} + e^{-\Delta_i\sqrt{(t-1)(\log M)/2}}\,.
\end{aligned} \tag{4.24}$$

Now, a simple analysis of functions shows that the functions $f_1(u) = ue^{-u}$ and $f_2(u) = ue^{-u^2/2}$ are decreasing on $[1, +\infty)$. Since $\Delta_i \geqslant \Delta$, this entails that

$$\Delta_i e^{-(t-1)\Delta_i^2/8} = \frac{2}{\sqrt{t-1}} f_2\left(\frac{\sqrt{t-1}\Delta_i}{2}\right) \leqslant \frac{2}{\sqrt{t-1}} f_2\left(\frac{\sqrt{t-1}\Delta}{2}\right) = \Delta e^{-(t-1)\Delta^2/8} \tag{4.25}$$

provided that $\frac{\sqrt{t-1}\Delta}{2} \geqslant 1$, *i.e.* $t \geqslant 1 + \frac{4}{\Delta^2}$, which is the case since $t \geqslant t_0 + 1 \geqslant 1 + \frac{8\log M}{\Delta^2}$. Likewise,

$$\Delta_i e^{-\Delta_i\sqrt{(t-1)(\log M)/2}} \leqslant \Delta e^{-\Delta\sqrt{(t-1)(\log M)/2}} \tag{4.26}$$

if $\Delta\sqrt{(t-1)(\log M)/2} \geqslant 1$, *i.e.* $t \geqslant 1 + \frac{2}{(\log M)\Delta^2}$, which is ensured by $t \geqslant t_0 + 1$. It follows from (4.21), (4.24), (4.25) and (4.26) that for every $t \geqslant t_0 + 1$:

$$\begin{aligned}
\mathbb{E}[\widehat{\ell}_t - \ell_{i^*,t}] &\leqslant M\Delta e^{-(t-1)\Delta^2/8} + M\Delta e^{-\Delta\sqrt{(t-1)(\log M)/2}} \\
&= \left(Me^{-t_0\Delta^2/8}\right)\left(\Delta e^{-(t-t_0-1)\Delta^2/8}\right) + \left(Me^{-\Delta\sqrt{(t-1)(\log M)/8}}\right)\left(\Delta e^{-\Delta\sqrt{(t-1)(\log M)/8}}\right) \\
&\leqslant \Delta e^{-(t-t_0-1)\Delta^2/8} + \Delta e^{-\Delta\sqrt{(t-1)/8}}
\end{aligned} \tag{4.27}$$

where inequality (4.27) comes from the bound $Me^{-t_0\Delta^2/8} \leqslant 1$ (since $t_0 \geqslant \frac{8\log M}{\Delta^2}$) and from the fact that $Me^{-\Delta\sqrt{(t-1)(\log M)/8}} \leqslant 1$ amounts to $t \geqslant 1 + \frac{8\log M}{\Delta^2}$, that is, to $t \geqslant t_0 + 1$. Summing inequality (4.27) yields, for every $T \geqslant t_0 + 1$,

$$\begin{aligned}
\mathbb{E}[\sum_{t=t_0+1}^{T}(\ell_t - \ell_{i^*,t})] &\leqslant \sum_{t=t_0+1}^{T}\left\{\Delta e^{-(t-t_0-1)\Delta^2/8} + \Delta e^{-\Delta\sqrt{(t-1)/8}}\right\} \\
&\leqslant \Delta\sum_{t\geqslant 0} e^{-t\Delta^2/8} + \Delta\sum_{t\geqslant 1} e^{-(\Delta/\sqrt{8})\sqrt{t}} \\
&\leqslant \Delta\left(1 + \frac{8}{\Delta^2}\right) + \Delta \times \frac{2}{(\Delta/\sqrt{8})^2} \qquad (4.28) \\
&\leqslant \frac{25}{\Delta} \qquad (4.29)
\end{aligned}$$

where inequality (4.28) comes from Lemma 4.1 below. Finally, combining inequalities (4.18) and (4.28) yields the pseudo-regret bound $\mathcal{R}_T \leqslant \frac{4\log M + 25}{\Delta}$.

Lemma 4.1. *For every $\alpha > 0$,*

$$\sum_{t\geqslant 1} e^{-\alpha t} \leqslant \frac{1}{\alpha} \qquad (4.30)$$

$$\sum_{t\geqslant 1} e^{-\alpha\sqrt{t}} \leqslant \frac{2}{\alpha^2}\,. \qquad (4.31)$$

Proof. Since the functions $t \mapsto e^{-\alpha t}$ and $t \mapsto e^{-\alpha\sqrt{t}}$ are decreasing on $\mathbf{R}^+$, we have

$$\begin{aligned}
\sum_{t\geqslant 1} e^{-\alpha t} &\leqslant \int_0^\infty e^{-\alpha t}\mathrm{d}t = \frac{1}{\alpha} \\
\sum_{t\geqslant 1} e^{-\alpha\sqrt{t}} &\leqslant \int_0^{+\infty} e^{-\alpha\sqrt{t}}\mathrm{d}t \underset{u=\alpha\sqrt{t}}{=} \frac{2}{\alpha^2}\int_0^{+\infty} ue^{-u}\mathrm{d}u = \frac{2}{\alpha^2}.
\end{aligned}$$

□

Remark 4.3. While the upper bound of Theorem 4.1 is stated for the pseudo-regret $\mathcal{R}_T$, a similar upper bound holds for the expected regret $\mathbb{E}[R_T]$. Indeed, under the assumptions of Theorem 4.1, for every $T \geqslant \frac{4\log M}{\Delta^2}$, we have $\mathbb{E}[R_T] \leqslant \mathcal{R}_T + \frac{1.1}{\Delta}$.

Proof. Note that $\mathbb{E}[R_T] - \mathcal{R}_T = \mathbb{E}[L_{i^*,T} - \min_{1\leqslant i\leqslant T} L_{i,T}]$. For every $a \geqslant 0$, Hoeffding's inequality (applied to the i.i.d. centered variables $\ell_{i^*,t} - \ell_{i,t} + \Delta_i \in [-1+\Delta_i, 1+\Delta_i]$, $1 \leqslant t \leqslant T$) entails

$$\begin{aligned}
\mathbb{P}\left(L_{i^*,T} - \min_{1\leqslant i\leqslant T} L_{i,T} \geqslant a\right) &\leqslant \sum_{i\neq i^*}\mathbb{P}\left(L_{i^*,T} - L_{i,T} + \Delta_i T \geqslant \Delta_i T + a\right) \\
&\leqslant \sum_{i\neq i^*} e^{-(\Delta_i T + a)^2/(2T)} \qquad (4.32) \\
&\leqslant Me^{-T\Delta^2/2}e^{-a^2/(2T)} \\
&\leqslant e^{-T\Delta^2/4}e^{-a^2/(2T)}\,, \qquad (4.33)
\end{aligned}$$

where inequality (4.33) comes from the fact that $Me^{-T\Delta^2/4} \leqslant 1$ since $T \geqslant \frac{4\log M}{\Delta^2}$. Since the random variable $L_{i^*,T} - \min_{1\leqslant i\leqslant T} L_{i,T}$ is nonnegative, this implies that

$$\begin{aligned}
\mathbb{E}\left[L_{i^*,T} - \min_{1\leqslant i\leqslant T} L_{i,T}\right] &= \int_0^\infty \mathbb{P}\left(L_{i^*,T} - \min_{1\leqslant i\leqslant T} L_{i,T} \geqslant a\right) \mathrm{d}a \\
&\leqslant e^{-T\Delta^2/4} \int_0^\infty e^{-a^2/(2T)} \mathrm{d}a \\
&= \sqrt{\frac{\pi}{2}} \cdot \sqrt{T} e^{-T\Delta^2/4} \\
&= \frac{\sqrt{\pi}}{\Delta}\left[\Delta\sqrt{T/2} \cdot e^{-(\Delta\sqrt{T/2})^2/2}\right] \\
&\leqslant \frac{\sqrt{\pi/e}}{\Delta} \qquad (4.34)
\end{aligned}$$

where inequality (4.34) comes from the fact that the function $u \mapsto ue^{-u^2/2}$ attains its maximum on $\mathbf{R}^+$ at $u = 1$. This concludes the proof, since $\sqrt{\pi/e} \leqslant 1.1$. □

4.7.2 Proof of Proposition 4.2

Fix M, Δ and T as in Proposition 4.2. For $i^* \in \{1, \ldots, M\}$, denote $\mathbb{P}_{i^*}$ the following distribution on $[0,1]^{M\times T}$: if $(\ell_{i,t})_{1\leqslant i\leqslant M, 1\leqslant t\leqslant T} \sim \mathbb{P}_{i^*}$, then the variables $\ell_{i,t}$ are independent Bernoulli variables, of parameter $\frac{1}{2} - \Delta$ if $i = i^*$ and $\frac{1}{2}$ otherwise; also, denote by $\mathbb{E}_{i^*}$ the expectation with respect to $\mathbb{P}_{i^*}$. Let $\mathcal{A} = (A_t)_{1\leqslant t\leqslant T}$ be any Hedging algorithm, where $A_t : [0,1]^{M\times(t-1)} \to \mathcal{P}_M$ maps past losses $(\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{t-1})$ to an element of the probability simplex $\mathcal{P}_M \subset \mathbf{R}^M$ on $\{1, \ldots, M\}$. For any $i^* \in \{1, \ldots, M\}$, let $\mathcal{R}_T(i^*, \mathcal{A})$ denote the pseudo-regret of algorithm $\mathcal{A}$ under the distribution $\mathbb{P}_{i^*}$. Since $\boldsymbol{\ell}_t$ is independent of $\boldsymbol{v}_t$ under $\mathbb{P}_{i^*}$, we have

$$\mathcal{R}_T(i^*, \mathcal{A}) = \sum_{t=1}^T \sum_{i\neq i^*} \mathbb{E}_{i^*}\left[v_{i,t}(\ell_{i,t} - \ell_{i^*,t})\right] = \Delta \sum_{t=1}^T \sum_{i\neq i^*} \mathbb{E}_{i^*}[v_{i,t}] = \Delta \sum_{t=1}^T \mathbb{E}_{i^*}[1 - v_{i^*,t}] \qquad (4.35)$$

with $\boldsymbol{v}_t := A_t(\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{t-1})$. It follows from Equation (4.35) that, for every $\mathcal{A}$ and i^*, $\mathcal{R}_T(i^*, \mathcal{A})$ increases with T. Hence, without loss of generality we may assume that $T = \lfloor(\log M)/(16\Delta^2)\rfloor$. The maximum pseudo-regret of $\mathcal{A}$ on the instances $\mathbb{P}_{i^*}$ is lower-bounded as follows:

$$\sup_{1\leqslant i^*\leqslant M} \mathcal{R}_T(i^*, \mathcal{A}) \geqslant \frac{1}{M} \sum_{1\leqslant i^*\leqslant M} \mathcal{R}_T(i^*, \mathcal{A}) = \frac{1}{M} \sum_{1\leqslant i^*\leqslant M} \Delta \sum_{t=1}^T \mathbb{E}_{i^*}[1 - v_{i^*,t}]\,. \qquad (4.36)$$

We now "randomize" the algorithm $\mathcal{A}$, by replacing it with a randomized algorithm which picks expert i at time t with probability $v_{i,t}$. Formally, let $\widetilde{P} = \mathcal{U}([0,1])^{\otimes T}$ be the distribution of T independent uniform random variables on $[0,1]$, and denote $\widetilde{\mathbb{P}}_{i^*} = \mathbb{P}_{i^*} \otimes \widetilde{P}$ for $i^* \in \{1, \ldots, M\}$. Furthermore, for every $\boldsymbol{v} \in \mathcal{P}_M$, let $I_{\boldsymbol{v}} : [0,1] \to \{1, \ldots, M\}$ be a measurable map such that $\mathbb{P}(I_{\boldsymbol{v}}(U) = i) = v_i$ for every $i \in \{1, \ldots, M\}$, where $U \sim \mathcal{U}([0,1])$. For every sequence of losses $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ and random variables $U_1, \ldots, U_T$ and every $1 \leqslant t \leqslant T$, let $I_t = I_{\boldsymbol{v}_t}(U_t)$, where $\boldsymbol{v}_t = A_t(\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_t)$.

Denote by $\widetilde{\mathbb{E}}_{i^*}$ the expectation with respect to $\widetilde{\mathbb{P}}_{i^*}$. By definition of $I_{\boldsymbol{v}}$, we have $\mathbb{E}_{i^*}[v_{i^*,t}] = \widetilde{\mathbb{E}}_{i^*}[\mathbf{1}(I_t = i^*)]$ so that, denoting $N_i = \sum_{i=1}^T \mathbf{1}(I_t = i)$ the number of times expert i is picked,

$$\sum_{t=1}^T \mathbb{E}_{i^*}[1 - v_{i^*,t}] = \widetilde{\mathbb{E}}_{i^*}[T - N_{i^*}] \geqslant \mathbb{P}_{i^*}(N_{i^*} \leqslant T/2) \cdot \frac{T}{2}\,.$$

Hence, letting $A_i \subseteq [0,1]^{M\times T} \times [0,1]^T$ be the event $\{N_i > T/2\}$, Equation (4.36) implies that

$$\sup_{1\leqslant i^*\leqslant M} \mathcal{R}_T(i^*, \mathcal{A}) \geqslant \frac{\Delta T}{2} \times \frac{1}{M} \sum_{1\leqslant i^*\leqslant M} \left(1 - \widetilde{\mathbb{P}}_{i^*}(A_{i^*})\right). \tag{4.37}$$

It now remains to upper bound $\frac{1}{M}\sum_{i^*} \widetilde{\mathbb{P}}_{i^*}(A_{i^*})$. To do this, first note that the events A_{i^*}, $1 \leqslant i^* \leqslant M$, are pairwise disjoint. Hence, Fano's inequality (see Gerchinovitz et al., 2017, p.2) implies that, for every distribution $\widetilde{\mathbb{Q}}$ on $[0,1]^{M\times T} \times [0,1]^T$,

$$\frac{1}{M} \sum_{1\leqslant i^*\leqslant M} \widetilde{\mathbb{P}}_{i^*}(A_{i^*}) \leqslant \frac{1}{\log M}\left\{\frac{1}{M}\sum_{1\leqslant i^*\leqslant M} \mathrm{KL}(\widetilde{\mathbb{P}}_{i^*}, \widetilde{\mathbb{Q}}) + \log 2\right\} \tag{4.38}$$

where $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ denotes the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$. Here, we take $\widetilde{\mathbb{Q}} = \mathbb{Q} \otimes \widetilde{P}$, where $\mathbb{Q}$ is the product of Bernoulli distributions $\mathcal{B}(1/2)^{\otimes T}$. This choice leads to

$$\mathrm{KL}(\widetilde{\mathbb{P}}_{i^*}, \widetilde{\mathbb{Q}}) = \mathrm{KL}(\mathbb{P}_{i^*}, \mathbb{Q}) = T \cdot \mathrm{KL}(\mathcal{B}(1/2 - \Delta), \mathcal{B}(1/2)) \leqslant 4T\Delta^2 \leqslant \frac{\log M}{4}\,,$$

where the first bound is obtained by comparing KL and χ^2 divergences (Tsybakov, 2009, Lemma 2.7). Hence, inequality (4.38) becomes (recalling that $M \geqslant 4$)

$$\frac{1}{M} \sum_{1\leqslant i^*\leqslant M} \widetilde{\mathbb{P}}_{i^*}(A_{i^*}) \leqslant \frac{(\log M)/4}{\log M} + \frac{\log 2}{\log M} \leqslant \frac{3}{4}\,;$$

plugging this into (4.37) yields, noting that $T = \lfloor (\log M)/(16\Delta^2) \rfloor \geqslant (\log M)/(32\Delta^2)$ since $(\log M)/(16\Delta^2) \geqslant 1$ (as $M \geqslant 4$ and $\Delta \leqslant \frac{1}{4}$),

$$\sup_{1\leqslant i^*\leqslant M} \mathcal{R}_T(i^*, \mathcal{A}) \geqslant \frac{\Delta T}{2} \times \frac{1}{4} \geqslant \frac{\log M}{256\Delta}\,.$$

This concludes the proof.

4.7.3 Proof of Theorem 4.2 and Corollary 4.1

Let t_0 be the smallest integer $t \geqslant 1$ such that $Me^{-c_0\Delta\sqrt{t\log(M)/8}} \leqslant \Delta$, namely $t_0 = \left\lceil \frac{8}{c_0^2\Delta^2} \frac{\log^2(M/\Delta)}{\log M} \right\rceil$. Note that $\sqrt{t_0} \leqslant \sqrt{1 + \frac{8}{c_0^2\Delta^2}\frac{\log^2(M/\Delta)}{\log M}} \leqslant 1 + \frac{\sqrt{8}}{c_0\Delta}\frac{\log(M/\Delta)}{\sqrt{\log M}}$. Let $t_1 := t_0 \vee \tau_0$. For every $T \leqslant t_1$, the regret bound in the assumption of Theorem 4.2 implies

$$\begin{aligned} R_T &\leqslant c_1\sqrt{T\log M} \\ &\leqslant c_1\sqrt{\tau_0 \log M} + c_1\sqrt{t_0 \log M} \\ &\leqslant c_1\sqrt{\tau_0 \log M} + c_1\sqrt{\log M} + \frac{\sqrt{8}\log(M/\Delta)}{c_0\Delta} \end{aligned} \tag{4.39}$$

which implies (4.9) with $c_2 = c_1 + \frac{\sqrt{8}}{c_0}$ and $c_3 = \frac{\sqrt{8}}{c_0}$ (since $1 \leqslant \sqrt{\log M} \leqslant \frac{\log M}{\Delta}$). From now on, assume that $T \geqslant t_1 + 1$. Since $T \geqslant \tau_0$, we have $R_T = \widehat{L}_T - L_{i^*,T}$, so that

$$R_T = \widehat{L}_{t_1} - L_{i^*,t_1} + \sum_{t=t_1+1}^{T} \left(\widehat{\ell}_t - \ell_{i^*,t}\right). \tag{4.40}$$

In addition, we have for $t \geqslant t_1 + 1$

$$\begin{aligned}
\widehat{\ell}_t - \ell_{i^*,t} &= \sum_{i \neq i^*} v_{i,t}(\ell_{i,t} - \ell_{i^*,t}) \\
&\leqslant \sum_{i \neq i^*} v_{i,t} \\
&= \sum_{i \neq i^*} \frac{e^{-\eta_t(L_{i,t-1} - L_{i^*,t-1})}}{1 + \sum_{j \neq i^*} e^{-\eta_t(L_{j,t-1} - L_{i^*,t-1})}} \\
&\leqslant \sum_{i \neq i^*} e^{-c_0\sqrt{(\log M)/t} \times \Delta(t-1)} && (4.41) \\
&\leqslant M e^{-c_0 \Delta \sqrt{(t-1)(\log M)/2}} \\
&\leqslant \left(M e^{-c_0\Delta\sqrt{t_0(\log M)/8}}\right) e^{-c_0\Delta\sqrt{(t-1)/8}} && (4.42) \\
&\leqslant \Delta e^{-c_0\Delta\sqrt{(t-1)/8}} && (4.43)
\end{aligned}$$

where (4.41) comes from the fact that $\eta_t \geqslant c_0\sqrt{(\log M)/t}$ and $L_{i,t-1} - L_{i^*,t-1} \geqslant \Delta(t-1)$ (since $t - 1 \geqslant t_1 \geqslant \tau_0$), (4.42) from the fact that $t - 1 \geqslant t_0$ and $\log M \geqslant 1$, and (4.43) from the fact that $M e^{-c_0\Delta\sqrt{t_0(\log M)/8}} \leqslant \Delta$. Summing inequality (4.43), we obtain

$$\begin{aligned}
\sum_{t=t_1+1}^{T} (\widehat{\ell}_t - \ell_{i^*,t}) &\leqslant \sum_{t=t_1+1}^{T} \Delta e^{-c_0\Delta\sqrt{(t-1)/8}} \\
&\leqslant \Delta \sum_{t \geqslant 1} e^{-c_0\Delta\sqrt{t/8}} \\
&\leqslant \Delta \times \frac{2}{(c_0\Delta/\sqrt{8})^2} && (4.44) \\
&= \frac{16}{c_0^2 \Delta} && (4.45)
\end{aligned}$$

where (4.44) follows from Lemma 4.1. Combining (4.40), (4.39) and (4.45) proves Theorem 4.2 with $c_2 = c_1 + \frac{\sqrt{8}}{c_0}$, $c_3 = \frac{\sqrt{8}}{c_0}$ and $c_4 = \frac{16}{c_0^2}$.

Proof of Corollary 4.1. Define $\tau = \sup\{t \geqslant 0, \exists i \neq i^*, L_{i,t} - L_{i^*,t} \leqslant \frac{\Delta t}{2}\}$. By Lemma 4.2 below, for every $\varepsilon > 0$ we have, with probability at least $1 - \varepsilon$, $\tau \leqslant 8(\log M + \log \varepsilon^{-1})/\Delta^2$. By Theorem 4.2, this implies that, with probability at least $1 - \varepsilon$,

$$\begin{aligned}
R_T &\leqslant c_1\sqrt{\tau \log M} + \frac{c_2 \log M + c_3 \log \Delta^{-1} + c_4}{\Delta/2} \\
&\leqslant \left(c_1\sqrt{8} + 2c_2\right)\frac{\log M}{\Delta} + c_1\frac{\sqrt{8 \log M \log \varepsilon^{-1}}}{\Delta} + 2c_3\frac{\log \Delta^{-1}}{\Delta} + \frac{2c_4}{\Delta}
\end{aligned}$$

where c_2, c_3, c_4 are the constants of Theorem 4.2. The bound (4.11) on the pseudo-regret is obtained similarly from Theorem 4.2, by using the fact that $\mathcal{R}_T \leqslant \mathbb{E}[R_T]$ and

$$\mathbb{E}[\sqrt{\tau \log M}] \leqslant \sqrt{\mathbb{E}[\tau] \log M} \leqslant \sqrt{\log M}\sqrt{1 + \frac{8(\log M + 1)}{\Delta^2}} \leqslant \sqrt{\log M}\Big(1 + \frac{\sqrt{8 \log M} + 1}{\Delta}\Big)$$

which is smaller than $(2 + \sqrt{8})(\log M)/\Delta \leqslant 5(\log M)/\Delta$ since $M \geqslant 3$ and $\Delta \leqslant 1$. □

Lemma 4.2. *Let $(\ell_{i,t})_{1 \leqslant i \leqslant M, t \geqslant 1}$ be as in Theorem 4.1. Denote $\tau = \sup\{t \geqslant 0, \exists i \neq i^*, L_{i,t} - L_{i^*,t} \leqslant \frac{\Delta t}{2}\}$. We have*

$$\mathbb{E}[\tau] \leqslant 1 + \frac{8(\log M + 1)}{\Delta^2}\,, \tag{4.46}$$

and for every $\varepsilon \in (0,1)$,

$$\mathbb{P}\Big(\tau \geqslant \frac{8(\log M + \log \varepsilon^{-1})}{\Delta^2}\Big) \leqslant \varepsilon\,. \tag{4.47}$$

Proof of Lemma 4.2. For every $i \neq i^*$ and $t \geqslant 1$, let $\Delta_{i,t} := \mathbb{E}[\ell_{i,t} - \ell_{i^*,t} | \mathcal{F}_{t-1}]$. Using the Hoeffding-Azuma's maximal inequality to the $(\mathcal{F}_t)_{t \geqslant 1}$-martingale difference sequence $Z_{i,t} = -(L_{i,t} - L_{i^*,t}) + \Delta_{i,t}$ (such that $\Delta_{i,t} - 1 \leqslant Z_{i,t} \leqslant \Delta_{i,t} + 1$), together with the fact that $\Delta_{i,t} \geqslant \Delta$, implies that

$$\mathbb{P}\left(\exists t \geqslant t_0, L_{i,t} - L_{i^*,t} \leqslant \frac{\Delta t}{2}\right) \leqslant \mathbb{P}\left(\sup_{t \geqslant t_0} \frac{1}{t}\left(\sum_{s=1}^{t} Z_{i,s}\right) \geqslant \frac{\Delta}{2}\right) \leqslant e^{-t_0 \Delta^2/8}\,. \tag{4.48}$$

By a union bound, equation (4.48) implies that

$$\mathbb{P}\left(\tau \geqslant t_0\right) \leqslant M e^{-t_0 \Delta^2/8}\,. \tag{4.49}$$

Solving for the probability level in (4.49) yields the high probability bound (4.47) on τ. The bound on τ in expectation (4.46) ensues by integrating the high-probability bound over ε. □

We recall Hoeffding-Azuma's maximal inequality for bounded martingale difference sequences (Hoeffding, 1963; Azuma, 1967). While it follows from a standard argument, we provide a short proof for completeness, since the inequality given in Proposition 4.5 below differs slightly from the one given in Hoeffding (1963).

Proposition 4.5 (Hoeffding-Azuma's maximal inequality)**.** *Let $(Z_t)_{t \geqslant 1}$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_t)_{t \geqslant 1}$. Assume that Z_t is a martingale difference sequence: $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$ for any $t \geqslant 1$, and that $A_t - 1 \leqslant Z_t \leqslant A_t + 1$ almost surely, where A_t is $\mathcal{F}_{t-1}$-measurable. Then, denoting $S_n := \sum_{t=1}^{n} Z_t$, we have for every $n \geqslant 1$ and $a \geqslant 0$:*

$$\mathbb{P}\left(\sup_{m \geqslant n} \frac{S_m}{m} \geqslant a\right) \leqslant e^{-na^2/2}\,. \tag{4.50}$$

Proof. Fix $\lambda > 0$. By Hoeffding's inequality, $\mathbb{E}[e^{\lambda Z_t} | \mathcal{F}_{t-1}] \leqslant e^{\lambda^2/2}$, so that the sequence $M_t^\lambda := \exp\left(\lambda S_t - \lambda^2 t/2\right)$ is a positive supermartingale. Hence, Doob's supermartingale inequality implies that for $\varepsilon \in (0,1]$:

$$\mathbb{P}\Big(\sup_{t \geqslant 1} M_t^\lambda \geqslant \frac{1}{\varepsilon}\Big) \leqslant \frac{\mathbb{E}[M_0^\lambda]}{1/\varepsilon} = \varepsilon\,. \tag{4.51}$$

Rearranging (4.51) and letting $\lambda = \sqrt{2\log(1/\varepsilon)/n}$ yields: with probability $1-\varepsilon$, for every $t \geqslant n$,

$$\frac{S_t}{t} \leqslant \frac{\log(1/\varepsilon)}{\lambda t} + \frac{\lambda}{2} = \sqrt{\frac{\log(1/\varepsilon)}{2}}\left(\frac{\sqrt{n}}{t} + \frac{1}{\sqrt{t}}\right) \leqslant \sqrt{\frac{2\log(1/\varepsilon)}{n}}\,. \tag{4.52}$$

Setting $\varepsilon = e^{-na^2/2}$ in (4.52) gives the desired bound. □

4.7.4 Proof of Proposition 4.3

Note that, since the loss vectors $\boldsymbol{\ell}_t$ are in fact deterministic, $\mathcal{R}_T = R_T$. Denoting $(v_{i,t})_{1\leqslant i\leqslant M}$ the weights selected by the Constant Hedge algorithm at time t, and letting $c = c_0\sqrt{\log M}$, we have

$$\begin{aligned} R_T &= \sum_{t=1}^{T}\sum_{i=2}^{M} v_{i,t}(\ell_{i,t} - \ell_{1,t}) \\ &= \sum_{t=1}^{T}\sum_{i=2}^{M} \frac{\exp\big(-\frac{c}{\sqrt{T}}(L_{i,t-1} - L_{1,t-1})\big)}{1 + \sum_{2\leqslant i'\leqslant M}\exp\big(-\frac{c}{\sqrt{T}}(L_{i',t-1} - L_{1,t-1})\big)} \\ &= \sum_{t=1}^{T} \frac{(M-1)\exp\big(-\frac{c}{\sqrt{T}}(t-1)\big)}{1 + (M-1)\exp\big(-\frac{c}{\sqrt{T}}(t-1)\big)}\,. \end{aligned} \tag{4.53}$$

Now, let $t_0 \geqslant 0$ be the largest integer such that $(M-1)\exp(-\frac{c}{\sqrt{T}}t) \geqslant 1/2$, namely

$$t_0 = \left\lfloor \frac{\sqrt{T}}{c}\log(2(M-1)) \right\rfloor.$$

It follows from Equation (4.53) that

$$R_T \geqslant \sum_{t=1}^{T\wedge(t_0+1)} \frac{(M-1)\exp\big(-\frac{c}{\sqrt{T}}(t-1)\big)}{1 + (M-1)\exp\big(-\frac{c}{\sqrt{T}}(t-1)\big)} \geqslant \frac{1}{3}\min(T, t_0+1) \tag{4.54}$$

where the second inequality comes from the fact that $\frac{x}{1+x} \geqslant \frac{1}{3}$ for $x \geqslant \frac{1}{2}$, which we apply to $x = (M-1)\exp(-\frac{c}{\sqrt{T}}(t-1)) \geqslant \frac{1}{2}$ for $t \leqslant T \wedge (t_0+1) \leqslant t_0+1$. In order to establish inequality (4.13), it remains to note that

$$t_0 + 1 \geqslant \frac{\sqrt{T}}{c}\log\big(2(M-1)\big) \geqslant \frac{\sqrt{T\log M}}{c_0}\,,$$

since $2(M-1) \geqslant M$ and $c = \sqrt{c_0 \log M}$.

Now, consider the Hedge algorithm with doubling trick. Assume that $T \geqslant 2$, and let $k \geqslant 1$ such that $T_k \leqslant T < T_{k+1}$. Since $R_T = \sum_{t=1}^{T}\sum_{2\leqslant i\leqslant M} v_{i,t}(\ell_{i,t} - \ell_{1,t})$ and each of the terms in the sum is nonnegative, R_T is lower bounded by the cumulative regret on the period $[\![T_{k-1}, T_k - 1]\!]$. During this period of length T_{k-1}, the algorithm reduces to the Hedge algorithm with constant learning rate $c_0\sqrt{\log(M)/T_{k-1}}$, so that the above bound (4.13) applies; further bounding $T_{k-1} \geqslant \frac{T}{4}$ establishes (4.14).

4.7.5 Proof of Proposition 4.4

By convexity of $x \mapsto x^2$ and concavity of $x \mapsto x^\beta$, we have:

$$\begin{aligned}
\mathbb{E}[(\widehat{\ell}_t - \ell_{i^*,t})^2] &\leqslant \mathbb{E}\left[\sum_{i=1}^M v_{i,t}(\ell_{i,t} - \ell_{i^*,t})^2\right] && (4.55)\\
&= \mathbb{E}\left[\sum_{i=1}^M v_{i,t}\mathbb{E}\left[(\ell_{i,t} - \ell_{i^*,t})^2|\mathcal{F}_{t-1}\right]\right]\\
&\leqslant B\mathbb{E}\left[\sum_{i=1}^M v_{i,t}\mathbb{E}\left[\ell_{i,t} - \ell_{i^*,t}|\mathcal{F}_{t-1}\right]^\beta\right] && (4.56)\\
&\leqslant B\mathbb{E}\left[\sum_{i=1}^M v_{i,t}\mathbb{E}\left[\ell_{i,t} - \ell_{i^*,t}|\mathcal{F}_{t-1}\right]\right]^\beta && (4.57)\\
&= B\mathbb{E}[\widehat{\ell}_t - \ell_{i^*,t}]^\beta && (4.58)
\end{aligned}$$

where inequalities (4.55) and (4.57) come from Jensen's inequality, and (4.56) from the Bernstein condition (4.15). Taking the expectation of the regret bound (4.16), we obtain

$$\begin{aligned}
\mathbb{E}[R_{i^*,T}] &\leqslant \mathbb{E}\left[C_1\sqrt{(\log M)\sum_{t=1}^T(\widehat{\ell}_t - \ell_{i^*,t})^2} + C_2\log M\right]\\
&\leqslant C_1\sqrt{(\log M)\sum_{t=1}^T\mathbb{E}\big[(\widehat{\ell}_t - \ell_{i^*,t})^2\big]} + C_2\log M && (4.59)\\
&\leqslant C_1\sqrt{(\log M)B\sum_{t=1}^T\mathbb{E}\big[\widehat{\ell}_t - \ell_{i^*,t}\big]^\beta} + C_2\log M\\
&= C_1\sqrt{BT\log M}\left(\frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[\widehat{\ell}_t - \ell_{i^*,t}\big]^\beta\right)^{1/2} + C_2\log M\\
&\leqslant C_1\sqrt{BT\log M}\left(\frac{\mathbb{E}[R_{i^*,T}]}{T}\right)^{\beta/2} + C_2\log M && (4.60)
\end{aligned}$$

where inequalities (4.59) and (4.60) come from Jensen's inequality. Letting $r = \mathbb{E}[R_{i^*,T}]/T$ and $u = (\log M)/T$, inequality (4.60) writes $r \leqslant C_1\sqrt{Bu}r^{\beta/2} + C_2u$. This implies that (depending on which of these two terms is larger) either $r \leqslant 2C_2u$, or $r \leqslant 2C_1\sqrt{Bu}r^{\beta/2}$, and the latter condition amounts to $r \leqslant (2C_1)^{2/(2-\beta)}(Bu)^{1/(2-\beta)}$. This entails that

$$r \leqslant (2C_1)^{\frac{2}{2-\beta}}(Bu)^{\frac{1}{2-\beta}} + 2C_2u\,,$$

which amounts to

$$\mathbb{E}[R_{i^*,T}] \leqslant C_3(B\log M)^{\frac{1}{2-\beta}}T^{\frac{1-\beta}{2-\beta}} + C_4\log M \qquad (4.61)$$

where $C_3 = (2C_1)^{2/(2-\beta)} \leqslant \max(1, 4C_1^2)$ and $C_4 = 2C_2$.

4.7.6 Proof of Theorem 4.3

Consider the constant losses $\ell_{1,t} = 0$, $\ell_{i,t} = \Delta$ where $\Delta = 1 \wedge c_0^{-1}\sqrt{(\log M)/T}$. These losses satisfy the $(1,1)$-Bernstein condition since, for every $i > 1$, $\mathbb{E}[(\ell_{i,t} - \ell_{1,t})^2] = \Delta^2 \leqslant \Delta = \mathbb{E}[\ell_{i,t} - \ell_{1,t}]$. On the other hand, the regret of the Hedge algorithm with learning rate $\eta_t = c_0\sqrt{(\log M)/t}$ writes

$$\begin{aligned}
\mathcal{R}_T &= \sum_{t=1}^{T}\sum_{i\neq 1}\mathbb{E}[v_{i,t}(\ell_{i,t} - \ell_{1,t})] \\
&= \Delta\sum_{t=1}^{T}\frac{(M-1)e^{-\eta_t\Delta(t-1)}}{1+(M-1)e^{-\eta_t\Delta(t-1)}} \\
&\geqslant \frac{\Delta}{3}\sum_{t=1}^{T}\mathbf{1}\Big((M-1)e^{-\eta_t\Delta(t-1)} \geqslant \frac{1}{2}\Big) \\
&\geqslant \frac{\Delta}{3}\sum_{t=1}^{T}\mathbf{1}\left(Me^{-c_0\Delta\sqrt{(t-1)\log M}} \geqslant 1\right) \qquad (4.62) \\
&\geqslant \frac{\Delta}{3}\times\min\left(\frac{\log M}{c_0^2\Delta^2}, T\right) \\
&= \frac{1}{3}\min\left(\frac{1}{c_0}\sqrt{T\log M}, T\right), \qquad (4.63)
\end{aligned}$$

where (4.62) relies on the inequalities $2(M-1) \geqslant M$ and $(t-1)/\sqrt{t} \leqslant \sqrt{t-1}$ for $M \geqslant 2, t \geqslant 1$, while (4.63) is obtained by noting that $(\log M)/(c_0^2\Delta^2) \geqslant T$ since $\Delta \leqslant c_0^{-1}\sqrt{(\log M)/T}$ and substituting for Δ.

4.7.7 Proof of Theorem 4.4

Assume that the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots$ are i.i.d., and denote $i^* = \arg\min_{1\leqslant i\leqslant M}\mathbb{E}[\ell_{i,t}]$ (which is assumed to be unique), $\Delta = \min_{i\neq i^*}\Delta_i > 0$ where $\Delta_i = \mathbb{E}[\ell_{i,t} - \ell_{i^*,t}]$ and $j \in \{1,\ldots,M\}$ such that $\Delta_j = \Delta$. The Decreasing Hedge algorithm with learning rate $\eta_t = c_0\sqrt{(\log M)/t}$ satisfies

$$\begin{aligned}
\mathcal{R}_T &= \sum_{t=1}^{T}\sum_{i\neq i^*}\mathbb{E}[v_{i,t}]\Delta_i \\
&\geqslant \Delta\sum_{t=1}^{T}\mathbb{E}\left[\frac{\sum_{i\neq i^*}e^{-\eta_t(L_{i,t-1}-L_{i^*,t-1})}}{1+\sum_{i\neq i^*}e^{-\eta_t(L_{i,t-1}-L_{i^*,t-1})}}\right] \\
&\geqslant \Delta\sum_{t=1}^{T}\mathbb{E}\left[\frac{e^{-\eta_t(L_{j,t-1}-L_{i^*,t-1})}}{1+e^{-\eta_t(L_{j,t-1}-L_{i^*,t-1})}}\right] \qquad (4.64) \\
&\geqslant \frac{\Delta}{3}\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left(e^{-\eta_t(L_{j,t-1}-L_{i^*,t-1})} \geqslant \frac{1}{2}\right)\right] \\
&= \frac{\Delta}{3}\sum_{t=1}^{T}\mathbb{P}\left(\eta_t(L_{j,t-1} - L_{i^*,t-1}) \leqslant \log 2\right) \qquad (4.65)
\end{aligned}$$

where (4.64) relies on the fact that the function $x \mapsto \frac{x}{1+x}$ is increasing on $\mathbf{R}^+$. Denoting $a = (\log 2)/(c_0\sqrt{\log M})$, we have for every $1 \leqslant t \leqslant 1 + \frac{a^2}{4\Delta^2}$:

$$\begin{aligned}\mathbb{P}\left(\eta_t(L_{j,t-1} - L_{i^*,t-1}) > \log 2\right) &= \mathbb{P}\left(L_{j,t-1} - L_{i^*,t-1} - \Delta(t-1) > a\sqrt{t} - \Delta(t-1)\right) \\ &\leqslant \mathbb{P}\left(L_{j,t-1} - L_{i^*,t-1} - \Delta(t-1) > \frac{a\sqrt{t-1}}{2}\right) \qquad (4.66) \\ &\leqslant e^{-a^2/8} \qquad (4.67)\end{aligned}$$

where inequality (4.66) stems from the fact that $\Delta(t-1) \leqslant \frac{a\sqrt{t-1}}{2}$ (since $t \leqslant 1 + \frac{a^2}{4\Delta^2}$), while (4.67) is a consequence of Hoeffding's bound applied to the i.i.d. $[-1-\Delta, 1-\Delta]$-valued random variables $\ell_{j,s} - \ell_{i^*,s} - \Delta$, $1 \leqslant s \leqslant t-1$. Assuming that $c_0 \geqslant 1$, we have $a \leqslant \sqrt{\log 2} \leqslant 1$, so that by concavity of the function $x \mapsto 1 - e^{-x/8}$, $1 - e^{-a^2/8} \geqslant (1 - e^{-1/8})a^2$. Combining this with inequalities (4.65) and (4.67) and using the fact that $\left\lfloor 1 + \frac{a^2}{4\Delta^2} \right\rfloor \geqslant \frac{a^2}{4\Delta^2}$, we obtain for $T \geqslant \frac{1}{4\Delta^2} \geqslant \frac{a^2}{4\Delta^2}$:

$$\mathbb{E}\left[R_T\right] \geqslant \frac{\Delta}{3} \min\left(\frac{a^2}{4\Delta^2}, T\right)(1 - e^{-1/8})a^2 = \frac{(1 - e^{-1/8})a^4}{12\Delta} \geqslant \frac{1}{450 c_0^4 (\log M)^2 \Delta}, \qquad (4.68)$$

where the last inequality comes from the fact that $(\log 2)^4(1 - e^{-1/8})/12 \geqslant \frac{1}{450}$.

Chapter 5

Efficient tracking of a growing number of experts

Abstract. We consider a variation on the problem of prediction with expert advice, where new forecasters that were unknown until then may appear at each round. As often in prediction with expert advice, designing an algorithm that achieves near-optimal regret guarantees is straightforward, using aggregation of experts. However, when the comparison class is sufficiently rich, for instance when the best expert and the set of experts itself changes over time, such strategies naively require to maintain a prohibitive number of weights (typically exponential with the time horizon). By contrast, designing strategies that both achieve a near-optimal regret and maintain a reasonable number of weights is highly non-trivial. We consider three increasingly challenging objectives (simple regret, shifting regret and sparse shifting regret) that extend existing notions defined for a fixed expert ensemble; in each case, we design strategies that achieve tight regret bounds, adaptive to the parameters of the comparison class, while being computationally inexpensive. Moreover, our algorithms are anytime, agnostic to the number of incoming experts and completely parameter-free. These results are made possible thanks to two simple but highly effective recipes: first, the "abstention trick" that comes from the *specialist* framework and enables to handle the least challenging notions of regret, but is limited when addressing more sophisticated objectives. Second, the "muting trick" that we introduce to give more flexibility. We show how to combine these two tricks in order to handle the most challenging class of comparison strategies.

Contents

5.1 Introduction

Aggregation of experts is a well-established framework in machine learning (Cesa-Bianchi and Lugosi, 2006; Vovk, 1998; Haussler et al., 1998), that provides a sound strategy to combine the forecasts of many different sources. This is classically considered in the sequential prediction setting, where at each time step, a learner receives the predictions of experts, uses them to provide his own forecast, and then observes the true value of the signal, which determines his loss and those of the experts. The goal is then to minimize the *regret* of the learner, which is defined as the difference between his cumulated loss and that of the best expert (or combination thereof), no matter what the experts' predictions or the values of the signal are.

A standard assumption in the existing literature is that the set of experts is known before the beginning of the game. In many situations, however, it is desirable to add more and more forecasters over time. For instance, in a non-stationary setting one could add new experts trained on a fraction of the signal, possibly combined with change point detection. Even in a stationary setting, a growing number of increasingly complex models enables to account for increasingly subtle properties of the signal without having to include them from the start, which can be needlessly costly computationally (as complex models, which take more time to fit, are not helpful in the first rounds) or even intractable in the case of an infinite number of models with no closed form expression. Additionally, in many realistic situations some completely novel experts may appear in an unpredicted way (possibly due to innovation, the discovery of better algorithms or the availability of new data), and one would want a way to safely incorporate them to the aggregation procedure.

In this chapter, we study how to amend aggregation of experts strategies in order to incorporate novel experts that may be added on the fly at any time step. Importantly, since we do not know in advance when new experts are made available, we put a strong emphasis on *anytime* strategies, that do not assume the time horizon is finite and known. Likewise, our algorithms should be agnostic to the total number of experts available at a given time. Three notions of regret of increasing complexity will be defined for growing expert sets, that extend existing notions to a growing expert set. Besides comparing against the best expert, it is natural in a growing experts setting to track the best expert; furthermore, when the number of experts gets large, it becomes profitable to track the best expert in a small pool of good experts. For each notion, we propose corresponding algorithms with tight regret bounds. As is often the case in structured aggregation of experts, the key difficulty is typically not to derive the regret bounds, but to obtain efficient algorithms. All our methods exhibit minimal time and space requirements that are linear in the number of present experts.

Related work. This work builds on the setting of prediction with expert advice (Cesa-Bianchi and Lugosi, 2006; Vovk, 1998; Herbster and Warmuth, 1998) that originates from the work on universal prediction (Ryabko, 1988; Merhav and Feder, 1998; Györfi et al., 1999). We make use of the notion of *specialists* (Freund et al., 1997; Chernov and Vovk, 2009) and its application to *sleeping experts* (Koolen et al., 2012), as well as the corresponding standard extensions (Fixed Share, Mixing Past Posteriors) of basic strategies to the problem of *tracking the best expert* (Herbster and Warmuth, 1998; Koolen and de Rooij, 2013; Bousquet and Warmuth, 2002); see also Willems (1996); Shamir and Merhav (1999) for related work in

the context of lossless compression. Note that, due to its versatility, aggregation of experts has been adapted successfully to a number of applications (Monteleoni et al., 2011; McQuade and Monteleoni, 2012; Stoltz, 2010). It should be noted that the literature on prediction with expert advice is split in two categories: the first one focuses on exp-concave loss functions, whereas the second studies convex bounded losses. While our work belongs to the first category, it should be possible to transport our regret bounds to the convex bounded case by using time-varying learning rates, as done e.g. by Hazan and Seshadhri (2009) and Gyorgy et al. (2012). In this case, the growing body of work on the automatic tuning of the learning rate (de Rooij et al., 2014; Koolen et al., 2014) as well as alternative aggregation schemes (Wintenberger, 2017; Koolen and van Erven, 2015; Luo and Schapire, 2015) might open the path for even further improvements.

The use of a growing expert ensemble was already proposed by Györfi et al. (1999) in the context of sequentially predicting an ergodic stationary time series, where new higher order Markov experts were introduced at exponentially increasing times (and the weights were reset as uniform); since consistency was the core focus of the chapter, this simple "doubling trick" could be used, something we cannot afford when new experts arrive more regularly. Closer to our approach, growing expert ensembles have been considered in contexts where the underlying signal may be non-stationary, see e.g. Hazan and Seshadhri (2009); Shalizi et al. (2011). Of special interest to our problem is Shalizi et al. (2011), which considers the particular case when one new expert is introduced every τ time steps, and propose a variant of the Fixed Share (FS) algorithm analogous to our GrowingMarkovHedge algorithm. However, their algorithms depend on parameters which have to be tuned depending on the parameters of the comparison class, whereas our algorithms are parameter-free and do not assume the prior knowledge of the comparison class. Moreover, we introduce several other algorithms tailored to different notions of regret; in particular, we address the problem of comparing to sequences of experts that alternate between a small number of experts, a refinement that is crucial when the total set of experts grows, and has not been obtained previously in this context.

Another related setting is that of "branching experts" considered by Gofer et al. (2013), where each incumbent expert is split into several experts that may diverge later on. Their results include a regret bound in terms of the number of *leading experts* (whose cumulated loss was minimal at some point). Our approach differs in that it does not assume such a tree-like structure: a new entering forecaster is not assumed to be associated to an incumbent expert. More importantly, while Gofer et al. (2013) compare to the leaders in terms of cumulated loss (since the beginning of the game), our methods compete instead with sequences of experts that perform well on some periods, but can predict arbitrarily bad on others; this is harder, since the loss of the optimal sequence of experts can be significantly smaller than that of the best expert.

Outline. This chapter is organized as follows. After introducing the setting, notations and the different comparison classes, we provide in Section 5.2 an overview of our results, stated in less general but more interpretable forms. Then, Section 5.3 introduces the exponential weights algorithm and its regret, a classical preliminary result that will be used throughout this chapter. Sections 5.4, 5.5 and 5.6 form the core of this chapter, and have the same structure: a generic result is first stated in the case of a fixed set of experts, before being turned into a strategy in the growing experts framework. Section 5.4 starts with the related *specialist* setting and adapts the algorithm into an anytime growing experts algorithm, with a more

general formulation and regret bound involving *unnormalized priors*. Section 5.5 proposes an alternative approach, which casts the growing experts problem as one of competing against *sequences* of experts; this approach proves more flexible and general for our task, but perhaps surprisingly we can also recover algorithms that are essentially equivalent to the aggregation of growing experts with an unnormalized prior. Finally, the two approaches are combined in Section 5.6 in the context of *sleeping experts*, where we reinterpret the algorithm of Koolen et al. (2012) and extend it to more general priors before adapting it to the growing experts setting.

5.2 Overview of results

Our work is framed in the classical setting of *prediction with expert advice* (Vovk, 1998; Cesa-Bianchi and Lugosi, 2006), which we adapt to account for a growing number of experts. The problem is characterized by its *loss function* $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$, where $\widehat{\mathcal{Y}}$ is a convex *prediction space*, and $\mathcal{Y}$ is the *signal* or *output space*.

Let M_t be the total number of experts at time t, and $m_t = M_t - M_{t-1}$ be the number of experts introduced at time t. We index experts by their entry order, so that expert i is the i^{th} introduced expert and denote $\tau_i = \min\{t \geqslant 1 : i \leqslant M_t\}$ its *entry time* (the time at which it is introduced). We say we are in the *fixed expert set* case when $M_t = M$ for every $t \geqslant 1$ and in the *growing experts setting* otherwise. At each step $t \geqslant 1$, the experts $i = 1, \dots, M_t$ output their predictions $\widehat{y}_{i,t} \in \widehat{\mathcal{Y}}$, which the learner uses to build $\widehat{y}_t \in \widehat{\mathcal{Y}}$; then, the environment decides the value of the signal $y_t \in \mathcal{Y}$, which sets the losses $\ell_t = \ell(\widehat{y}_t, y_t)$ of the learner and $\ell_{i,t} = \ell(\widehat{y}_{i,t}, y_t)$ of the experts.

Notations. Let $\mathcal{P}_M$ be the *probability simplex*, *i.e.* the set of probability measures over the set of experts $\{1, \dots, M\}$. We denote by $\mathrm{KL}(\cdot, \cdot)$ the *Kullback-Leibler divergence*, defined for $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{P}_M$ by $\mathrm{KL}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{M} u_i \log \frac{u_i}{v_i} \geqslant 0$.

Loss function. Throughout this text, we make the following standard assumption[1] on the loss function (Cesa-Bianchi and Lugosi, 2006).

Assumption 5.1. The loss function ℓ is η-*exp-concave* for some $\eta > 0$, in the sense that $\exp(-\eta\, \ell(\cdot, y))$ is concave on $\widehat{\mathcal{Y}}$ for every observation $y \in \mathcal{Y}$. This is equivalent to the inequality

$$\ell\left(\sum_{i=1}^{M} v_i\, \widehat{y}_i, y\right) \leqslant -\frac{1}{\eta} \log \sum_{i=1}^{M} v_i\, e^{-\eta\, \ell(\widehat{y}_i, y)} \tag{5.1}$$

for every $y \in \mathcal{Y}$, $\widehat{\boldsymbol{y}} = (\widehat{y}_i)_{1 \leqslant i \leqslant M} \in \widehat{\mathcal{Y}}^M$ and $\boldsymbol{v} = (v_i)_{1 \leqslant i \leqslant M} \in \mathcal{P}_M$.

Remark 5.1. An important example in the case when $\widehat{\mathcal{Y}}$ is the set of probability measures over $\mathcal{Y}$ is the *logarithmic* or *self-information* loss $\ell(\widehat{y}, y) = -\log \widehat{y}(\{y\})$ for which the inequality holds with $\eta = 1$, and is actually an equality. Another example of special interest is the quadratic loss on a bounded interval: indeed, for $\widehat{\mathcal{Y}} = \mathcal{Y} = [a, b] \subset \mathbf{R}$, $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ is $\frac{1}{2(b-a)^2}$-exp-concave.

[1] This could be readily replaced (up to some cosmetic changes in the statements and their proofs) by the more general η-*mixability* condition (Vovk, 1998), which allows to use higher learning rates η for some losses (such as the square loss, but not the logarithmic loss) by using more sophisticated combination functions.

Several notions of regret can be considered in the growing expert setting. We review here three of them, each corresponding to a specific comparison class; we show the kind of bounds that our algorithms achieve, to illustrate the more general results stated in the subsequent sections. We provide more uniform bounds in Section 5.7.5, and compare them with information-theoretic bounds.

Constant experts. Since the experts only output predictions after their entry time, it is natural to consider the *regret* with respect to each expert $i \geqslant 1$ over its time of activity, namely the quantity

$$\sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t}) \tag{5.2}$$

for every $T \geqslant \tau_i$. Note that this is equivalent to controlling (5.2) for every $T \geqslant 1$ and $i \leqslant M_T$. Algorithm GrowingHedge is particularly relevant in this context; with the choice of (unnormalized) prior weights $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$, it achieves the following regret bound: for every $T \geqslant 1$ and $i \leqslant M_T$,

$$\sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log m_{\tau_i} + \frac{1}{\eta} \log \tau_i + \frac{1}{\eta} \log(1 + \log T)\,. \tag{5.3}$$

This bound has the merit of being simple, virtually independent of T and independent of the number of experts $(m_t)_{t>\tau_i}$ added after i. Several other instantiations of the general regret bound of GrowingHedge (Theorem 5.1) are given in Section 5.4.2.

Sequences of experts. Another way to study growing expert sets is to view them through the lens of sequences of experts. Given a sequence of experts $i^T = (i_1, \ldots, i_T)$, we measure the performance of a learning algorithm against it in terms of the *cumulative regret*:

$$L_T - L_T(i^T) = \sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t}\,. \tag{5.4}$$

In order to derive meaningful regret bounds, some constraints have to be imposed on the comparison sequence; hence, we consider in the sequel different types of comparison classes that lead to different notions of regret, from the least to the most challenging one:

(a) Sequences of fresh experts. These are *admissible* sequences of experts i^T, in the sense that $i_t \leqslant M_t$ for $1 \leqslant t \leqslant T$ (so that $\ell_{i_t,t}$ is always well-defined) that only switch to *fresh* (newly entered) experts, *i.e.* if $i_t \neq i_{t-1}$, then $M_{t-1} + 1 \leqslant i_t \leqslant M_t$. More precisely, for each $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$ with $1 < \sigma_1 < \cdots < \sigma_k \leqslant T$, $\mathcal{S}_T^{(f)}(\boldsymbol{\sigma})$ denotes the set of sequences of fresh experts whose only shifts occur at times $\sigma_1, \ldots, \sigma_k$. Both the switch times $\boldsymbol{\sigma}$ and the number of shifts k are assumed to be unknown, although to obtain controlled regret one typically needs $k \ll T$. Comparing to sequences of fresh experts is essentially equivalent to comparing against constant experts; algorithms GrowingHedge and FreshMarkovHedge with $\pi_i = \frac{1}{m_{\tau_i}}$ achieve, for *every* $T \geqslant 1$, $k \leqslant T-1$ and $\boldsymbol{\sigma} = (\sigma_j)_{1 \leqslant j \leqslant k}$ (Theorems 5.1 and 5.2):

$$L_T - \inf_{i^T \in \mathcal{S}_T^{(f)}(\boldsymbol{\sigma})} L_T(i^T) \leqslant \frac{1}{\eta} \left\{ \log m_1 + \sum_{j=1}^{k} (\log m_{\sigma_j} + \log \sigma_j) + \log T \right\}. \tag{5.5}$$

In particular, the regret with respect to any sequence of fresh experts with k shifts is bounded by

$$\frac{1}{\eta}\left((k+1)\log\max_{1\leqslant t\leqslant T} m_t + (k+1)\log T\right).$$

(b) Arbitrary admissible sequences of experts. Like before, these are admissible sequences of experts that are piecewise constant with a typically small number of shifts k, except that shifts to *incumbent* (previously introduced) experts $i_t \leqslant M_{t-1}$ are now authorized. Specifically, given $\boldsymbol{\sigma}^0 = (\sigma^0_1, \dots, \sigma^0_{k_0})$ and $\boldsymbol{\sigma}^1 = (\sigma^1_1, \dots, \sigma^1_{k_1})$, we denote by $\mathcal{S}_T^{(a)}(\boldsymbol{\sigma}^0; \boldsymbol{\sigma}^1)$ the class of admissible sequences whose switches to fresh (resp. incumbent) experts occur only at times $\sigma^0_1 < \dots < \sigma^0_{k_0}$ (resp. $\sigma^1_1 < \dots < \sigma^1_{k_1}$). By Theorem 5.3, algorithm GrowingMarkovHedge with $\pi_i = \frac{1}{m_{\tau_i}}$ and $\alpha_t = \frac{1}{t}$ satisfies, for *every* $T \geqslant 1$, k_0, k_1 with $k_0 + k_1 \leqslant T-1$ and $\boldsymbol{\sigma}^0, \boldsymbol{\sigma}^1$:

$$L_T - \inf_{i^T \in \mathcal{S}_T^{(a)}(\boldsymbol{\sigma}^0;\boldsymbol{\sigma}^1)} L_T(i^T) \leqslant \frac{1}{\eta}\left\{\log m_1 + \sum_{j=1}^{k}(\log m_{\sigma_j} + \log \sigma_j) + \sum_{j=1}^{k_1}\log \sigma^1_j + 2\log T\right\} \quad (5.6)$$

where $k = k_0 + k_1$ and $\sigma_1 < \dots < \sigma_k$ denote *all* shifts (either in $\boldsymbol{\sigma}^0$ or in $\boldsymbol{\sigma}^1$). Note that the upper bound (5.6) may be further relaxed as

$$\frac{1}{\eta}\left((k+1)\log\max_{1\leqslant t\leqslant T} m_t + (k_0 + 2k_1 + 2)\log T\right).$$

(c) Sparse sequences of experts. These are admissible sequences i^T of experts that are additionally *sparse*, in the sense that they alternate between a small number $n \ll M_T$ of experts; again, n may be unknown in advance. Denoting $\mathcal{S}_T^{(s)}(\boldsymbol{\sigma}, E)$ the class of sequences with shifts in $\boldsymbol{\sigma}$ and taking values in the subset of experts $E = \{e_1, \dots, e_n\}$, algorithm GrowingSleepingMarkovHedge with $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$ and $\alpha_t = \beta_t = \frac{1}{t}$ achieves, for every $T \geqslant 1$, $E \subset \{1, \dots, M_T\}$ and $\boldsymbol{\sigma}$,

$$L_T - \inf_{i^T \in \mathcal{S}_T^{(s)}(\boldsymbol{\sigma},E)} L_T(i^T) \leqslant \frac{1}{\eta}\sum_{p=1}^{n}\left(\log \tau_{e_p} + \log\frac{m_{\tau_{e_p}}}{n}\right) + \frac{1}{\eta}n\log(2T) + \frac{2}{\eta}\sum_{j=1}^{k}\log\sigma_j. \quad (5.7)$$

In particular, the regret with respect to every admissible sequence of T experts with at most k shifts and taking at most n values is bounded by

$$\frac{1}{\eta}\left(n\log\frac{\max_{1\leqslant t\leqslant T} m_t}{n} + 2n\log(\sqrt{2}T) + 2k\log T\right).$$

The main results of this text are Theorem 5.3, a powerful parameter-free generalization of Shalizi et al. (2011, Theorem 2), and Theorem 5.4, which adapts results of Bousquet and Warmuth (2002); Koolen et al. (2012) to sequentially incoming forecasters, and has no precedent in this context.

5.3 Preliminary: the exponential weights algorithm

First, we introduce the simple but fundamental *exponential weights* or *Hedge algorithm* (Vovk, 1998; Cesa-Bianchi and Lugosi, 2006), designed to control the regret $L_T - L_{i,T} = \sum_{t=1}^{T}\ell_t -$

$\sum_{t=1}^{T} \ell_{i,t}$ for a fixed set of experts $\{1, \dots, M\}$. The algorithm depends on a *prior distribution* $\boldsymbol{\pi} \in \mathcal{P}_M$ on the experts and predicts as

$$\widehat{y}_t = \frac{\sum_{i=1}^{M} w_{i,t}\, \widehat{y}_{i,t}}{\sum_{i=1}^{M} w_{i,t}} \quad \text{with} \quad w_{i,t} = \pi_i\, e^{-\eta L_{i,t-1}}\,. \tag{5.8}$$

Equivalently, it forecasts $\widehat{y}_t = \sum_{i=1}^{M} v_{i,t}\, \widehat{y}_{i,t}$, where the weights $\boldsymbol{v}_t \in \mathcal{P}_M$ are sequentially updated in the following way: $\boldsymbol{v}_1 = \boldsymbol{\pi}$ and, after each round $t \geqslant 1$, $\boldsymbol{v}_{t+1}$ is set to the *posterior* distribution $\boldsymbol{v}_t^m$ of $\boldsymbol{v}_t$ given the losses $(\ell_{i,t})_{1 \leqslant i \leqslant M}$, defined by

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^{M} v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,. \tag{5.9}$$

All subsequent regret bounds will rely on the following standard regret bound (see Section 5.7.1), by reducing complex forecasting strategies to the aggregation of experts under a suitable prior.

Proposition 5.1 (Cesa-Bianchi and Lugosi (2006, Corollary 3.1))**.** *Irrespective of the values of the signal and the experts' predictions, the exponential weights algorithm* (5.8) *with prior* $\boldsymbol{\pi}$ *achieves*

$$L_T - L_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i} \tag{5.10}$$

for each $i = 1, \dots, M$ *and* $T \geqslant 1$*. More generally, for each probability vector* $\boldsymbol{\rho} \in \mathcal{P}_M$*,*

$$L_T - \sum_{i=1}^{M} \rho_i\, L_{i,T} \leqslant \frac{1}{\eta}\, \mathrm{KL}(\boldsymbol{\rho}, \boldsymbol{\pi})\,. \tag{5.11}$$

Choosing a uniform prior $\boldsymbol{\pi} = \frac{1}{M}\mathbf{1}$ yields a $\frac{1}{\eta} \log M$ regret bound with respect to the best expert.

5.4 Growing experts and specialists: the "abstention trick"

A natural idea to tackle the problem of a growing number of experts is to cast it in the related setting of *specialists*, introduced by Freund et al. (1997). We present the specialist setting and the related "specialist trick" identified by Chernov and Vovk (2009) (which we will call the "abstention trick"), which enables to convert any expert aggregation algorithm into a specialist aggregation algorithm. These ideas are then applied to the growing expert ensemble setting, which allows us to control the regret with respect to *constant* experts of equation (5.2); a refinement is introduced along the way, the use of *unnormalized priors*, that gives more flexibility to the algorithm and its regret bounds.

5.4.1 Specialists and their aggregation

In the specialist setting, we have access to *specialists* $i \in \{1, \dots, M\}$ that only output predictions at certain steps, while refraining from predicting the rest of the time. In other words, at each step $t \geqslant 1$, only a subset $A_t \subset \{1, \dots, M\}$ of *active* experts output a prediction $\widehat{y}_{i,t} \in \widehat{\mathcal{Y}}$.

In order to adapt any expert aggregation strategy to the specialists setting, a crucial idea due to Chernov and Vovk (2009) is to "complete" the specialists' predictions by attributing

to inactive specialists $i \notin A_t$ a forecast equal to that of the aggregating algorithm. Although this seems circular, it can be made precise by observing that the only way to simultaneously satisfy the conditions

$$\widehat{y}_t = \sum_{i=1}^{M} v_{i,t}\, \widehat{y}_{i,t} \quad \text{and} \quad \widehat{y}_{i,t} = \widehat{y}_t \ \text{ for any } i \notin A_t \tag{5.12}$$

is to take

$$\widehat{y}_t = \widehat{y}_{i,t} = \frac{\sum_{i \in A_t} v_{i,t}\, \widehat{y}_{i,t}}{\sum_{i \in A_t} v_{i,t}} \quad \text{for } i \notin A_t\,. \tag{5.13}$$

We call this technique the "abstention trick", since it consists in attributing to inactive specialists a forecast that will not affect the voting outcome. In the case of the exponential weights algorithm, this leads to the *specialist aggregation* algorithm with prior $\boldsymbol{\pi}$, which forecasts

$$\widehat{y}_t = \frac{\sum_{i \in A_t} w_{i,t}\, \widehat{y}_{i,t}}{\sum_{i \in A_t} w_{i,t}} \quad \text{with} \quad w_{i,t} = \pi_i\, e^{-\eta L_{i,t-1}}\,, \tag{5.14}$$

where we denote, for each specialist i and $t \geqslant 1$, $L_{i,t} := \sum_{s \leqslant t\,:\, i \in A_s} \ell_{i,s} + \sum_{s \leqslant t\,:\, i \notin A_s} \ell_s$.

Remark 5.2. The exp-concavity inequality $e^{-\eta\, \ell_t} \geqslant \sum_{i=1}^{M} v_{i,t}\, e^{-\eta\, \ell_{i,t}}$ shows that $v_{i,t+1} \geqslant v_{i,t}$ for any $i \notin A_t$. In the case of the logarithmic loss, for $\eta = 1$ this inequality becomes an equality, thus the weights of inactive specialists remain unchanged: $v_{i,t+1} = v_{i,t}$.

Since the specialist aggregation consists of the exponential weights on the extended predictions (5.13), and since for this extension one has $\sum_{t=1}^{T} (\ell_t - \ell_{i,t}) = \sum_{t \leqslant T\,:\, i \in A_t} (\ell_t - \ell_{i,t})$, Proposition 5.1 implies:

Proposition 5.2 (Freund et al. (1997, Theorem 1))**.** *The specialist aggregation with prior $\boldsymbol{\pi}$ achieves the following regret bound: for each specialist i and every $T \geqslant 1$,*

$$\sum_{t \leqslant T\,:\, i \in A_t} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i}\,. \tag{5.15}$$

Moreover, for each probability vector $\boldsymbol{\rho} \in \mathcal{P}_M$,

$$\sum_{i=1}^{M} \rho_i \sum_{t \leqslant T\,:\, i \in A_t} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta}\, \mathrm{KL}(\boldsymbol{\rho}, \boldsymbol{\pi})\,.$$

Remark 5.3. Note that the sets A_t of active specialists need not be known in advance.

5.4.2 Adaptation to growing expert ensembles: GrowingHedge

Growing experts can naturally be seen as specialists, by setting $A_t := \{1, \dots, M_t\}$; moreover, through this equivalence, the quantity controlled by Proposition 5.2 is precisely the regret (5.2) with respect to *constant experts*. In order to apply the results on specialist aggregation to the growing expert setting, it remains to specify exactly which total set of specialists is considered.

Fixed time horizon. In the simplest case when both the time horizon T and the eventual number of experts M_T are known, the eventual set of experts (at time T) is known, and we can take the finite specialist set to be $\{1, \dots, M_T\}$. Therefore, given any probability vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{M_T})$, we can use the aggregation of specialists, with the regret bound (5.15). In particular, the choice of $\pi_i = \frac{1}{M_T}$ for $i = 1, \dots, M_T$ yields the uniform regret bound $\frac{1}{\eta} \log M_T$.

Anytime algorithm, normalized prior. The fixed horizon approach is somewhat unsatisfactory, since we are typically interested in algorithms that are anytime and agnostic to M_t. To achieve this goal, a better choice is to take the infinite set of specialists $\mathbf{N}^*$. Crucially, the aggregation of this infinite number of specialists can be implemented in finite time, by introducing the weight of an expert *only when it enters*. Given a probability vector $\boldsymbol{\pi} = (\pi_i)_{i \geqslant 1}$ on $\mathbf{N}^*$, this leads to the anytime strategy GrowingHedge described below. A straightforward adaptation of Propositions 5.1 and 5.2 to a countably infinite set of experts shows that this strategy achieves, now for *every* $T \geqslant 1$ and $i \leqslant M_T$, the regret bound (5.15). However, we are constrained by the fact that $\boldsymbol{\pi}$ must be a probability on $\mathbf{N}^*$.

Anytime algorithm, unnormalized prior. We now turn to the most general analysis, which improves the previous two. Let $\boldsymbol{\pi} = (\pi_i)_{i \geqslant 1}$ denote a sequence of *arbitrary* positive weights, that are no longer assumed to sum to 1. These weights do not need to be set in advance: the weight π_i can be chosen when expert i enters, so that at this step τ_i, $(m_t)_{t \leqslant \tau_i}$ and $(M_t)_{t \leqslant \tau_i}$ are known, even if they were unknown at the beginning; in particular, π_i may depend on these quantities. We now consider the anytime algorithm GrowingHedge.

Algorithm 8 GrowingHedge — Anytime aggregation of growing experts

1: **Parameters:** Learning rate $\eta > 0$, weights on the experts $\boldsymbol{\pi} = (\pi_i)_{i \geqslant 1}$.
2: **Initialization:** Set $w_{i,1} = \pi_i$ for $i = 1, \dots, M_1$.
3: **for** $t = 1, 2, \dots$ **do**
4: Receive predictions $(\widehat{y}_{1,t}, \dots, \widehat{y}_{M_t,t}) \in \widehat{\mathcal{Y}}^{M_t}$ from the experts, and predict

$$\widehat{y}_t = \frac{\sum_{i=1}^{M_t} w_{i,t}\, \widehat{y}_{i,t}}{\sum_{i=1}^{M_t} w_{i,t}}\,. \tag{5.16}$$

5: Observe $y_t \in \mathcal{Y}$, and derive the losses $\ell_t = \ell(\widehat{y}_t, y_t)$ and $\ell_{i,t} = \ell(\widehat{y}_{i,t}, y_t)$.
6: Update the weights by $w_{i,t+1} = w_{i,t}\, e^{-\eta\, \ell_{i,t}}$ for $i = 1, \dots, M_t$. Moreover, introduce the weights $w_{i,t+1} = \pi_i\, e^{-\eta\, L_t}$ for $M_t + 1 \leqslant i \leqslant M_{t+1}$.
7: **end for**

Theorem 5.1. *Let $\boldsymbol{\pi} = (\pi_i)_{i \geqslant 1}$ be an arbitrary sequence of positive weights. Then, algorithm* GrowingHedge *achieves the following regret bound: for every $T \geqslant 1$ and $i \leqslant M_T$,*

$$\sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log\left(\frac{1}{\pi_i} \sum_{j=1}^{M_T} \pi_j\right). \tag{5.17}$$

Additionally, its time and space complexity at each step $t \geqslant 1$ is $O(M_t)$.

Theorem 5.1 is proved in Section 5.7.2. Let us now discuss a few choices of priors, with the corresponding regret bounds (5.17) (omitting the $\frac{1}{\eta}$ factor).

• With $\pi_i = 1$, we get $\log M_T$, but now with an anytime algorithm. Since $\sum_{i=1}^{M} \frac{1}{i} \leqslant 1 + \sum_{i=2}^{M} \int_{i-1}^{i} \frac{dx}{x} = 1 + \log M$, the choice of $\pi_i = \frac{1}{i}$ yields $\log i + \log(1 + \log M_T)$.

• The above bounds depend on the index $i \geqslant 1$, and hence arbitrarily distinguish experts entered at the same time. More natural bounds would only depend on the entry time τ_i,

which is achievable since π_i can be chosen when i enters, and thus depend on τ_i. Setting[2] $\pi_i = \frac{1}{m_{\tau_i}}\nu_{\tau_i}$, where $\boldsymbol{\nu} = (\nu_t)_{t\geqslant 1}$ is a positive sequence set in advance, we get

$$\log m_{\tau_i} + \log\frac{1}{\nu_{\tau_i}} + \log\sum_{t=1}^{T}\nu_t\,. \tag{5.18}$$

Letting $\nu_t = 1$, (5.18) becomes $\log m_{\tau_i} + \log T$, while $\nu_t = \frac{1}{t}$ yields the improved bound $\log m_{\tau_i} + \log\tau_i + \log(1+\log T)$. Note that neither choice is summable.

Regret against sequences of fresh experts. Theorem 5.1 provides a regret bound against any *static* expert, *i.e.* any constant choice of expert, albeit in a growing experts setting. However, this means that the regret is controlled only on the period $[\![\tau_i, T]\!]$ when the expert actually emits predictions. An alternative way to state Theorem 5.1 is in terms of *sequences of fresh experts*. Indeed, Theorem 5.1 implies that, for every sequence of fresh experts i^T with switching times $\sigma_1 < \cdots < \sigma_k$ (with the additional conventions $\sigma_0 := 1$ and $\sigma_{k+1} := T+1$), algorithm GrowingHedge achieves:

$$L_T - L_T(i^T) = \sum_{j=0}^{k}\sum_{t=\sigma_j}^{\sigma_{j+1}-1}(\ell_t - \ell_{i_{\sigma_j}}) \leqslant \frac{1}{\eta}\sum_{j=0}^{k}\log\frac{\Pi_{M_{\sigma_{j+1}-1}}}{\pi_{i_{\sigma_j}}} \tag{5.19}$$

since $\sigma_j = \tau_{i_{\sigma_j}}$, and where we denote $\Pi_M = \sum_{i=1}^{M}\pi_i$ for each $M \geqslant 1$. Taking $\pi_i = 1$, this bound reduces to $\frac{1}{\eta}\sum_{j=0}^{k}\log M_{\sigma_{j+1}-1} \leqslant \frac{1}{\eta}(k+1)\log M_T$. Taking $\pi_i = 1/m_{\tau_i}$, so that $\Pi_{M_t} = t$, and further bounding $\Pi_{M_{\sigma_{j+1}-1}} = \sigma_{j+1} - 1 \leqslant \sigma_{j+1}$ for $0 \leqslant j \leqslant k-1$ and $\Pi_{M_{\sigma_{k+1}-1}} = T$, we recover the bound (5.5) stated in the overview.

5.5 Growing experts and sequences of experts: the "muting trick"

Algorithm GrowingHedge, based on the specialist viewpoint, guarantees good regret bounds against *fresh* sequences of experts and admits an efficient implementation. Instead of comparing only against fresh sequences of experts, it may be preferable to target *arbitrary* admissible sequences of experts, that contain transitions to incumbent experts; this could be beneficial when some experts start predicting well after a few rounds. A natural approach consists in applying the abstention trick to algorithms for a fixed expert set that target arbitrary sequences of experts (such as Fixed Share, see Section 5.7.3). As it turns out, such an approach would require to maintain weights for unentered experts (which may be in unknown, even infinite, number in an anytime setting): the fact that one could obtain an efficient algorithm such as GrowingHedge was specific to the exponential weights algorithm, and does not extend to more sophisticated algorithms that perform weight sharing.

In this section, we adopt a "dual" point of view, which proves more flexible. Indeed, in the growing expert ensemble setting, there are two ways to cope with the fact that some experts' predictions are undefined at each step. The abstention trick amounts to attributing

[2] In fact, this can be slightly refined when $m_t = 0$ for most steps t. In this case, denoting for $t \geqslant 1$: $s(t) = |\{t' \leqslant t \mid m_{t'} \geqslant 1\}|$, we can take $\pi_i = 1/(s(\tau_i)\,m_{\tau_i})$ and get a regret bound $\frac{1}{\eta}\{\log m_{\tau_i} + \log s(\tau_i) + \log(1+\log s(T))\}$.

predictions to the experts which have not entered yet, so that they do not affect the learner's forecast. Another option is to design a prior on *sequences* of experts so that the *weight* of unentered experts is 0, and hence their predictions are irrelevant[3]; we call this the "muting trick".

After reviewing the well-known setting of aggregation of sequences of experts for a fixed set of experts (Section 5.5.1) and presenting the generic algorithm MarkovHedge with its regret bound, we adapt it to the growing experts setting by providing FreshMarkovHedge (Section 5.5.2) and GrowingMarkovHedge (Section 5.5.3), that compete respectively with fresh and arbitrary sequences.

5.5.1 Aggregating sequences of experts

The problem of controlling the regret with respect to sequences of experts, known as *tracking the best expert*, was introduced by Herbster and Warmuth (1998), who proposed the simple *Fixed Share* algorithm with good regret guarantees. A key fact, first recognized by Vovk (1999), is that Fixed Share, and in fact many other weight sharing algorithms (Koolen and de Rooij, 2013), can be interpreted as the exponential weights on sequences of experts under a suitable prior. We will state this result in the general form of Lemma 5.1, which implies the regret bound of Proposition 5.3.

Markov prior. If $i^T = (i_1, \dots, i_T)$ is a finite sequence of experts, its predictions up to time T are derived from those of the base experts $i \in \{1, \dots, M\}$ in the following way: $\widehat{y}_t(i^T) = \widehat{y}_{i_t,t}$ for $1 \leqslant t \leqslant T$. Given a prior $\pi = (\pi(i^T))_{i^T}$, we can in principle consider exponentially weighted aggregation of sequences under this prior; however, such an algorithm is intractable even for moderately low values of T, since it requires to store and update $O(M^T)$ weights. Fortunately, when $\pi(i_1, \dots, i_T) = \theta_1(i_1)\,\theta_2(i_2|i_1) \cdots \theta_T(i_T|i_{T-1})$ is a Markov probability distribution with initial measure $\boldsymbol{\theta}_1$ and transition matrices $\boldsymbol{\theta}_t$, $2 \leqslant t \leqslant T$, exponentially weighted aggregation under prior π collapses to the efficient algorithm MarkovHedge.

Remark 5.4. Algorithm MarkovHedge only requires to store and update $O(M)$ weights. Due to the matrix product (5.21), the update may take $O(M^2)$ time; however, all transition matrices we consider lead to a simple update in $O(M)$ time.

Lemma 5.1. *For every $T \geqslant 1$, the forecasts of algorithm* MarkovHedge *coincide up to time T with those of the exponential aggregation of finite sequences of experts $i^T = (i_1, \dots, i_T)$ under the Markov prior with initial distribution $\boldsymbol{\theta}_1$ and transition matrices $\boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_T$.*

Lemma 5.1 – proven in Section 5.7.3 – and Proposition 5.1 directly imply the following regret bound.

Proposition 5.3. *Algorithm* MarkovHedge*, with initial distribution $\boldsymbol{\theta}_1$ and transition matrices $\boldsymbol{\theta}_t$, guarantees the following regret bound: for every $T \geqslant 1$ and any sequence of experts $(i_1, \dots, i_T)$,*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} \leqslant \frac{1}{\eta} \log \frac{1}{\theta_1(i_1)} + \frac{1}{\eta} \sum_{t=2}^{T} \log \frac{1}{\theta_t(i_t|i_{t-1})}\,. \tag{5.22}$$

It is worth noting that the transition probabilities $\boldsymbol{\theta}_t$ only intervene at step t in algorithm MarkovHedge, and hence they can be chosen at this time.

[3] In this case, the learner's predictions do not depend on the way we complete the experts' predictions, so the algorithm may be defined even when experts with zero weight do not output predictions.

Algorithm 9 MarkovHedge — Aggregation of sequences of experts under a Markov prior

Parameters: Learning rate $\eta > 0$, initial weights $\boldsymbol{\theta}_1 = (\theta_1(i))_{1 \leqslant i \leqslant M}$, and transition probabilities $\boldsymbol{\theta}_t = \big(\theta_t(i|j)\big)_{1 \leqslant i,j \leqslant M}$ for all $t \geqslant 2$.
Initialization: Set $\boldsymbol{v}_1 = \boldsymbol{\theta}_1$.
for $t = 1, 2, \dots$ **do**
  Receive predictions $\widehat{\boldsymbol{y}}_t \in \widehat{\mathcal{Y}}^M$ from the experts, and predict $\widehat{y}_t = \boldsymbol{v}_t \cdot \widehat{\boldsymbol{y}}_t$.
  Observe $y_t \in \mathcal{Y}$, then derive the losses $\ell_t = \ell(\widehat{y}_t, y_t)$ and $\ell_{i,t} = \ell(\widehat{y}_{i,t}, y_t)$ and the posteriors

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,. \tag{5.20}$$

  Update the weights by $\boldsymbol{v}_{t+1} = \boldsymbol{\theta}_{t+1}\, \boldsymbol{v}_t^m$, *i.e.*

$$v_{i,t+1} = \sum_{j=1}^M \theta_{t+1}(i|j)\, v_{j,t}^m\,. \tag{5.21}$$

end for

Notable examples. In Section 5.7.3, we discuss particular instances of MarkovHedge that lead to well-known algorithms (such as Fixed Share), and recover their regret bounds using Proposition 5.3.

5.5.2 Application to sequences of fresh experts

We now explain how to specify the generic algorithm MarkovHedge in order to adapt it to the growing experts setting. This adaptation relies on the "muting trick": to obtain a strategy which is well-defined for growing experts, one has to ensure that experts who do not predict have zero weight, which amounts to saying that all weight is put to *admissible* sequences of experts. Importantly, this is possible even when the numbers M_t are not known from the beginning, since *the transition matrices $\boldsymbol{\theta}_t$ can be chosen at time t*, when M_t is revealed.

We start in this section by designing an algorithm FreshMarkovHedge that compares to sequences of fresh experts; to achieve this, it is natural to design a prior that assigns full probability to sequences of fresh experts. It turns out that we can recover an algorithm similar to the algorithm GrowingHedge, with the same regret guarantees, through this different viewpoint.

Let $\boldsymbol{\pi} = (\pi_i)_{i \geqslant 1}$ be an *unnormalized prior* as in Section 5.4.2. For each $M \geqslant 1$, we denote $\Pi_M = \sum_{i=1}^M \pi_i$. We consider the following transition matrices $\boldsymbol{\theta}_t$ in strategy MarkovHedge:

$$\theta_1(i) = \frac{\pi_i}{\Pi_{M_1}} \mathbf{1}_{i \leqslant M_1}\ ; \quad \theta_{t+1}(i|j) = \frac{\Pi_{M_t}}{\Pi_{M_{t+1}}} \mathbf{1}_{i=j} + \frac{\pi_i}{\Pi_{M_{t+1}}} \mathbf{1}_{M_t+1 \leqslant i \leqslant M_{t+1}} \tag{5.23}$$

for every $i \geqslant 1$, $t \geqslant 1$ and $j \in \{1, \dots, M_t\}$. The other transition probabilities $\theta_{t+1}(i|j)$ for $j > M_t$ are irrelevant; indeed, a simple induction shows that $v_{j,t} = 0$ for every $j > M_t$, so that the instantiation of algorithm MarkovHedge with the transition probabilities (5.23) leads to the forecasts

$$\widehat{y}_t = \sum_{i=1}^{M_t} v_{i,t}\, \widehat{y}_{i,t} \tag{5.24}$$

(which do not depend on the undefined prediction of the experts $i > M_t$) where the weights $(v_{i,t})_{1\leqslant i\leqslant M_t}$ are recursively defined by $v_{i,1} = \frac{\pi_i}{\Pi_{M_1}}$ $(1 \leqslant i \leqslant M_1)$ and the update

$$v_{i,t+1} = \frac{\Pi_{M_t}}{\Pi_{M_{t+1}}} v_{i,t}^m \quad (1 \leqslant i \leqslant M_t)\,; \quad v_{i,t+1} = \frac{\pi_i}{\Pi_{M_{t+1}}} \quad (M_t + 1 \leqslant i \leqslant M_{t+1})\,, \tag{5.25}$$

where we set $v_{i,t}^m = \dfrac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^{M_t} v_{j,t}\, e^{-\eta\, \ell_{j,t}}}$ for $1 \leqslant i \leqslant M_t$. We call this algorithm FreshMarkovHedge.

Theorem 5.2. *Algorithm* FreshMarkovHedge *using weights* $\boldsymbol{\pi}$ *achieves the following regret bound: for every* $T \geqslant 1$ *and sequence of fresh experts* $i^T = (i_1, \ldots, i_T)$ *with shifts at times* $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$,

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \sum_{j=0}^{k} \log \frac{1}{\pi_{i_{\sigma_j}}} + \frac{1}{\eta} \sum_{j=1}^{k} \log \Pi_{M_{\sigma_j - 1}} + \frac{1}{\eta} \log \Pi_{M_T}\,. \tag{5.26}$$

Additionally, the time and space complexity of the algorithm at each time step $t \geqslant 1$ *is* $O(M_t)$.

Proof. For any sequence of fresh experts $i^T \in \mathcal{S}_T^{(f)}(\boldsymbol{\sigma})$, replacing in the bound (5.22) of Proposition 5.3 the conditional probabilities $\theta_{t+1}(i_{t+1}|i_t)$ by their values (defined by (5.23)), we get

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \sum_{j=0}^{k} \left\{ \log \left(\frac{1}{\pi_{i_{\sigma_j}}} \Pi_{M_{\sigma_j}} \right) + \sum_{t=\sigma_j+1}^{\sigma_{j+1}-1} \log \frac{\Pi_{M_t}}{\Pi_{M_{t-1}}} \right\} = \frac{1}{\eta} \sum_{j=0}^{k} \log \frac{\Pi_{M_{\sigma_{j+1}-1}}}{\pi_{i_{\sigma_j}}}$$

which is precisely the desired bound (5.26). □

Remark 5.5. The regret bound (5.26) of the FreshMarkovHedge algorithm against sequences of fresh experts is exactly the same as that of the GrowingHedge algorithm (5.19). This is not a coincidence: the two algorithms are almost identical, except that expert i is introduced with weight $\pi_i / (\sum_{i=1}^{M_{\tau_i}} \pi_i)$ by FreshMarkovHedge and $\pi_i e^{-\eta\, L_{\tau_i - 1}} / (\sum_{j=1}^{M_{\tau_i}} \pi_j e^{-\eta\, L_{j,\tau_i - 1}})$ by GrowingHedge. For the logarithmic loss (with $\eta = 1$), these two weights are equal (see Remark 5.2), and hence the strategies GrowingHedge and FreshMarkovHedge coincide.

5.5.3 Regret against arbitrary sequences of experts

We now consider the more ambitious objective of comparing to *arbitrary* admissible sequences of experts. This can be done by using another choice of transition matrices, which puts all the weight to admissible sequences of experts (and not just sequences of fresh experts).

Algorithm GrowingMarkovHedge instantiates MarkovHedge on the transition matrices

$$\theta_1(i) = \frac{\pi_i}{\Pi_{M_1}} \mathbf{1}_{i\leqslant M_1}\,; \quad \theta_{t+1}(i|j) = \alpha_{t+1} \frac{\pi_i}{\Pi_{M_{t+1}}} + (1 - \alpha_{t+1})\, \theta_{t+1}^{(f)}(i|j) \tag{5.27}$$

where $\boldsymbol{\theta}_t^{(f)}$ denote the transition matrices of algorithm FreshMarkovHedge. As before, this leads to a well-defined growing experts algorithm which predicts $\widehat{y}_t = \sum_{i=1}^{M_t} v_{i,t}\, \widehat{y}_{i,t}$, where

the weights $(v_{i,t})_{1\leqslant i\leqslant M_t}$ are recursively defined by $v_{i,1} = \frac{\pi_i}{\Pi_{M_1}}$ $(1 \leqslant i \leqslant M_1)$ and the update

$$\begin{aligned} v_{i,t+1} &= (1-\alpha_{t+1})\frac{\Pi_{M_t}}{\Pi_{M_{t+1}}}\, v_{i,t}^m + \alpha_{t+1}\frac{\pi_i}{\Pi_{M_{t+1}}} \quad (1 \leqslant i \leqslant M_t)\,; \\ v_{i,t+1} &= \frac{\pi_i}{\Pi_{M_{t+1}}} \quad (M_t+1 \leqslant i \leqslant M_{t+1})\,, \end{aligned} \tag{5.28}$$

where again $v_{i,t}^m = v_{i,t}\, e^{-\eta\,\ell_{i,t}} / \sum_{j=1}^{M_t} v_{j,t}\, e^{-\eta\,\ell_{j,t}}$ for $1 \leqslant i \leqslant M_t$. In this case, Proposition 5.3 yields:

Theorem 5.3. *Algorithm* GrowingMarkovHedge *based on the weights* $\boldsymbol{\pi}$ *and parameters* $(\alpha_t)_{t\geqslant 2}$ *achieves the following regret bound: for every* $T \geqslant 1$*, and every admissible sequence of experts* $i^T = (i_1, \ldots, i_T)$ *with shifts at times* $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$*,*

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta}\left\{\sum_{j=0}^{k} \log\frac{\Pi_{M_{\sigma_{j+1}-1}}}{\pi_{i_{\sigma_j}}} + \sum_{j=1}^{k_1}\log\frac{1}{\alpha_{\sigma_j^1}} + \sum_{2\leqslant t\leqslant T\,:\, t\notin\boldsymbol{\sigma}} \log\frac{1}{1-\alpha_t}\right\}. \tag{5.29}$$

where $\boldsymbol{\sigma}^0 = (\sigma_1^0, \ldots, \sigma_{k_0}^0)$ *(resp.* $\boldsymbol{\sigma}^1 = (\sigma_1^1, \ldots, \sigma_{k_1}^1)$*) denotes the shifts to fresh (resp. incumbent) experts, with* $k = k_0 + k_1$*. Moreover, it has a* $O(M_t)$ *time and space complexity at each step* $t \geqslant 1$*.*

Remark 5.6. Note that by choosing $\alpha_t = \frac{1}{t}$, we have, since $\frac{1}{1-1/t} = \frac{t}{t-1}$,

$$\sum_{j=1}^{k_1}\log\frac{1}{\alpha_{\sigma_j^1}} + \sum_{2\leqslant t\leqslant T\,:\, t\notin\boldsymbol{\sigma}} \log\frac{1}{1-\alpha_t} \leqslant \sum_{j=1}^{k_1}\log\sigma_j^1 + \sum_{t=2}^{T}\log\frac{t}{t-1} = \sum_{j=1}^{k_1}\log\sigma_j^1 + \log T\,.$$

Additionally, by setting $\pi_i = 1$ the bound (5.29) becomes $\frac{1}{\eta}(\sum_{j=0}^{k}\log M_{\sigma_{j+1}-1} + \sum_{j=1}^{k_1}\log\sigma_j^1 + \log T)$, which is lower than $\frac{1}{\eta}(k+1)\log M_T + \frac{1}{\eta}(k_1+1)\log T$. We can also recover the bound (5.6) by setting $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$, since in this case we have $\Pi_{M_{\sigma_{j+1}-1}} \leqslant \Pi_{M_T} \leqslant \sum_{t=1}^{T}\frac{1}{t} \leqslant 1 + \log T$.

5.6 Combining growing experts and sequences of sleeping experts

Sections 5.4 and 5.5 studied the problem of growing experts using tools from two different settings (specialists and sequences of experts). Drawing on ideas from Koolen et al. (2012), we show in this section how to combine these two frameworks, in order to address the more challenging problem of controlling the regret with respect to *sparse sequences of experts* in the growing experts setting. Note that the refinement to sparse sequences of experts is particularly relevant in the context of a growing experts ensemble, since in this context the total number of experts will typically be large.

5.6.1 Sleeping experts: generic result

The problem of comparing to sparse sequences of experts, or *tracking a small pool of experts*, is a refinement on the problem of tracking the best expert. The seminal paper (Bousquet and Warmuth, 2002) proposed an ad-hoc strategy with essentially optimal regret bounds, the

Mixing Past Posteriors (MPP) algorithm (see also Cesa-Bianchi et al., 2012). A full "bayesian" interpretation of this algorithm in terms of the aggregation of "sleeping experts" was given by Koolen et al. (2012), which enabled the authors to propose a more efficient alternative. Here, by reinterpreting this construction, we propose a more general algorithm and regret bound (Proposition 5.4); this extension will be crucial to adapt this strategy to the growing experts setting (Section 5.6.2).

Given a fixed set of experts $\{1, \dots, M\}$, we call *sleeping expert* a couple $(i, a) \in \{1, \dots, M\} \times \{0, 1\}$; we endow the set of sleeping experts with a specialist structure by deciding that (i, a) is active if and only if $a = 1$, and that $\widehat{y}_t(i, 1) := \widehat{y}_{i,t}$ is the prediction of expert i. A key insight from Koolen et al. (2012) is to decompose the regret with respect to a sparse sequence $i^T = (i_1, \dots, i_T)$ of experts, taking values in the set $\{e_p \mid 1 \leqslant p \leqslant n\}$, in the following way:

$$\sum_{t=1}^{T} (\ell_t - \ell_{i_t,t}) = \sum_{p=1}^{n} \sum_{t \leqslant T \,:\, i_t = e_p} (\ell_t - \ell_{e_p,t}) = \sum_{p=1}^{n} \sum_{t=1}^{T} (\ell_t - \ell_t(e_p, a_{p,t})) = n \sum_{\iota^T} u(\iota^T)(L_T - L_T(\iota^T))$$

where $a_{p,t} := \mathbf{1}_{i_t = e_p}$, and u is the probability distribution on the sequences ι^T of sleeping experts which is uniform on the n sequences $\iota_p^T = (e_p, a_{p,t})_{1 \leqslant t \leqslant T}$, $p = 1, \dots, n$. Note that in the second equality we used the "abstention trick", which attributes to inactive sleeping experts $(e_p, 0)$ the prediction $\widehat{y}_t$ of the algorithm.

We can now aggregate sequences of sleeping experts under a Markov prior, given initial weights $\theta_1(i, a)$ and transition probabilities $\theta_{t+1}(i_{t+1}, a_{t+1} | i_t, a_t)$, recalling that $\boldsymbol{\theta}_t$ can be chosen at step t. For convenience, we restrict here to transitions that only occur between sleeping experts (i, a) with the same base expert, and denote $\theta_{i,t}(a|b) = \theta_t(i, a | i, b)$ for $a, b \in \{0, 1\}$. This leads to the algorithm SleepingMarkovHedge.

Remark 5.7. The structure of our prior is slightly more general than the one used by Koolen et al. (2012), which considered priors on couples (i, a^T) with an independence structure: $\pi(i, a^T) = \pi(i)\,\pi(a^T)$, with $\pi(a^T)$ a Markov distribution, which amounts to saying that the transition probabilities $\theta_{i,t}(a|b)$ could not depend on i. This additional flexibility will enable in Section 5.6.2 the "muting trick", which allows to convert SleepingMarkovHedge to the growing experts setting.

Proposition 5.4. *Strategy* SleepingMarkovHedge *guarantees the following regret bound: for each sequence i^T of experts taking values in the pool $\{e_p \mid 1 \leqslant p \leqslant n\}$, denoting $a_{p,t} = \mathbf{1}_{i_t = e_p}$*

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \sum_{p=1}^{n} \left(\log \frac{1/n}{\pi_{e_p}} + \log \frac{1}{\theta_{e_p,1}(a_{p,1})} + \sum_{t=2}^{T} \log \frac{1}{\theta_{e_p,t}(a_{p,t} | a_{p,t-1})} \right). \tag{5.33}$$

The proof of Proposition 5.4 is given in Section 5.7.4.

5.6.2 Sparse shifting regret for growing experts

We show here how to instantiate algorithm SleepingMarkovHedge in order to adapt it to the growing experts setting. Again, we use a "muting trick" which attributes a zero weight to experts that have not entered.

Let us consider prior weights $\boldsymbol{\pi} = (\pi_i)_{i \geqslant 1}$ on the experts, which may be unnormalized and chosen at entry time. Let $\alpha_t, \beta_t \in (0, 1)$ for $t \geqslant 2$. We set $\theta_{i,1}(1) = \frac{1}{2}$ for $i = 1, \dots, M_1$

Algorithm 10 SleepingMarkovHedge: sequences of sleeping experts under a Markov prior

1: **Parameters:** Learning rate $\eta > 0$, (normalized) prior $\boldsymbol{\pi}$ on experts, initial wake/sleep probabilities $\theta_{i,1}(a)$, transition probabilities $\boldsymbol{\theta}_{i,t} = \big(\theta_{i,t}(a|b)\big)_{a,b\in\{0,1\}}$ for $t \geqslant 2$, $1 \leqslant i \leqslant M$.

2: **Initialization:** Set $v_1(i,a) = \pi_i\, \theta_{i,1}(a)$ for $i = 1, \dots, M$ and $a \in \{0,1\}$.

3: **for** $t = 1, 2, \dots$ **do**

4: Receive predictions $\widehat{\boldsymbol{y}}_t \in \widehat{\mathcal{Y}}^M$ from the experts, and predict

$$\widehat{y}_t = \frac{\sum_{i=1}^M v_t(i,1)\, \widehat{y}_{i,t}}{\sum_{i=1}^M v_t(i,1)}\,. \tag{5.30}$$

5: Observe $y_t \in \mathcal{Y}$, then derive the losses $\ell_t(i,0) = \ell_t = \ell(\widehat{y}_t, y_t)$, $\ell_t(i,1) = \ell_{i,t} = \ell(\widehat{y}_{i,t}, y_t)$ and the posteriors

$$v_t^m(i,a) = \frac{v_t(i,a)\, e^{-\eta\, \ell_t(i,a)}}{\sum_{i',a'} v_t(i',a')\, e^{-\eta\, \ell_t(i',a')}}\,. \tag{5.31}$$

6: Update the weights by

$$v_{t+1}(i,a) = \sum_{b\in\{0,1\}} \theta_{i,t+1}(a|b)\, v_t^m(i,b)\,. \tag{5.32}$$

7: **end for**

and 0 otherwise; moreover, for every $t \geqslant 1$, we take $\theta_{i,t+1}(1|\cdot) = 0$ for $i > M_{t+1}$ (recall that $\boldsymbol{\theta}_{i,t+1}$ can be chosen at step $t+1$), $\theta_{i,t+1}(1|\cdot) = \frac{1}{2}$ if $M_t + 1 \leqslant i \leqslant M_{t+1}$, and for $i \leqslant M_t$: $\theta_{i,t+1}(0|1) = \alpha_{t+1}$, $\theta_{i,t+1}(1|0) = \beta_{t+1}$. The algorithm obtained with these choices, which we call GrowingSleepingMarkovHedge, is well-defined and predicts

$$\widehat{y}_t = \frac{\sum_{i=1}^{M_t} v_t(i,1)\, \widehat{y}_{i,t}}{\sum_{i=1}^{M_t} v_t(i,1)}\,, \tag{5.34}$$

where the weights $(v_t(i,a))_{1\leqslant i\leqslant M_t,\, a\in\{0,1\}}$ are defined by $v_1(i,a) = \frac{1}{2}\,\pi_i$ $(1 \leqslant i \leqslant M_1)$ and by the update

$$\begin{aligned} v_{t+1}(i,a) &= \sum_{b\in\{0,1\}} \theta_{i,t+1}(a|b)\, v_t^m(i,b) \quad (1 \leqslant i \leqslant M_t, a \in \{0,1\})\,; \\ v_{t+1}(i,a) &= \frac{1}{2}\,\pi_i \quad (M_t + 1 \leqslant i \leqslant M_{t+1}, a \in \{0,1\})\,, \end{aligned}$$

with $v_t^m(i,a) = v_t(i,a)\, e^{-\eta\, \ell_t(i,a)} / \sum_{i=1}^{M_t} \sum_{a'\in\{0,1\}} v_t(i',a')\, e^{-\eta\, \ell_t(i',a')}$ for $1 \leqslant i \leqslant M_t$.

Theorem 5.4. *Algorithm* GrowingSleepingMarkovHedge *guarantees the following: for each $T \geqslant 1$ and any sequence i^T of experts taking values in the pool $\{e_p \mid 1 \leqslant p \leqslant n\}$, denoting*

$a_{p,t} = \mathbf{1}_{i_t = e_p}$

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \sum_{p=1}^{n} \log \frac{\Pi_{M_T}/n}{\pi_{e_p}} + \frac{1}{\eta} n \log 2 + \frac{1}{\eta} \sum_{t=2}^{T} \left[\log \frac{1}{1-\alpha_t} + (n-1) \log \frac{1}{1-\beta_t} \right] + \frac{1}{\eta} \sum_{j=1}^{k} \left(\log \frac{1}{\alpha_{\sigma_j}} + \log \frac{1}{\beta_{\sigma_j}} \right) \tag{5.35}$$

where $\boldsymbol{\sigma} = \sigma_1 < \cdots < \sigma_k$ *denote the shifting times of* i^T*. Moreover, the algorithm has a* $O(M_t)$ *time and space complexity at step* t*, for every* $t \geqslant 1$*.*

In particular, Theorem 5.4 enables to recover the bound (5.7) for $\alpha_t = \beta_t = 1/t$ and $\pi_i = 1/(\tau_i m_{\tau_i})$.

Proof. Note that algorithm GrowingSleepingMarkovHedge is invariant under any change of prior $\boldsymbol{\pi} \leftarrow \lambda \boldsymbol{\pi}$ due to the renormalisation in the formula (5.35) defining $\widehat{y}_t$. In particular, setting $\lambda = 1/\Pi_{M_T}$, we see that it coincides up to time T with algorithm SleepingMarkovHedge with set of experts $\{1, \ldots, M_T\}$ and (normalized) prior weights π_i/Π_{M_T}. The bound (5.35) is now a consequence of the general regret bound (5.33), by substituting for the values of $\boldsymbol{\theta}_{i,t+1}$. □

Conclusion. In this chapter, we extended aggregation of experts to the *growing expert* setting, where novel experts are made available at any time. In this context when the set of experts itself varies, it is natural to seek to track the best expert; different comparison classes of increasing complexity were considered. In order to obtain efficient algorithms with a per-round complexity linear in the current number of experts, we started with generic reformulation of existing algorithms for fixed expert set, and identified two orthogonal techniques (the "abstention trick" from the specialist literature, and the "muting trick") to adapt them to sequentially incoming forecasters. Combined with a proper tuning of the parameters of the prior, this enabled us to obtain tight regret bounds, adaptive to the parameters of the comparison class. Along the way, we recovered several key results from the literature as special cases of our analysis, in a somewhat unified approach.

Although we considered the exp-concave assumption to avoid distracting the reader from the main challenges of the growing expert setting, extending our results to the bounded convex case in which the parameter η needs to be adaptively tuned seems possible and is left for future work. In addition, building on the recent work of Jun et al. (2017) might bring further improvements in this case. Another natural extension of our work would be to address the same questions in the framework of online convex optimization (Shalev-Shwartz, 2012; Hazan, 2016), in the case where the dimension of the constraint set may increase over time as new features are added.

5.7 Proofs

5.7.1 Proof of Proposition 5.1

Proof. Since the loss function is η-exp-concave and $\widehat{y}_t = \sum_{i=1}^M v_{i,t}\,\widehat{y}_{i,t}$, we have

$$e^{-\eta\,\ell(\widehat{y}_t, y_t)} \geqslant \sum_{i=1}^M v_{i,t}\, e^{-\eta\,\ell(\widehat{y}_{i,t}, y_t)}, \qquad i.e. \qquad \ell_t \leqslant -\frac{1}{\eta}\log\left(\sum_{i=1}^M v_{i,t}\, e^{-\eta\,\ell_{i,t}}\right).$$

This yields, introducing the posterior weights $v_{i,t}^m$ defined by (5.9),

$$\ell_t - \ell_{i,t} \leqslant -\frac{1}{\eta}\log\left(\sum_{j=1}^M v_{j,t}\, e^{-\eta\,\ell_{j,t}}\right) - \ell_{i,t} = \frac{1}{\eta}\log\left(\frac{e^{-\eta\,\ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\,\ell_{j,t}}}\right) = \frac{1}{\eta}\log\frac{v_{i,t}^m}{v_{i,t}}.$$

Now recalling that the exponentially weighted average forecaster uses $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t^m$, this writes: $\ell_t - \ell_{i,t} \leqslant \frac{1}{\eta}\log\frac{v_{i,t+1}}{v_{i,t}}$ which, summing over $t = 1, \ldots, T$, yields $L_T - L_{i,T} \leqslant \frac{1}{\eta}\log\frac{v_{i,T+1}}{v_{i,1}}$. Since $v_{i,1} = \pi_i$ and $v_{i,T+1} \leqslant 1$, this proves (5.10); moreover, noting that $\log\frac{v_{i,T+1}}{v_{i,1}} = \log\frac{\rho_i}{v_{i,1}} - \log\frac{\rho_i}{v_{i,T+1}}$, this implies

$$\sum_{i=1}^M \rho_i\,(L_T - L_{i,T}) \leqslant \frac{1}{\eta}\sum_{i=1}^M \rho_i\,\log\frac{v_{i,T+1}}{v_{i,1}} = \frac{1}{\eta}\big(\mathrm{KL}(\boldsymbol{\rho}, \boldsymbol{v}_1) - \mathrm{KL}(\boldsymbol{\rho}, \boldsymbol{v}_{T+1})\big),$$

which establishes (5.11) since $\boldsymbol{v}_1 = \boldsymbol{\pi}$ and $\mathrm{KL}(\boldsymbol{\rho}, \boldsymbol{v}_{T+1}) \geqslant 0$. □

Remark 5.8. We can recover the bound (5.10) from inequality (5.11) by considering $\boldsymbol{\rho} = \delta_i$. Conversely, inequality (5.10) implies, by convex combination,

$$L_T - \sum_{i=1}^M \rho_i\, L_{i,T} \leqslant \frac{1}{\eta}\sum_{i=1}^M \rho_i \log\frac{1}{\pi_i};$$

inequality (5.11) is actually an improvement on this bound, which replaces the terms $\log\frac{1}{\pi_i}$ by $\log\frac{\rho_i}{\pi_i}$. Following Koolen et al. (2012), this refinement is used in Section 5.6.1 to obtain a tighter regret bound.

5.7.2 Proof of Theorem 5.1

Theorem 5.1 is in fact a corollary of the more general Proposition 5.5, valid in the specialist setting.

Proposition 5.5. *Assume we are given a set $\mathcal{M}$ of specialists, as well as a positive weight function $\pi : \mathcal{M} \to \mathbf{R}_+^*$. Assume that, at each time step $t \geqslant 1$, the set A_t of active specialists is finite. Then, denoting $A_{\leqslant t} = \bigcup_{1 \leqslant s \leqslant t} A_s$, the aggregation of specialists*[4]

$$\widehat{y}_t = \frac{\sum_{i \in A_t} \pi(i)\, e^{-\eta\, L_{i,t-1}}\, \widehat{y}_{i,t}}{\sum_{i \in A_t} \pi(i)\, e^{-\eta\, L_{i,t-1}}} \tag{5.36}$$

[4]Denoting, as in equation (5.14), $L_{i,t} = \sum_{s \leqslant t\,:\, i \in A_s} \ell_{i,s} + \sum_{s \leqslant t\,:\, i \notin A_s} \ell_s$ for each specialist i and $t \geqslant 1$.

achieves the following regret bound: for each $T \geqslant 1$ and $i \in \mathcal{M}$, we have

$$\sum_{t \leqslant T\,:\, i \in A_t} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log\left(\frac{1}{\pi(i)} \sum_{j \in A_{\leqslant T}} \pi(j) \right). \tag{5.37}$$

Proof of Proposition 5.5. Fix $T \geqslant 1$, and denote $\Pi_T := \sum_{i \in A_{\leqslant T}} \pi(i)$. For $t = 1, \dots, T$, the forecast (5.36) may be rewritten as

$$\widehat{y}_t = \frac{\sum_{i \in A_t} \frac{\pi(i)}{\Pi_T} e^{-\eta\, L_{i,t-1}}\, \widehat{y}_{i,t}}{\sum_{i \in A_t} \frac{\pi(i)}{\Pi_T} e^{-\eta\, L_{i,t-1}}}$$

which corresponds precisely to the aggregation of the set of specialists $A_{\leqslant T}$ with prior weights $\pi(i)/\Pi_T$ and active specialists $A_t \subset A_{\leqslant T}$ (up to time T). (5.37) now follows from Proposition 5.2. □

Proof of Theorem 5.1. It suffices to notice that the weights of GrowingHedge are, for $i \leqslant M_t$, $w_{i,t} = \pi_i\, e^{-\eta L_{i,t-1}}$ with $L_{i,t-1} = L_{\tau_i - 1} + \sum_{s=\tau_i}^{t} \ell_{i,s}$; hence, the forecasts of GrowingHedge are those of equation (5.36), and we can apply Proposition 5.5. □

5.7.3 Proof of Lemma 5.1 and instantiations of MarkovHedge

Proof of Lemma 5.1. Denote, for each $t \geqslant 1$, $\pi^t(i_1, \dots, i_t) = \theta_1(i_1)\, \theta_2(i_2|i_1) \cdots \theta_t(i_t|i_{t-1})$. Let $T \geqslant 1$ be arbitrary. We need to show that the predictions $\widehat{y}_t$ of the exponentially weighted aggregation of sequences of experts i^T under the prior π^T at times $t = 1, \dots, T$ coincide with those of algorithm MarkovHedge.

First note that, by definition and since $L_{t-1}(i^T) = \sum_{s=1}^{t-1} \ell_{i_s,s} =: L_{t-1}(i^{t-1})$ does not depend on $i_t^T = (i_t, \dots, i_T)$, we have for $1 \leqslant t \leqslant T$

$$\begin{aligned}
\widehat{y}_t &= \frac{\sum_{i^T} \pi^T(i^T)\, e^{-\eta\, L_{t-1}(i^T)}\, \widehat{y}_t(i^T)}{\sum_{i^T} \pi^T(i^T)\, e^{-\eta\, L_{t-1}(i^T)}} = \frac{\sum_{i^t, i_{t+1}^T} \pi^T(i^t, i_{t+1}^T)\, e^{-\eta\, L_{t-1}(i^{t-1})}\, \widehat{y}_{i_t,t}}{\sum_{i^t, i_{t+1}^T} \pi^T(i^t, i_{t+1}^T)\, e^{-\eta\, L_{t-1}(i^{t-1})}} \\
&\underset{(\star)}{=} \frac{\sum_{i^t} \pi^t(i^t)\, e^{-\eta\, L_{t-1}(i^{t-1})}\, \widehat{y}_{i_t,t}}{\sum_{i^t} \pi^t(i^t)\, e^{-\eta\, L_{t-1}(i^{t-1})}} = \frac{\sum_{i^{t-1}, i} \pi^t(i^{t-1}, i)\, e^{-\eta\, L_{t-1}(i^{t-1})}\, \widehat{y}_{i,t}}{\sum_{i^{t-1}, i} \pi^t(i^{t-1}, i)\, e^{-\eta\, L_{t-1}(i^{t-1})}}
\end{aligned}$$

where $(\star)$ is a consequence of the identity $\sum_{i_{t+1}^T} \pi^T(i^t, i_{t+1}^T) = \pi^t(i^t)$. Hence, denoting $w_t(i^t) := \pi^t(i^t)\, e^{-\eta L_{t-1}(i^{t-1})}$, we have

$$\widehat{y}_t = \frac{\sum_i \sum_{i^{t-1}} w_t(i^{t-1}, i)\, \widehat{y}_{i,t}}{\sum_i \sum_{i^{t-1}} w_t(i^{t-1}, i)} = \frac{\sum_{i=1}^M w_{i,t}\, \widehat{y}_{i,t}}{\sum_{i=1}^M w_{i,t}} = \sum_{i=1}^M v_{i,t}\, \widehat{y}_{i,t}$$

where we set $w_{i,t} := \sum_{i^{t-1}} \pi^t(i^{t-1}, i)\, e^{-\eta L_{t-1}(i^{t-1})}$ and $v_{i,t} := w_{i,t} / \big(\sum_{j=1}^M w_{j,t}\big)$. To conclude the proof, it remains to show that the weights $\boldsymbol{v}_t$ are those computed by algorithm MarkovHedge.

We proceed by induction on $t \geqslant 1$. For $t = 1$, we have for every $i = 1, \dots, M$, $w_{i,1} = w_1(i) = \pi^1(i) = \theta_1(i)$ and hence $v_{i,1} = \theta_1(i)$, *i.e.* $\boldsymbol{v}_1 = \boldsymbol{\theta}_1$. Moreover, for every $t \geqslant 1$, the

identity $\pi^{t+1}(i^{t+1}) = \pi^t(i^t)\,\theta_{t+1}(i_{t+1}|i_t)$ implies

$$\begin{aligned} w_{t+1}(i^{t+1}) &= \pi^{t+1}(i^{t+1})\,e^{-\eta L_t(i^t)} \\ &= \theta_{t+1}(i_{t+1}|i_t)\,\pi^t(i^t)\,e^{-\eta L_{t-1}(i^{t-1})}\,e^{-\eta\,\ell_{i_t,t}} \\ &= \theta_{t+1}(i_{t+1}|i_t)\,w_t(i^t)\,e^{-\eta\,\ell_{i_t,t}} \end{aligned}$$

i.e., for every i, j and i^{t-1}, $w_{t+1}(i^{t-1}, j, i) = \theta_{t+1}(i|j)\,w_t(i^{t-1}, j)\,e^{-\eta\,\ell_{j,t}}$. Summing over i^{t-1} and j, this yields:

$$w_{i,t+1} = \sum_{j=1}^{M} \theta_{t+1}(i|j)\,w_{j,t}\,e^{-\eta\,\ell_{j,t}}\,. \tag{5.38}$$

Summing (5.38) over $i = 1, \dots, M$ gives $\sum_{i=1}^M w_{i,t+1} = \sum_{j=1}^M w_{j,t}\,e^{-\eta\,\ell_{j,t}}$ (since $\sum_{i=1}^M \theta_{t+1}(i|j) = 1$) and therefore

$$\begin{aligned} v_{i,t+1} = \frac{w_{i,t+1}}{\sum_{j=1}^M w_{j,t+1}} &= \frac{\sum_{j=1}^M \theta_{t+1}(i|j)\,w_{j,t}\,e^{-\eta\,\ell_{j,t}}}{\sum_{j=1}^M w_{j,t}\,e^{-\eta\,\ell_{j,t}}} = \frac{\sum_{j=1}^M \theta_{t+1}(i|j)\,v_{j,t}\,e^{-\eta\,\ell_{j,t}}}{\sum_{j=1}^M v_{j,t}\,e^{-\eta\,\ell_{j,t}}} \\ &= \sum_{j=1}^{M} \theta_{t+1}(i|j)\,v_{j,t}^m \end{aligned}$$

where $\boldsymbol{v}_t^m$ is the posterior distribution, defined by equation (5.9). This corresponds precisely to the update of the MarkovHedge algorithm, which completes the proof. □

We now instantiate the generic algorithm MarkovHedge and Proposition 5.3 on specific choices of prior weights and transition probabilities. This enables to recover a number of results from the literature. For concreteness, we take $\boldsymbol{\theta}_1 = \frac{1}{M}\mathbf{1}$.

Corollary 5.1 (Fixed share)**.** *Setting $\theta_t(i|j) = (1-\alpha)\,\mathbf{1}_{i=j} + \alpha\,\frac{1}{M}$ with $\alpha \in (0,1)$, this leads to the Fixed-Share algorithm of Herbster and Warmuth (1998) with update $\boldsymbol{v}_{t+1} = (1-\alpha)\,\boldsymbol{v}_t^m + \alpha\,\frac{1}{M}\mathbf{1}$ and regret bound*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} \leqslant \frac{k+1}{\eta}\log M + \frac{k}{\eta}\log\frac{1}{\alpha} + \frac{T-k-1}{\eta}\log\frac{1}{1-\alpha}\,, \tag{5.39}$$

where $k = k(i^T)$ denotes the number of shifts, $1 < \sigma_1 < \cdots < \sigma_k \leqslant T$ these shifts (such that $i_{\sigma_j} \neq i_{\sigma_j - 1}$) and $\sigma_0 = 1$. When T and k are fixed and known, this bound is minimized by choosing $\alpha = \frac{k}{T-1}$ and becomes, denoting $H(p) = -p\log p - (1-p)\log(1-p)$ the binary entropy function,

$$\frac{k+1}{\eta}\log M + \frac{T-1}{\eta}H\Big(\frac{k}{T-1}\Big) \leqslant \frac{k+1}{\eta}\log M + \frac{k}{\eta}\log\frac{T-1}{k} + \frac{k}{\eta}\,. \tag{5.40}$$

Remark 5.9. The quantity of equation (5.40), *i.e.* the bound on the regret of fully tuned Fixed Share algorithm, is essentially equal to the optimal bound $\frac{1}{\eta}\log\binom{T-1}{k}M^{k+1} \approx \frac{k+1}{\eta}\log M + \frac{k}{\eta}\log\frac{T-1}{k}$, obtained by aggregating all sequences of experts with at most k shifts (which would require to maintain a prohibitively large number of weights).

Corollary 5.2 (Decreasing share)**.** *Consider the special case of algorithm* MarkovHedge *where* $\theta_t(i|j) = (1-\alpha_t)\,\mathbf{1}_{i=j} + \frac{\alpha_t}{M}$*, so that the update becomes* $\boldsymbol{v}_{t+1} = (1-\alpha_{t+1})\,\boldsymbol{v}_t^m + \frac{\alpha_{t+1}}{M}\,\mathbf{1}$*. For every* $T \geqslant 1$*,* $0 \leqslant k \leqslant T$*, and every sequence of experts* $i^T = (i_1, \dots, i_T)$ *with* k *shifts at times* $\sigma_1 < \cdots < \sigma_k$*,*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} \leqslant \frac{k+1}{\eta} \log M + \frac{1}{\eta} \sum_{j=1}^{k} \log \frac{1}{\alpha_{\sigma_j}} + \frac{1}{\eta} \sum_{t=2}^{T} \log \frac{1}{1-\alpha_t} \tag{5.41}$$

In the special case[5] *when* $\alpha_t = \frac{1}{t}$*, this bound becomes, for* every *T, k and i^T:*

$$\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} \leqslant \frac{k+1}{\eta} \log M + \frac{1}{\eta} \sum_{j=1}^{k} \log \sigma_j + \frac{1}{\eta} \log T \leqslant \frac{k+1}{\eta} \log M + \frac{k+1}{\eta} \log T\,. \tag{5.42}$$

Remark 5.10. The result of Corollary 5.2 is worth emphasizing: at no computational overhead, the use of decreasing transition probabilities gives a bound essentially in $\frac{1}{\eta}(k+1)\log M + \frac{1}{\eta} k \log T$ valid for every T and k, which is close to the bound $\frac{1}{\eta}(k+1)\log M + \frac{1}{\eta} k \log \frac{T}{k}$ one gets by optimally tuning α as a function of T and k in the Fixed Share algorithm, particularly when $k \ll T$ (in this latter case of rare shifts, the first bound of equation (5.42) is sharper).

Proof of Corollaries 5.1 and 5.2. We consider the Decreasing Share algorithm, with time-varying transition probabilities $\alpha_t \in (0,1)$ (the Fixed Share algorithm corresponds to the special case $\alpha_t = \alpha$). Let $i^T = (i_1, \dots, i_T)$ be a sequence of experts with shifts at times $\sigma_1 < \cdots < \sigma_k$. By Proposition 5.3, we have

$$\begin{aligned}\sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t} &\leqslant \frac{1}{\eta} \log \frac{1}{1/M} + \frac{1}{\eta} \sum_{j=1}^{k} \log \frac{1}{\alpha_{\sigma_j}/M} + \frac{1}{\eta} \sum_{t \neq \sigma_j} \log \frac{1}{1-\alpha_{\sigma_j} + \alpha_{\sigma_j}/M} \\ &\leqslant \frac{k+1}{\eta} \log M + \frac{1}{\eta} \sum_{j=1}^{k} \log \frac{1}{\alpha_{\sigma_j}} + \frac{1}{\eta} \sum_{t \neq \sigma_j} \log \frac{1}{1-\alpha_{\sigma_j}}\end{aligned}$$

Corollary 5.1 directly follows by taking $\alpha_t = \alpha$ in the above inequality, whereas the bound (5.41) of Corollary 5.2 is obtained by bounding $\sum_{t \neq \sigma_j} \log \frac{1}{1-\alpha_t} \leqslant \sum_{t=2}^{T} \log \frac{1}{1-\alpha_t}$. In the case when $\alpha_t = \frac{1}{t}$, we recover (5.42) by substituting for α_t and noting that

$$\sum_{t=2}^{T} \log \frac{1}{1-1/t} = \sum_{t=2}^{T} \log \frac{t}{t-1} = \log T\,. \qquad \square$$

5.7.4 Proof of Proposition 5.4

Proof. Since $\widehat{y}_t(i,1) = \widehat{y}_{i,t}$ and $\widehat{y}_t(i,0) = \widehat{y}_t$, equation (5.30) implies that the forecast $\widehat{y}_t$ of SleepingMarkovHedge satisfies:

$$\widehat{y}_t = \sum_{i=1}^{M} \sum_{a \in \{0,1\}} v_t(i,a)\, \widehat{y}_t(i,a)\,.$$

[5] Which we consider because of the simplicity of the bound as well as its proof, involving a telescoping simplification; it is akin to Theorem 10 of Koolen and de Rooij (2013), which uses $\alpha_t = 1 - e^{-c/t}$.

Hence, SleepingMarkovHedge reduces to algorithm MarkovHedge over the sleeping experts, *i.e.* (by Lemma 5.1, up to time T) to the exponentially weighted aggregation of sequences of sleeping experts under the Markov prior

$$\pi((i,a_t)_{1\leqslant t\leqslant T}) = \theta_{i,1}(a_1)\prod_{t=2}^{T}\theta_{i,t}(a_t|a_{t-1})$$

(and 0 for other sequences). Hence, if u is the uniform probability on the n sequences $(e_p,a_{p,t})_{1\leqslant t\leqslant T}$, $1\leqslant p\leqslant n$, we have by Proposition 5.1:

$$\sum_{\iota^T} u(\iota)\,(L_T - L_T(\iota^T)) \leqslant \frac{1}{\eta}\mathrm{KL}(u,\pi) = \frac{1}{\eta}\frac{1}{n}\sum_{p=1}^{n}\log\frac{1/n}{\pi((e_p,a_{p,t})_{1\leqslant t\leqslant T})} \tag{5.43}$$

As shown in the reformulation of the regret with respect to sparse sequences of experts of Section 5.6.1, the left hand side of equation (5.43) equals $\frac{1}{n}(L_T - L_T(i^T))$. The desired regret bound (5.33) follows by substituting for π in the right-hand side. □

5.7.5 Uniform bounds and optimality

In this section, we provide simple bounds derived from Theorems 5.1, 5.2, 5.3 and 5.4 that are not quite as adaptive to the parameters of the comparison class as the ones provided in Section 5.2, but are more uniform and hence more interpretable. We then discuss the optimality of these bounds, by relating them either to theoretical lower bounds or to information-theoretic upper bounds (obtained by naively aggregating all elements of the comparison class, which is computationally prohibitive).

Constant experts. Consider the algorithm GrowingHedge with the uniform (unnormalized) prior: $\pi_i = 1$ for each $i\geqslant 1$. By Theorem 5.1, this algorithm achieves the regret bound

$$\frac{1}{\eta}\log M_T$$

with respect to each constant expert. This regret bound cannot be improved in general: indeed, consider the logarithmic loss on $\mathbf{N}^*$, defined by $\ell(\widehat{y},y) = -\log\widehat{y}(y)$ for every $y\in\mathbf{N}^*$ and every probability distribution x on $\mathbf{N}^*$. Fix $T\geqslant 1$, and consider the sequence $y_t = \widehat{y}_{i,t} = 1$ ($1\leqslant t<T$, $1\leqslant i\leqslant M_t$) and $y_t\in\{1,\dots,M_T\}$ and $\widehat{y}_{i,T} = i$ for $i=1,\dots,M_T$. For each $i=1,\dots,M_T$, we have $\sup_{1\leqslant i\leqslant M_T}(L_T - L_{i,T}) = \sup_{1\leqslant i\leqslant M_T} -\log\frac{\widehat{y}_t(y_t)}{\widehat{y}_{i,t}(y_t)} = -\log\widehat{y}_t(y_t)$. Now whatever $\widehat{y}_t$ is, there exists $y_t\in\{1,\dots,M_T\}$ such that $\widehat{y}_t(y_t)\leqslant\frac{1}{M_T}$ (since $\widehat{y}_t$ sums to 1). Since y_t is picked by an adversary after $\widehat{y}_t$ is chosen, the adversary can always ensure a regret of at least $\log M_T$.

Arbitrary admissible sequences of experts. By Theorem 5.3, algorithm GrowingMarkovHedge with uniform prior $\boldsymbol{\pi}$ and transition probabilities $\alpha_t = \frac{1}{t}$ achieves, for every admissible sequence i^T

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta}\sum_{j=0}^{k}\log M_{\sigma_{j+1}-1} + \frac{1}{\eta}\sum_{j=1}^{k_1}\log\sigma_j^1 + \frac{1}{\eta}\log T \leqslant \frac{1}{\eta}(k+1)\log M_T + \frac{1}{\eta}(k_1+1)\log T.$$

where $\boldsymbol{\sigma}^0 = (\sigma_1^0, \dots, \sigma_{k_0}^0)$ (resp. $\boldsymbol{\sigma}^1 = (\sigma_1^1, \dots, \sigma_{k_1}^1)$) denotes the shifts to fresh (resp. incumbent) experts, with $k = k_0 + k_1$.

This simple bound is close to the information-theoretic bound obtained by aggregating all admissible sequences of experts: indeed, the number of such sequences is bounded by (with equality if $M_T = M_1$) $M_T^{k+1}\binom{T-1}{k_1}$ (an admissible sequence is determined by its switches to fresh experts – at most $M_T^{k_0+1}$ possibilities – and its switches to incumbent experts – at most $M_T^{k_1}$ possibilities for the choices of the experts, and at most $\binom{T-1}{k_1}$ choices for the switches to incumbent experts). The regret bound corresponding to the aggregation of this large expert class is therefore of order

$$\frac{1}{\eta} \log M_T^{k+1} \binom{T-1}{k_1} \approx \frac{1}{\eta}(k+1) \log M_T + \frac{1}{\eta} k_1 \log \frac{T-1}{k_1} \,,$$

which is close to the bound of GrowingHedge, especially if $k_1 \ll T$.

Sparse admissible sequences. Finally, Theorem 5.4 implies that algorithm GrowingSleepingMarkovHedge, with uniform weights $\boldsymbol{\pi}$ and transition probabilities $\alpha_t = \beta_t = \frac{1}{t \log t}$, has a regret bound of

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} n \log \frac{M_T}{n} + \frac{1}{\eta} n (\log 2 + c_T \log\log T) + \frac{2}{\eta} k \log T + \frac{1}{\eta} 2k \log\log T \,.$$

for any sparse admissible sequence i^T with at most k shifts and taking values in a pool of n experts, where $c_T := (\log\log T)^{-1} \sum_{t=2}^{T} \log \frac{1}{1-\alpha_t} \longrightarrow_{T\to\infty} 1$. Again, for $k \ll T$, this is close to the information-theoretic upper bound obtained by aggregating all sparse sequences with k shifts in a pool of n experts, of approximately $n \log \frac{M_T}{n} + (k+1) \log n + k \log \frac{T}{k}$. The main difference, namely the doubling of the term $k \log T$ in the regret bound of GrowingSleepingMarkovHedge, is not specific to the growing experts setting, and also appears in the context of a fixed set of experts (Bousquet and Warmuth, 2002; Koolen et al., 2012).

Part III

Density estimation, least squares and logistic regression

Chapter 6

Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices

Abstract. The first part of this chapter is devoted to the decision-theoretic analysis of random-design linear prediction. It is known from Tsybakov (2003) that, under boundedness constraints on the response or regression coefficients, minimax excess risk scales, up to constants, as $\sigma^2 d/n$ in dimension d with n samples and noise σ^2. Here, we do not restrict the optimal regression parameter, and study the expected excess risk with respect to the full linear class. First, the ordinary least squares estimator is exactly minimax optimal in the well-specified case for every distribution of covariates. Further, we express the minimax risk in terms of the distribution of *statistical leverage scores* of individual samples. We deduce a precise minimax lower bound of $\sigma^2 d/(n-d+1)$ for general covariate distribution, nearly matching the risk for Gaussian design. We then obtain nonasymptotic upper bounds on the minimax risk for covariates that satisfy a "small ball"-type regularity condition, scaling as $(1+o(1))\sigma^2 d/n$ as $d = o(n)$ both in the well-specified and misspecified cases.

Our main technical contribution is the study of the lower tail of the smallest singular value of empirical covariance matrices around 0. We establish a lower bound on this lower tail, valid for any distribution in dimension $d \geqslant 2$, together with a matching upper bound under a necessary regularity condition. Our proof relies on the PAC-Bayesian technique for controlling empirical processes, and extends an analysis of Oliveira (2016) devoted to a different part of the lower tail. Equivalently, the operator norm of the inverse sample covariance matrix has bounded L^q norm up to $q \asymp n$, and this exponent is unimprovable. Finally, we show that the regularity condition naturally holds for independent coordinates.

Contents

6.1 Introduction

The linear least-squares problem, also called *random-design linear regression* or *linear aggregation*, is a standard problem in Statistics and Learning Theory. Specifically, given a random pair (X, Y) where X is a covariate vector in $\mathbf{R}^d$ and Y is a scalar response, the aim is to predict Y using a linear function $\langle \beta, X \rangle = \beta^\top X$ (with $\beta \in \mathbf{R}^d$) of X as well as possible, in a sense measured by the prediction risk with squared error $R(\beta) = \mathbb{E}[(Y - \langle \beta, X \rangle)^2]$. The best prediction is achieved by the population risk minimizer β^*, which equals:

$$\beta^* = \Sigma^{-1} \mathbb{E}[YX]$$

where $\Sigma := \mathbb{E}[XX^\top]$, assuming that both Σ and $\mathbb{E}[YX]$ are well-defined and that Σ is invertible. In the statistical setting considered here, the joint distribution P of the pair (X, Y) is unknown. The goal is then, given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of n i.i.d. realizations of P, to find a predictor (also called *estimator*) $\widehat{\beta}_n$ with small *excess risk*

$$\mathcal{E}(\widehat{\beta}_n) := R(\widehat{\beta}_n) - R(\beta^*) = \|\widehat{\beta}_n - \beta^*\|_\Sigma^2 \,,$$

where we define $\|\beta\|_\Sigma^2 := \langle \Sigma\beta, \beta \rangle = \|\Sigma^{1/2}\beta\|^2$. Arguably the most common procedure is the *Ordinary Least Squares* (OLS) estimator (that is, the empirical risk minimizer), defined by

$$\widehat{\beta}_n^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \beta, X_i \rangle)^2 \right\} = \widehat{\Sigma}_n^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i \,,$$

with $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n X_i X_i^\top$ the sample covariance matrix.

Linear classes are of fundamental importance to regression problems, both for themselves and since they naturally appear in the context of nonparametric estimation (Györfi et al., 2002; Tsybakov, 2009). In this chapter, we perform a precise decision-theoretic analysis of this problem, focusing on the minimax excess risk with respect to the full linear class $\mathcal{F} = \{x \mapsto \langle \beta, x \rangle : \beta \in \mathbf{R}^d\}$. This minimax perspective is relevant when little is known (or assumed) on the optimal parameter β^*. Specifically, define the *minimax excess risk* (see, e.g., Lehmann and Casella, 1998) with respect to $\mathcal{F}$ under a set $\mathcal{P}$ of joint distributions P on (X, Y) as:

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathcal{E}(\widehat{\beta}_n)] = \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}} \left(\mathbb{E}[R(\widehat{\beta}_n)] - \inf_{\beta \in \mathbf{R}^d} R(\beta) \right) , \tag{6.1}$$

where the infimum in (6.1) spans over all estimators $\widehat{\beta}_n$ based on n samples, while the expectation and the risk R depend the underlying distribution P. Our aim is to understand the influence on the hardness of the problem of the distribution P_X of covariates, as well as the noise level. Hence, our considered classes $\mathcal{P}$ of distributions are obtained by fixing those parameters, and letting the optimal regression parameter β^* vary freely in $\mathbf{R}^d$ (see Section 6.2).

Some minimal regularity condition on the distribution P_X is required to ensure even finitess of the minimax risk (6.1) in the random-design setting. Indeed, assume that the distribution P_X charges some positive mass on a hyperplane $H \subset \mathbf{R}^d$ (we call such a distribution *degenerate*, see Definition 6.1). Then, with positive probability, all points $X_1, \dots, X_n$ in the sample

lie within H, so that the component of the optimal parameter β^* which is orthogonal to H cannot be estimated. Note that such a component matters for out-of-sample prediction, in case the point X for which one wishes to compute the prediction does not belong to H. Such a degeneracy (or quantitative variants, where P_X puts too much mass at the neighborhood of a hyperplane) turns out to be the main obstruction to achieving controlled uniform excess risk over $\mathbf{R}^d$.

The second part of this chapter (Section 6.3) is devoted to the study of the *sample covariance matrix*

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top , \tag{6.2}$$

where $X_1, \dots, X_n$ are i.i.d. samples from P_X. Indeed, upper bounds on the minimax risk require a control of relative deviations of the empirical covariance matrix $\widehat{\Sigma}_n$ with respect to its population counterpart Σ, in the form of *negative moments* of the rescaled covariance matrix $\widetilde{\Sigma}_n := \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}$, namely

$$\mathbb{E}[\lambda_{\min}(\widetilde{\Sigma}_n)^{-q}] \tag{6.3}$$

where $q \geqslant 1$ and $\lambda_{\min}(A)$ is the smallest eigenvalue of symmetric matrix A.

Control of lower relative deviations of $\widehat{\Sigma}_n$ with respect to Σ can be expressed in terms of lower-tail bounds, of the form

$$\mathbb{P}\big(\lambda_{\min}(\widetilde{\Sigma}_n) \leqslant t\big) \leqslant \delta\,, \tag{6.4}$$

where $t, \delta \in (0,1)$. Sub-Gaussian tail bounds for $\lambda_{\min}(\widetilde{\Sigma}_n)$, of the form (6.4) with

$$\delta = \exp\Big(-cn\Big(1 - C\sqrt{\frac{d}{n}} - t\Big)_+^2\Big)$$

for some constants c, C depending on P_X, as well as similar bounds for the largest eigenvalue $\lambda_{\max}(\widetilde{\Sigma}_n)$, can be obtained under the (strong) assumption that X is sub-Gaussian (see, e.g., Vershynin, 2012). Remarkably, it has been shown by Oliveira (2016); Koltchinskii and Mendelson (2015) that such bounds can be obtained for the *smallest* eigenvalue under much weaker assumptions on X, namely bounded fourth moments of the linear marginals of X.

While sub-Gaussian bounds provide a precise control of deviations (6.4) for $t \in (c, 1 - C\sqrt{d/n})$ (for some constants c, C), they do not suffice to control moments of $\lambda_{\min}(\widetilde{\Sigma}_n)^{-1}$. Indeed, such bounds "saturate", in the sense that $\delta = \delta(t)$ does not tend to 0 as $t \to 0$; in other words, they provide no nonvacuous guarantee (6.4) with $t > 0$ as the confidence level $1 - \delta$ tends to 1. This prevents one from integrating such tail bounds and deduce a control of moments of the form (6.3). In Section 6.3, we complement the sub-Gaussian tail bounds by a study of non-asymptotic large deviation bounds (6.4) with $\delta = \exp(-n\psi(t))$ for small values of t, namely $t \in (0, c)$.

6.1.1 Summary of contributions

Let us provide an overview of our results on least squares regression, which appear in Section 6.2:

1. We determine the minimax excess risk in the well-specified case (where the true regression function $x \mapsto \mathbb{E}[Y|X = x]$ is linear) for every distribution P_X of features and noise

level σ^2. For some *degenerate* distributions (Definition 6.1), the minimax risk is infinite (Proposition 6.1); while for non-degenerate ones, the OLS estimator is exactly minimax (Theorem 6.1) irrespective of P_X, σ^2.

2. We express the minimax risk in terms of the distribution of *statistical leverage scores* of samples drawn from P_X (Theorem 6.2). Quite intuitively, distributions for which leverage scores are uneven are seen to be harder from a minimax point of view. We deduce from this a precise minimax lower bound of $\sigma^2 d/(n-d+1)$, valid for *every* distribution P_X of covariates. This lower bound nearly matches the $\sigma^2 d/(n-d-1)$ risk for centered Gaussian covariates, in both low ($d/n \to 0$) and moderate ($d/n \to \gamma \in (0,1)$) dimensions; hence, Gaussian covariates are almost the "easiest" ones in terms of minimax risk. This provides a counterpart to "universality" results obtained in the moderate-dimensional regime for *independent* covariates from the Marchenko-Pastur law.

3. We then turn to *upper bounds* on the minimax excess risk. Under some quantitative variant of the non-degeneracy assumption (Assumption 6.1) together with a fourth-moment condition on P_X (Assumption 6.2 or 6.3), we show that the minimax risk is finite and scales as $(1+o(1))\sigma^2 d/n$ for $d = o(n)$, both in the well-specified (Theorem 6.3) and misspecified (Proposition 6.3) cases. This shows in particular that OLS is asymptotically minimax in the misspecified case as well, as $d/n \to 0$. To the best of our knowledge, these are the first bounds on the expected risk of the OLS estimator with non-Gaussian random design.

The previous upper bounds rely on the study of the lower tail of the sample covariance matrix $\widehat{\Sigma}_n$, carried in Section 6.3. Our contributions here are the following (assuming, to simplify notations, that $\mathbb{E}[XX^\top] = I_d$):

4. First, we establish a *lower bound* on the lower tail of $\lambda_{\min}(\widehat{\Sigma}_n)$, for $d \geqslant 2$ and *any* distribution P_X such that $\mathbb{E}[XX^\top] = I_d$, of the form: $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \geqslant (ct)^{n/2}$ for some numerical constant c and every $t \in (0,1)$ (Proposition 6.4). We also exhibit a "small-ball" condition (Assumption 6.1) which is necessary to achieve similar upper bounds.

5. Under Assumption 6.1, we show a matching *upper bound* on the lower tail $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$, valid for all $t \in (0,1)$, and in particular for small t. This result (Theorem 6.4) is the core technical contribution of this chapter. Its proof relies the PAC-Bayesian technique for controlling empirical processes, which was used by Oliveira (2016) to control a different part of the lower tail; however, some non-trivial refinements (such as non-Gaussian smoothing) are needed to handle small values of t. This result can be equivalently stated as an upper bound on moments of $\lambda_{\min}(\widehat{\Sigma}_n)^{-1}$, namely $\|\lambda_{\min}(\widehat{\Sigma}_n)^{-1}\|_{L^q} = O(1)$ for $q \asymp n$ (Corollary 6.4).

6. Finally, we discuss in Section 6.3.3 the case of independent covariates. In this case, the "small-ball" condition (Assumption 6.1) holds naturally under mild regularity assumptions on the distribution of individual coordinates. A result of Rudelson and Vershynin (2014) establishes this for coordinates with bounded density; we complement it by a general anti-concentration result for linear combination of independent variables (Proposition 6.6), implying Assumption 6.1 for sufficiently "non-atomic" coordinates.

6.1.2 Related work

Linear least squares regression is a classical statistical problem, and the literature on this topic is too vast to be surveyed here; we refer to Györfi et al. (2002); Audibert and Catoni (2010); Hsu et al. (2014) (and reference therein) for a more thorough overview. Analysis of least squares regression is most standard and straightforward in the *fixed design* setting, where the covariates $X_1, \dots, X_n$ are treated as deterministic and the risk is evaluated within-sample; in this case, the expected excess risk of the OLS estimator is bounded by $\sigma^2 d/n$ (see, e.g., Wasserman, 2006).

In the random design setting considered here, a classical result (Györfi et al., 2002, Theorem 11.3) states that, if $\mathrm{Var}(\varepsilon|X) \leqslant \sigma^2$ and the true regression function $g^*(x) := \mathbb{E}[Y|X = x]$ satisfies $|g^*(X)| \leqslant L^2$ almost surely, then the risk of the (nonlinear) *truncated* ERM estimator, defined by $\widehat{g}_n^L(x) = \min(-L, \max(L, \langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle))$, is at most

$$\mathbb{E}[R(\widehat{g}_n^L)] - R(g^*) \leqslant 8\big(R(\beta^*) - R(g^*)\big) + C \max(\sigma^2, L^2)\frac{d(\log n + 1)}{n} \tag{6.5}$$

for some universal constant $C > 0$. This result is an *inexact oracle inequality*, where the risk is bounded by a constant times that of the best linear predictor β^*. Such guarantees are adequate in a nonparametric setting, where the approximation error $R(\beta^*) - R(g^*)$ of the linear model is itself of order $O(d/n)$ (Györfi et al., 2002). On the other hand, when no assumption is made on the magnitude of the approximation error, this bound does not ensure that the risk of the estimator approaches that of β^*. By contrast, in the *linear aggregation* problem as defined by Nemirovski (2000) and studied by Tsybakov (2003); Catoni (2004); Bunea et al. (2007); Audibert and Catoni (2011); Hsu et al. (2014); Lecué and Mendelson (2016); Mendelson (2015); Oliveira (2016), one seeks to obtain excess risk bounds, also called *exact* oracle inequalities (where the constant 8 in the bound (6.5) is replaced by 1), with respect to the linear class. In this setting, Tsybakov (2003) showed that the minimax rate of aggregation is of order $O(d/n)$, under boundedness assumptions on the regression function and on covariates. It is also worth noting that bounds on the regression function also implicitly constrain the optimal regression parameter to lie in some ball. This contrasts with the approach considered here, where minimax risk with respect to the full linear class is considered. Perhaps most different from the point of view adopted here is the approach from Foster (1991); Vovk (2001); Azoury and Warmuth (2001); Shamir (2015); Bartlett et al. (2015), who consider worst-case covariates (either in the individual sequences or in the agnostic learning setting) under boundedness assumptions on both covariates and outputs, and investigate achievable excess risk (or regret) bounds with respect to bounded balls in this case. By contrast, we take the distribution of covariates as given and allow the optimal regression parameter to be arbitrary, and study under which conditions on the covariates uniform bounds are achievable. Another type of non-uniform guarantees over linear classes is achieved by Ridge regression (Hoerl, 1962; Tikhonov, 1963) in the context of finite-dimensional or nonparametric reproducing kernel Hilbert spaces (Cucker and Smale, 2002a,b; De Vito et al., 2005; Caponnetto and De Vito, 2007; Smale and Zhou, 2007; Steinwart et al., 2009; Audibert and Catoni, 2011; Hsu et al., 2014), where the bounds do not depend explicitly on the dimension d, but rather on spectral properties of Σ and some norm of β^*.

This work is concerned with the expected risk. Risk bounds in probability are obtained, among others, by Audibert and Catoni (2011); Hsu et al. (2014); Hsu and Sabato (2016); Oliveira (2016); Mendelson (2015); Lecué and Mendelson (2016). While such bounds hold

with high probability, the probability is upper bounded and cannot be arbitrarily close to 1, so that they cannot be integrated to control the expected risk. Indeed, some additional regularity conditions are required in order to have finite minimax risk, as will be seen below. To the best of our knowledge, the only available uniform expected risk bounds for random-design regression are obtained in the case of Gaussian covariates, where they rely on the knowledge of the closed-form distribution of inverse covariance matrices (Stein, 1960; Breiman and Freedman, 1983; Anderson, 2003). One reason for considering the expected risk is that it is a single scalar, which can be more tightly controlled (in terms of matching upper and lower bounds) and compared across distributions than quantiles. In addition, random-design linear regression is a classical statistical problem, which justifies its precise decision-theoretic analysis. On the other hand, expected risk provides little indication on the tails of the risk in the high-confidence regime: in the case of heavy-tailed noise, the OLS estimator may perform poorly, and dedicated robust estimators may be required (see, e.g., the references in Lugosi and Mendelson, 2019).

Another line of work (El Karoui, 2013; Dicker, 2016; Donoho and Montanari, 2016; El Karoui, 2018; Dobriban and Wager, 2018; Hastie et al., 2019) considers the limiting behavior of regression procedures in the high-dimensional asymptotic regime where d, n tend to infinity at a proportional rate, with their ratio kept constant (Huber, 1973). The results in this setting take the form of a convergence in probability of the risk to a limit depending on the ratio d/n as well as the properties of β^*. With the notable exception of El Karoui (2018), the previous results hold under the assumption that the covariates are either Gaussian, or have a joint independence structure that leads to the same limiting behavior in high dimension. In this approach one also lets $d, n \to \infty$ while fixing (some property of) the parameter β^*, while here we consider non-asymptotic bounds valid for fixed n, d and uniformly over $\beta^* \in \mathbf{R}^d$.

The study of spectral properties of sample covariance matrices has a rich history (see for instance Bai and Silverstein, 2010; Anderson et al., 2010; Tao, 2012 and references therein); we refer to Rudelson and Vershynin (2010) for an overview of results (up to 2010) on the non-asymptotic control of the smallest eigenvalue of sample covariance matrices, which is the topic of Section 6.3. It is well-known (Vershynin, 2012) that sub-Gaussian tail bounds on both the smallest and the largest eigenvalues can be obtained under sub-Gaussian assumptions on the covariates. A series of work obtained control on these quantities under weaker assumptions (Adamczak et al., 2010; Mendelson and Paouris, 2014; Srivastava and Vershynin, 2013). A critical observation, which has been exploited in a series of work (Srivastava and Vershynin, 2013; Koltchinskii and Mendelson, 2015; Oliveira, 2016; Yaskov, 2014, 2015; van de Geer and Muro, 2014), is that the smallest eigenvalue can be controlled under much weaker tail assumptions than the largest one. Our study follows this line of work, but considers a different part of the lower tail, which poses some additional technical difficulties. In addition, we also provide a universal lower bound on the lower tail.

Notations. Throughout this text, the transpose of an $m \times n$ real matrix A is denoted $A^\top$, its trace $\mathrm{Tr}(A)$, and vectors in $\mathbf{R}^d$ are identified with $d \times 1$ column vectors. In addition, the coordinates of a vector $x \in \mathbf{R}^d$ are indicated as superscripts: $x = (x^j)_{1 \leqslant j \leqslant d}$. We also denote $\langle x, z \rangle := x^\top z = \sum_{j=1}^d (x^j) \cdot (z^j)$ the canonical scalar product of $x, z \in \mathbf{R}^d$, and $\|x\| := \langle x, x \rangle^{1/2}$ the associated Euclidean norm. In addition, for any symmetric and positive $d \times d$ matrix A, we define the scalar product $\langle x, z \rangle_A := \langle Ax, z \rangle$ and norm $\|x\|_A := \langle Ax, x \rangle^{1/2} = \|A^{1/2} x\|$. The $d \times d$ identity matrix is denoted I_d, while $S^{d-1} = \{x \in \mathbf{R}^d : \|x\| = 1\}$ refers to the unit

sphere. The smallest and largest eigenvalues of a symmetric matrix A are denoted $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively; if A is positive, then $\lambda_{\max}(A) = \|A\|_{\mathrm{op}}$ is the operator norm of A (with respect to $\|\cdot\|$), while $\lambda_{\min}(A) = \|A^{-1}\|_{\mathrm{op}}^{-1}$.

6.2 Exact minimax analysis of least-squares regression

This section is devoted to the minimax analysis of the linear least-squares problem, and in particular on the dependence of its hardness on the distribution P_X of covariates. In Section 6.2.1, we provide the exact minimax risk and estimator in the well-specified case, namely on the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$. In Section 6.2.2, we express the minimax risk in terms of the distribution of statistical leverage scores, and deduce a general lower bound. Finally, Section 6.2.3 provides upper bounds on the minimax risk under some regularity condition on the distribution P_X, both in the well-specified and misspecified cases.

Throughout the chapter, we assume that the covariate vector X satisfies $\mathbb{E}[\|X\|^2] < +\infty$, and denote $\Sigma = \mathbb{E}[XX^\top]$ its covariance matrix (by a slight but common abuse of terminology, we refer to Σ as the covariance matrix of X even when X is not centered). In addition, we assume that Σ is invertible, or equivalently that the support of X is not contained in any hyperplane; this assumption is not restrictive — since otherwise one can simply restrict to the span of the support of X (a linear subspace of $\mathbf{R}^d$) — and merely serves to simplify notations. Then, for every distribution of Y given X such that $\mathbb{E}[Y^2] < +\infty$, the risk $R(\beta) = \mathbb{E}[(\langle \beta, X\rangle - Y)^2]$ of any $\beta \in \mathbf{R}^d$ is finite; this risk is uniquely minimized by $\beta^* = \Sigma^{-1}\mathbb{E}[YX]$, where $\mathbb{E}[YX]$ is well-defined since $\mathbb{E}[\|YX\|] \leqslant \mathbb{E}[Y^2]^{1/2}\mathbb{E}[\|X\|^2]^{1/2} < +\infty$ by Cauchy-Schwarz's inequality. The response Y may then be written as

$$Y = \langle \beta^*, X\rangle + \varepsilon\,, \tag{6.6}$$

where ε is the *error*, with $\mathbb{E}[\varepsilon X] = \mathbb{E}[YX] - \Sigma\beta^* = 0$. The distribution P of (X, Y) is then characterized by the distribution P_X of X, the coefficient $\beta^* \in \mathbf{R}^d$ as well as the conditional distribution of ε given X, which satisfies $\mathbb{E}[\varepsilon^2] \leqslant \mathbb{E}[Y^2] < +\infty$ and $\mathbb{E}[\varepsilon X] = 0$. Now, given a distribution P_X of covariates and a bound σ^2 on the conditional second moment of the error, define the following three classes where Y is given by (6.6):

$$\begin{aligned}
\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) &= \left\{ P_{(X,Y)} : X \sim P_X,\ \beta^* \in \mathbf{R}^d,\ \varepsilon | X \sim \mathcal{N}(0, \sigma^2) \right\} \\
\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2) &= \left\{ P_{(X,Y)} : X \sim P_X,\ \beta^* \in \mathbf{R}^d,\ \mathbb{E}[\varepsilon|X] = 0,\ \mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2 \right\} \\
\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2) &= \left\{ P_{(X,Y)} : X \sim P_X,\ \beta^* \in \mathbf{R}^d,\ \mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2 \right\}.
\end{aligned} \tag{6.7}$$

The class $\mathcal{P}_{\mathrm{Gauss}}$ corresponds to the standard case of independent Gaussian noise, while $\mathcal{P}_{\mathrm{well}}$ includes all *well-specified* distributions, such that the true regression function $x \mapsto \mathbb{E}[Y|X = x]$ is linear. Finally, $\mathcal{P}_{\mathrm{mis}}$ corresponds to the general *misspecified* case, where the regression function $x \mapsto \mathbb{E}[Y|X = x]$ is not assumed to be linear.

6.2.1 Minimax analysis of linear least squares

Let us start with the following definition.

Definition 6.1. Let $n \geqslant d$. The following properties of the distribution P_X are equivalent:

1. For every linear hyperplane $H \subset \mathbf{R}^d$, $\mathbb{P}(X \in H) = 0$ (equivalently, for every $\theta \in S^{d-1}$, $\mathbb{P}(\langle \theta, X \rangle = 0) = 0$);
2. The sample covariance matrix $\widehat{\Sigma}_n$ is invertible almost surely;
3. The ordinary least-squares (OLS) estimator

$$\widehat{\beta}_n^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \sum_{i=1}^{n} (\langle \beta, X_i \rangle - Y_i)^2 \tag{6.8}$$

is uniquely defined almost surely, and equals $\widehat{\beta}_n^{\mathrm{LS}} = \widehat{\Sigma}_n^{-1} n^{-1} \sum_{i=1}^n Y_i X_i$.

When either of these properties does not hold, we say that P_X is *degenerate*.

Proof. The equivalence between the second and third points is standard: the empirical risk being convex, its global minimizers are the critical points β characterized by $\widehat{\Sigma}_n \beta = n^{-1} \sum_{i=1}^n Y_i X_i$.

We now prove that the second point implies the first, by contraposition. If $\mathbb{P}(\langle \theta, X \rangle = 0) = p > 0$ for some $\theta \in S^{d-1}$, then with probability p^n, $\langle \theta, X_i \rangle = 0$ for $i = 1, \dots, n$, so that $\widehat{\Sigma}_n \theta = n^{-1} \sum_{i=1}^n \langle \theta, X_i \rangle X_i = 0$ and thus $\widehat{\Sigma}_n$ is not invertible.

Conversely, let us now show that the first point implies the second one. Note that the latter amounts to saying that $X_1, \dots, X_n$ span $\mathbf{R}^d$ almost surely. In particular, it suffices to show it for $n = d$, which we do by showing that, almost surely, $V_k := \operatorname{span}(X_1, \dots, X_k)$ is of dimension k for $0 \leqslant k \leqslant d$, by induction on k. The case $k = 0$ is clear. Now, assume that $k \leqslant d$ and that V_{k-1} is of dimension $k - 1 \leqslant d - 1$ almost surely. Then, V_{k-1} is contained in a hyperplane of $\mathbf{R}^d$, and since X_k is independent of V_{k-1}, the first point implies that $\mathbb{P}(X_k \in V_{k-1}) = 0$, so that V_k is of dimension k almost surely. This concludes the proof. □

Remark 6.1 (Intercept). Assume that $X = (X^j)_{1 \leqslant j \leqslant d}$, where $X^d \equiv 1$ is an intercept variable. Then, the distribution P_X is degenerate if and only if there exists $\theta = (\theta^j)_{1 \leqslant j < d} \in \mathbf{R}^{d-1} \setminus \{0\}$ and $c \in \mathbf{R}$ such that $\sum_{j=1}^{d-1} \theta^j X^j = c$ with positive probability. This amounts to say that $(X^1, \dots, X^{d-1})$ belongs to some fixed affine hyperplane of $\mathbf{R}^{d-1}$ with positive probability.

The following result shows that non-degeneracy of the design distribution is necessary to obtain finite minimax risk.

Proposition 6.1 (Degenerate case)**.** *Assume that either $n < d$, or that the distribution P_X of X is degenerate, in the sense of Definition 6.1. Then, the minimax excess risk with respect to the class $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ is infinite.*

The fact that the minimax excess risk is infinite means that some dependence on the true parameter β^* (for instance, through its norm) is unavoidable in the expected risk of any estimator $\widehat{\beta}_n$. From now on and until the rest of this section, let us assume that the distribution P_X is non-degenerate, and that $n \geqslant d$. In particular, the OLS estimator is well-defined, and the empirical covariance matrix $\widehat{\Sigma}_n$ is invertible almost surely. Theorem 6.1 below provides the exact minimax excess risk and estimator in the well-specified case.

Theorem 6.1. *Assume that P_X is non-degenerate and $n \geqslant d$. The minimax risks over classes $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ coincide, and equal*

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] = \frac{\sigma^2}{n} \cdot \mathbb{E}\big[\operatorname{Tr}(\widetilde{\Sigma}_n^{-1})\big]\,, \tag{6.9}$$

where $\widetilde{\Sigma}_n = \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}$ *is the rescaled empirical covariance matrix. In addition, the OLS estimator* (6.8) *achieves at most this risk over the class* $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$*, and is therefore minimax optimal over classes* $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ *and* $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ *for every* P_X *and* σ^2*.*

The proof of Theorem 6.1 and Proposition 6.1 is provided in Section 6.4.2, and relies on simple decision-theoretic arguments. First, an upper bound (in the non-degenerate case) over $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ is obtained through the risk of the OLS estimator. Then, a matching lower bound on the minimax risk over the subclass $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ is established by considering the Bayes risk under Gaussian prior on β^* and using a monotone convergence argument.

Remark 6.2 (Linear changes of covariates). The minimax risk is invariant under invertible linear transformations of the covariates x. This can be argued a priori, by noting that the class of linear functions of x is invariant under linear changes of variables. To see it from Theorem 6.1, let $X' = AX$, where A is an invertible $d \times d$ matrix. Since $\Sigma' := \mathbb{E}[X'X'^\top]$ equals $A\Sigma A^\top$ and $\widehat{\Sigma}'_n := n^{-1}\sum_{i=1}^n X'_i X'^\top_i$ equals $A\widehat{\Sigma}_n A^\top$, we have

$$\widehat{\Sigma}'^{-1}_n \Sigma' = ((A^\top)^{-1}\widehat{\Sigma}_n^{-1}A^{-1})(A\Sigma A^\top) = (A^\top)^{-1}(\widehat{\Sigma}_n^{-1}\Sigma)A^\top\,,$$

which is conjugate to $\widehat{\Sigma}_n^{-1}\Sigma$ and hence has the same trace. By Theorem 6.1 (and since $\operatorname{Tr}(\widetilde{\Sigma}_n^{-1}) = \operatorname{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)$), this implies that the minimax risk is the same for the covariates X and X'. In particular, the minimax risk for the design X is the same as the one for $\widetilde{X} = \Sigma^{-1/2}X$.

Let us point out that the OLS estimator $\widehat{\beta}_n^{\text{LS}}$ is minimax optimal for every distribution of covariates P_X and noise level σ^2. This establishes the optimality of this procedure in a wide sense, and shows that the knowledge of neither of those properties of the distribution of (X, Y) is helpful to achieve improved risk uniformly over the linear class. On the other hand, when some additional knowledge on the optimal parameter β^* is available, OLS may no longer be optimal, and the knowledge of the noise level σ^2 may be helpful (this is for instance the case when β^* is drawn from a Gaussian prior, as in the proof of Theorem 6.1 in Section 6.4.2: the optimal estimator is then a Ridge estimator, which depends on σ^2).

Another consequence of Theorem 6.1 is that independent Gaussian noise is the least favorable noise structure (in terms of minimax risk) in the well-specified case for a given noise level σ^2.

Finally, the convexity of the map $A \mapsto \operatorname{Tr}(A^{-1})$ on positive matrices (Bhatia, 2009) implies (by Jensen's inequality combined with the identity $\mathbb{E}[\widetilde{\Sigma}_n] = I_d$) that the minimax risk (6.9) is always at least as large as $\sigma^2 d/n$, which is the minimax risk in the fixed-design case. We will however show in what follows that a strictly better lower bound can be obtained for $d \geqslant 2$.

6.2.2 Connection with leverage score and distribution-independent lower bound

In this section, we provide another expression for the minimax risk over the classes $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ and $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$, by relating it to the notion of *statistical leverage score* (Hoaglin and Welsch, 1978; Chatterjee and Hadi, 1988; Huber, 1981).

Theorem 6.2 (Minimax risk and leverage score)**.** *Under the assumptions of Theorem 6.1, the minimax risk* (6.9) *over the classes* $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ *and* $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ *is equal to*

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] = \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\right] \tag{6.10}$$

where the expectation holds over an i.i.d. sample $X_1, \dots, X_{n+1}$ *drawn from* P_X*, and where* $\widehat{\ell}_{n+1}$ *denotes the* statistical leverage score *of* X_{n+1} *among* $X_1, \dots, X_{n+1}$*, defined by:*

$$\widehat{\ell}_{n+1} = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle . \tag{6.11}$$

The leverage score $\widehat{\ell}_{n+1}$ of X_{n+1} among $X_1, \dots, X_{n+1}$ measures the influence of the response Y_{n+1} on the associated fitted value $\widehat{Y}_{n+1} = \langle \widehat{\beta}^{\mathrm{LS}}_{n+1}, X_{n+1} \rangle$: $\widehat{Y}_{n+1}$ is an affine function of Y_{n+1}, with slope $\widehat{\ell}_{n+1} = \partial \widehat{Y}_{n+1} / \partial Y_{n+1}$ (Hoaglin and Welsch, 1978; Chatterjee and Hadi, 1988). Theorem 6.2 shows that the minimax predictive risk under the distribution P_X is characterized by the distribution of leverage scores of samples drawn from this distribution. Intuitively, uneven leverage scores (with some points having high leverage) imply that the estimator $\widehat{\beta}^{\mathrm{LS}}_n$ is determined by a smaller number of points, and therefore has higher variance. This is consistent with the message from robust statistics that points with high leverage (typically seen as outliers) can be detrimental to the performance of the least squares estimator (Hoaglin and Welsch, 1978; Chatterjee and Hadi, 1988; Huber, 1981), see also Raskutti and Mahoney (2016).

Proof of Theorem 6.2. By Theorem 6.1, the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ equals, letting $X_{n+1} \sim P_X$ be independent from $X_1, \dots, X_n$:

$$\begin{aligned}
\frac{\sigma^2}{n} \cdot \mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] &= \frac{\sigma^2}{n} \cdot \mathbb{E}\big[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\big] \\
&= \sigma^2 \cdot \mathbb{E}\big[\mathrm{Tr}\big((n\widehat{\Sigma}_n)^{-1} X_{n+1} X_{n+1}^\top\big)\big] \\
&= \sigma^2 \cdot \mathbb{E}\big[\langle (n\widehat{\Sigma}_n)^{-1} X_{n+1}, X_{n+1} \rangle\big] \\
&= \sigma^2 \cdot \mathbb{E}\left[\frac{\langle (n\widehat{\Sigma}_n + X_{n+1}X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1} \rangle}{1 - \langle (n\widehat{\Sigma}_n + X_{n+1}X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1} \rangle}\right] \qquad (6.12) \\
&= \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\right],
\end{aligned}$$

where equality (6.12) follows from Lemma 6.1 below, with $S = n\widehat{\Sigma}_n$ and $v = X_{n+1}$. □

Lemma 6.1. *Let* S *be a symmetric positive* $d \times d$ *matrix, and* $v \in \mathbf{R}^d$*. Then,*

$$\langle S^{-1} v, v \rangle = \frac{\langle (S + vv^\top)^{-1} v, v \rangle}{1 - \langle (S + vv^\top)^{-1} v, v \rangle} . \tag{6.13}$$

Proof. Since $S + vv^\top \succcurlyeq S$ is positive, it is invertible, and the Sherman-Morrison formula (Horn and Johnson, 1990) shows that

$$(S + vv^\top)^{-1} = S^{-1} - \frac{S^{-1} v v^\top S^{-1}}{1 + v^\top S^{-1} v},$$

so that

$$\langle (S + vv^\top)^{-1} v, v \rangle = v^\top S^{-1} v - \frac{v^\top S^{-1} v v^\top S^{-1} v}{1 + v^\top S^{-1} v} = \langle S^{-1} v, v \rangle - \frac{\langle S^{-1} v, v \rangle^2}{1 + \langle S^{-1} v, v \rangle} = \frac{\langle S^{-1} v, v \rangle}{1 + \langle S^{-1} v, v \rangle},$$

which implies that $\langle (S + vv^\top)^{-1} v, v \rangle \in [0, 1)$. Inverting this equality yields equation (6.13). □

We now deduce from Theorem 6.2 a precise lower bound on the minimax risk (6.9), valid for every distribution of covariates P_X. By Proposition 6.1, it suffices to consider the case when $n \geqslant d$ and P_X is nondegenerate (since otherwise the minimax risk is infinite).

Corollary 6.1 (Minimax lower bound)**.** *Under the assumptions of Theorem 6.1, the minimax risk* (6.9) *over* $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ *satisfies*

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] \geqslant \frac{\sigma^2 d}{n - d + 1}\,. \tag{6.14}$$

Proof of Corollary 6.1. By Theorem 6.2, the minimax excess risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ writes:

$$\sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\right] \geqslant \sigma^2 \cdot \frac{\mathbb{E}[\widehat{\ell}_{n+1}]}{1 - \mathbb{E}[\widehat{\ell}_{n+1}]}\,, \tag{6.15}$$

where the inequality follows from the convexity of the map $x \mapsto x/(1-x) = 1 - 1/(1-x)$ on $[0, 1)$. Now, observe that, by exchangeability of $(X_1, \dots, X_{n+1})$,

$$\begin{aligned}
\mathbb{E}[\widehat{\ell}_{n+1}] &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\left[\left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_i, X_i \right\rangle\right] \\
&= \frac{1}{n+1} \mathbb{E}\left[\mathrm{Tr}\left\{ \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{n+1} X_i X_i^\top \right) \right\}\right] = \frac{d}{n+1}\,.
\end{aligned} \tag{6.16}$$

Plugging equation (6.16) into (6.15) yields the lower bound (6.14). □

Since $n - d + 1 \geqslant n$, Corollary 6.1 implies a lower bound of $\sigma^2 d/n$. The minimax risk for linear regression has been determined under some additional boundedness assumptions on Y (and thus on β^*) by Tsybakov (2003), showing that it scales as $\Theta(\sigma^2 d/n)$ up to numerical constants. The proof of the lower bound relies on information-theoretic arguments, and in particular on Fano's inequality (Tsybakov, 2009). Although widely applicable, such techniques often lead to loose constant factors. By contrast, the approach relying on Bayesian decision theory leading to Corollary 6.1 recovers the optimal leading constant, owing to the analytical tractability of the problem.

In fact, the lower bound of Corollary 6.1 is more precise than the $\sigma^2 d/n$ lower bound, in particular when the dimension d is commensurate to n. Indeed, in the case of centered Gaussian design, namely when $X \sim \mathcal{N}(0, \Sigma)$ for some positive matrix Σ, the risk of the OLS estimator (and thus, by Theorem 6.1, the minimax risk) can be computed exactly (Anderson, 2003; Breiman and Freedman, 1983), and equals

$$\mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n^{\mathrm{LS}})\big] = \frac{\sigma^2 d}{n - d - 1}\,. \tag{6.17}$$

The distribution-independent lower bound of Corollary 6.1 is very close to the above whenever $n - d \gg 1$. Hence, it is almost the best possible distribution-independent lower bound on the minimax risk. This also shows that Gaussian design is almost the easiest design distribution, in terms of minimax risk. This can be understood as follows: degeneracy (a large value of $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})$) occurs whenever the rescaled sample covariance matrix $\widetilde{\Sigma}_n$ is small in some direction; this occurs if either the direction of $\widetilde{X} = \Sigma^{-1/2} X$ is far from uniform (so that the projection

of $\widetilde{X}$ in some direction is small), or if its norm can be small. If $\widetilde{X} \sim \mathcal{N}(0, I_d)$, then $\widetilde{X}/\|\widetilde{X}\|$ is uniformly distributed on the unit sphere, while $\|\widetilde{X}\| = \sqrt{\sum_{j=1}^d (\widetilde{X}^j)^2}$ is sharply concentrated around $\sqrt{d}$: with exponential probability, $\|\widetilde{X}\| = \sqrt{d} + O(1)$ (see, e.g., Vershynin, 2018).

In particular, in the "dense" high-dimensional regime where d and n are large and commensurate, namely $d, n \to \infty$ and $d/n \to \gamma$, the lower bound of Corollary 6.1 matches the minimax risk (6.17) in the Gaussian case, which converges to $\sigma^2\gamma/(1-\gamma)$. The limit $\sigma^2\gamma/(1-\gamma)$ is also known to be *universal* in the high-dimensional regime: for covariates with independent coordinates, the excess risk converges almost surely to this limit under mild assumptions Tulino and Verdú (2004); Bai and Silverstein (2010) (note however that almost sure convergence does not imply convergence in expectation: indeed, the minimax risk may be infinite, for instance for degenerate distributions). However, the "universality" of this limit behavior is questionable (El Karoui and Kösters, 2011; El Karoui, 2018), since it relies on the independence assumption, which induces in high dimension a very specific geometry of the covariates due to the concentration of measure phenomenon (Ledoux, 2001; Boucheron et al., 2013). For instance, El Karoui (2018) obtains different limiting risks for robust regression in high dimension when considering non-independent coordinates. Corollary 6.1 shows that, if not universal, the limiting excess risk obtained in the independent case provides a *lower bound* for general design distributions.

Finally, the property of the design distribution that leads to the minimal excess risk in high dimension can be formulated succinctly in terms of leverage scores, using Theorem 6.2.

Corollary 6.2. *Let $(d_n)_{n\geqslant 1}$ be a sequence of positive integers such that $d_n/n \to \gamma \in (0,1)$, and $(P_X^{(n)})_{n\geqslant 1}$ a sequence of non-degenerate distributions on $\mathbf{R}^{d_n}$. Assume that the minimax excess risk (6.9) over $\mathcal{P}_{\mathrm{well}}(P_X^{(n)}, \sigma^2)$ converges to $\sigma^2\gamma/(1-\gamma)$. Then, the distribution of the leverage score $\widehat{\ell}_{n+1}^{(n)}$ of one sample among $n+1$ under $P_X^{(n)}$ converges in probability to γ.*

Proof. Let $\phi(x) = x/(1-x)$ for $x \in [0,1)$, and $\psi(x) := \phi(x) - \phi(\gamma) - \phi'(\gamma)(x-\gamma)$ (with $\psi(\gamma) = 0$). Since ϕ is strictly convex, $\psi(x) > 0$ for $x \neq \gamma$, and ψ is also strictly convex. Hence, ψ is decreasing on $[0,\gamma]$ and increasing on $[\gamma,1)$. In particular, for every $\varepsilon > 0$, $\eta_\varepsilon := \inf_{|x-\gamma|\geqslant\varepsilon} \psi(x) > 0$.

By Theorem 6.2, the assumption of Corollary 6.2 means that $\mathbb{E}[\phi(\widehat{\ell}_{n+1}^{(n)})] \to \phi(\gamma)$. Since in addition $\mathbb{E}[\widehat{\ell}_{n+1}^{(n)}] = d_n/(n+1) \to \gamma$ (the first equality, used in the proof of Corollary 6.1, holds for $d_n \leqslant n+1$, hence for n large enough since $\gamma < 1$), we have $\mathbb{E}[\psi(\widehat{\ell}_{n+1}^{(n)})] \to 0$. Now, for every $\varepsilon > 0$, $\psi(x) \geqslant \eta_\varepsilon \cdot \mathbf{1}(|x-\gamma| \geqslant \varepsilon)$, so that $\mathbb{P}(|\widehat{\ell}_{n+1}^{(n)} - \gamma| \geqslant \varepsilon) \leqslant \eta_\varepsilon^{-1}\mathbb{E}[\psi(\widehat{\ell}_{n+1}^{(n)})] \to 0$. □

6.2.3 Upper bounds on the minimax risk

In this section, we complement the lower bound of Corollary 6.1 by providing matching *upper bounds* on the minimax risk. Since by Proposition 6.1 the minimax risk is infinite when the design distribution is degenerate, some condition is required in order to control this quantity. We therefore introduce the following quantitative version of the non-degeneracy condition:

Assumption 6.1 (Small-ball condition)**.** The whitened design $\widetilde{X} := \Sigma^{-1/2}X$ satisfies the following: there exist constants $C \geqslant 1$ and $\alpha \in (0,1]$ such that, for every linear hyperplane H of $\mathbf{R}^d$ and $t > 0$,

$$\mathbb{P}\big(\mathrm{dist}(\widetilde{X}, H) \leqslant t\big) \leqslant (Ct)^\alpha . \tag{6.18}$$

Equivalently, for every $\theta \in \mathbf{R}^d \setminus \{0\}$ and $t > 0$,

$$\mathbb{P}\big(|\langle \theta, X\rangle| \leqslant t\|\theta\|_\Sigma\big) \leqslant (Ct)^\alpha\,. \tag{6.19}$$

Note that the equivalence between (6.18) and (6.19) comes from the fact that the distance $\mathrm{dist}(\widetilde{X}, H)$ of $\widetilde{X}$ to the hyperplane H equals $|\langle \theta', \widetilde{X}\rangle|$, where $\theta' \in S^{d-1}$ is a normal vector to H. Condition (6.19) is then recovered by letting $\theta = \Sigma^{-1/2}\theta'$ (such that $\|\theta\|_\Sigma = \|\theta'\| = 1$) and by homogeneity.

Assumption 6.1 states that $\widetilde{X}$ does not lie too close to any fixed hyperplane. This assumption is a strengthened variant of the "small ball" condition introduced by Koltchinskii and Mendelson (2015); Mendelson (2015); Lecué and Mendelson (2016) in the analysis of sample covariance matrices and least squares regression, which amounts to assuming (6.19) for a *single* value of $t < C^{-1}$. This latter condition amounts to a uniform equivalence between the L^1 and L^2 norms of one-dimensional marginals $\langle \theta, X\rangle$ ($\theta \in \mathbf{R}^d$) of X (Koltchinskii and Mendelson, 2015). Here, we require that the condition holds for arbitrarily small t; the reason for this is that in order to control the minimax excess risk (6.9) (and thus $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$), we are lead to control the lower tail of the rescaled covariance matrix $\widetilde{\Sigma}_n$ at all confidence levels. The study of the lower tail of $\widetilde{\Sigma}_n$ (on which the results of this section rely) is deferred to Section 6.3. We also illustrate Assumption 6.1 in Section 6.3.3, by discussing conditions under which it holds in the case of independent coordinates.

First, Assumption 6.1 itself suffices to obtain an upper bound on the minimax risk of $O(\sigma^2 d/n)$, without additional assumptions on the upper tail of $XX^\top$ (apart from integrability).

Proposition 6.2. *If Assumption 6.1 holds, then for every $P \in \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$, letting $C' := 3C^4 e^{1+9/\alpha}$ we have:*

$$\mathbb{E}[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})] \leqslant 2C' \cdot \frac{\sigma^2 d}{n}\,. \tag{6.20}$$

Proposition 6.2 (which is a consequence of Corollary 6.4 from Section 6.3.2) is optimal in terms of the rate of convergence; however, it exhibits the suboptimal $2C'$ factor in the leading term. As we show next, it is possible to obtain an optimal constant in the first-order term (as well as a second-order term of the correct order) under a modest additional assumption.

Assumption 6.2 (Norm kurtosis)**.** $\mathbb{E}[\|\Sigma^{-1/2}X\|^4] \leqslant \kappa d^2$ for some $\kappa > 0$.

Remark 6.3. Since $\mathbb{E}[\|\Sigma^{-1/2}X\|^2] = d$, Assumption 6.2 is a bound on the kurtosis of the variable $\|\Sigma^{-1/2}X\|$. This condition is implied by the following L^2-L^4 equivalence for one-dimensional marginals of X: for every $\theta \in \mathbf{R}^d$, $\mathbb{E}[\langle \theta, X\rangle^4]^{1/4} \leqslant \kappa^{1/4} \cdot \mathbb{E}[\langle \theta, X\rangle^2]^{1/2}$ (Assumption 6.3 below). Indeed, assuming that the latter holds, then taking $\theta = \Sigma^{-1/2}e_j$ (where $(e_j)_{1\leqslant j\leqslant d}$ denotes the canonical basis of $\mathbf{R}^d$), so that $\langle \theta, X\rangle$ is the j-th coordinate $\widetilde{X}^j$ of $\widetilde{X}$, we get $\mathbb{E}[(\widetilde{X}^j)^4] \leqslant \kappa \widetilde{\mathbb{E}}[(\widetilde{X}^j)^2]^2 = \kappa$ (since $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = I_d$). This implies that

$$\begin{aligned}\mathbb{E}\big[\|\widetilde{X}\|^4\big] &= \mathbb{E}\bigg[\bigg(\sum_{j=1}^d (\widetilde{X}^j)^2\bigg)^2\bigg] = \sum_{1\leqslant j,k\leqslant d} \mathbb{E}\big[(\widetilde{X}^j)^2(\widetilde{X}^k)^2\big] \\ &\leqslant \sum_{1\leqslant j,k\leqslant d} \mathbb{E}\big[(\widetilde{X}^j)^4\big]^{1/2}\mathbb{E}\big[(\widetilde{X}^k)^4\big]^{1/2} \leqslant \sum_{1\leqslant j,k\leqslant d} \kappa^{1/2}\cdot\kappa^{1/2} = \kappa\cdot d^2\,,\end{aligned}$$

where the first inequality above comes from the Cauchy-Schwarz inequality. The converse is false: if $\widetilde{X}$ is uniform on $\{\sqrt{d}e_j : 1 \leqslant j \leqslant d\}$, then the first condition holds with $\kappa = 1$, while the second only holds for $\kappa \geqslant d$ (taking $\theta = e_1$). Hence, Assumption 6.2 on the upper tail of X is weaker than an L^2-L^4 equivalence of the one-dimensional marginals of X; on the other hand, we do require a small-ball condition (Assumption 6.1) on the lower tail of X.

Theorem 6.3 (Upper bound in the well-specified case)**.** *Grant Assumptions 6.1 and 6.2. Let $C' = 3C^4 e^{1+9/\alpha}$ (which only depends on α, C). If $n \geqslant \min(6\alpha^{-1}d, 12\alpha^{-1}\log(12\alpha^{-1}))$, then*

$$\frac{1}{n}\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] \leqslant \frac{d}{n} + 8C'\kappa\Big(\frac{d}{n}\Big)^2. \tag{6.21}$$

In particular, the minimax excess risk over the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ satisfies:

$$\frac{\sigma^2 d}{n} \leqslant \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{well}}(P_X,\sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] \leqslant \frac{\sigma^2 d}{n}\Big(1 + 8C'\frac{\kappa d}{n}\Big). \tag{6.22}$$

The proof of Theorem 6.3 is given in Section 6.4.3; it relies in particular on Lemma 6.6 herein and on Theorem 6.4 from Section 6.3. As shown by the lower bound (established in Corollary 6.1), the constant in the first-order term in (6.22) is tight; in addition, one could see from a higher-order expansion (under additional moment assumptions) that the second-order term is also tight, up to the constant $8C'$ factor. This suggests that Assumption 6.2 is essentially a minimal condition on the upper tail of $XX^\top$ to obtain a second-order term in $O((d/n)^2)$.

Let us now consider the general misspecified case, namely the class $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$. Here, we will need the slightly stronger Assumption 6.3.

Assumption 6.3 (L^2-L^4 norm equivalence)**.** There exists a constant $\kappa > 0$ such that, for every $\theta \in \mathbf{R}^d$, $\mathbb{E}[\langle\theta, X\rangle^4] \leqslant \kappa \cdot \mathbb{E}[\langle\theta, X\rangle^2]^2$.

Proposition 6.3 (Upper bound in the misspecified case)**.** *Assume that P_X satisfies Assumptions 6.1 and 6.3, and that*

$$\chi := \mathbb{E}\big[\mathbb{E}[\varepsilon^2|X]^2\|\Sigma^{-1/2}X\|^4\big]/d^2 < +\infty$$

(note that $\chi \leqslant \mathbb{E}[(Y - \langle\beta^*, X\rangle)^4\|\Sigma^{-1/2}X\|^4]/d^2$). Then, for $n \geqslant \max(96, 6d)/\alpha$, the risk of the OLS estimator satisfies

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{1}{n}\mathbb{E}\big[(Y - \langle\beta^*, X\rangle)^2\|\Sigma^{-1/2}X\|^2\big] + 276C'^2\sqrt{\kappa\chi}\Big(\frac{d}{n}\Big)^{3/2}. \tag{6.23}$$

In particular, we have

$$\frac{\sigma^2 d}{n} \leqslant \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{mis}}(P_X,\sigma^2)} \mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n)\big] \leqslant \frac{\sigma^2 d}{n}\Big(1 + 276C'^2\kappa\sqrt{\frac{d}{n}}\Big). \tag{6.24}$$

The proof of Proposition 6.3 is provided in Section 6.4.4; it combines the results from Section 6.3 with a tail bound from Oliveira (2016). Proposition 6.3 shows that, under Assumptions 6.1 and 6.3, minimax excess risk over the class $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$ scales as $(1+o(1))\sigma^2 d/n$ as $d/n \to 0$. It also shows that the OLS estimator is asymptotically minimax on the misspecified class $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$ as $d = o(n)$, and that independent Gaussian noise is asymptotically the least favorable structure for the error ε.

6.2.4 Parameter estimation

Let us briefly discuss how the results of this section obtained for prediction can be adapted to the problem of parameter estimation, where the loss of an estimate $\widehat{\beta}_n$ given β^* is $\|\widehat{\beta}_n - \beta^*\|^2$.

By the same proof as that of Theorem 6.1 (replacing the norm $\|\cdot\|_\Sigma$ by $\|\cdot\|$), the minimax excess risk over the classes $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ is $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})]$, achieved by the OLS estimator. By convexity of $A \mapsto \mathrm{Tr}(A^{-1})$ over positive matrices (Bhatia, 2009), this quantity is larger than $\sigma^2\mathrm{Tr}(\Sigma^{-1})/n$.

In the case of centered Gaussian covariates, $\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})] = \mathrm{Tr}(\Sigma^{-1}\mathbb{E}[\widetilde{\Sigma}_n^{-1}]) = \mathrm{Tr}(\Sigma^{-1})n/(n-d-1)$ (Anderson, 2003), so that the minimax risk is $\sigma^2\mathrm{Tr}(\Sigma^{-1})/(n-d-1)$. On the other hand, the improved lower bound for general design of Corollary 6.1 for prediction does not appear to extend to estimation. The reason for this is that the map $A \mapsto A/(1-\mathrm{Tr}(A))$ is not convex over positive matrices for $d \geqslant 2$ (where convexity is defined with respect to the positive definite order, see e.g. Boyd and Vandenberghe (2004) for a definition), although its trace is.

Finally, the results of Section 6.3 on the lower tail of $\widetilde{\Sigma}_n$ can be used to obtain upper bounds in a similar fashion as for prediction. For instance, an analogue of Proposition 6.2 can be directly obtained by bounding $\mathrm{Tr}(\widehat{\Sigma}_n^{-1}) \leqslant \lambda_{\min}(\widetilde{\Sigma}_n)^{-1} \cdot \mathrm{Tr}(\Sigma^{-1})$. Since this chapter is primarily focused on prediction, we do not elaborate further in this direction.

6.3 Bounding the lower tail of a sample covariance matrix at all probability levels

Throughout this section, up to replacing X by $\Sigma^{-1/2}X$, we assume unless otherwise stated that $\mathbb{E}[XX^\top] = I_d$. Our aim is to obtain non-asymptotic large deviation inequalities of the form:

$$\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant e^{-n\psi(t)}$$

where $\psi(t) \to \infty$ as $t \to 0^+$. Existing bounds (Vershynin, 2012; Srivastava and Vershynin, 2013; Koltchinskii and Mendelson, 2015; Oliveira, 2016) are typically sub-Gaussian bounds with $\psi(t) = c(1 - C\sqrt{d/n} - t)_+^2$ for some constants $c, C > 0$, which "saturate" for small t. In this section, we study the behavior of the large deviations for small values of t, namely $t \in (0, c)$, where $c < 1$ is a fixed constant. In Section 6.3.1, we provide a lower bound on these tail probabilities, namely an upper bound on ψ, valid for every distribution of X when $d \geqslant 2$. In Section 6.3.2, we show that Assumption 6.1 is necessary and sufficient to obtain tail bounds of the optimal order. Finally, in Section 6.3.3 we show that Assumption 6.1 is naturally satisfied in the case of independent coordinates, under a mild regularity condition on their distributions.

6.3.1 A general lower bound on the lower tail

First, Proposition 6.4 below shows that in dimension $d \geqslant 2$, the probability of deviations of $\lambda_{\min}(\widehat{\Sigma}_n)$ cannot be arbitrarily small.

Proposition 6.4. *Assume that $d \geqslant 2$. Let X be a random vector in $\mathbf{R}^d$ such that $\mathbb{E}[XX^\top] = I_d$. Then, for every $t \leqslant 1$,*

$$\sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X \rangle| \leqslant t) \geqslant 0.16 \cdot t\,, \tag{6.25}$$

and therefore

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\big) \geqslant (0.025 \cdot t)^{n/2}\,. \tag{6.26}$$

The assumption that $d \geqslant 2$ is necessary since for $d = 1$, if $X = 1$ almost surely, then $\lambda_{\min}(\widehat{\Sigma}_n) = 1$ almost surely. Proposition 6.4 is proved in Section 6.6.1 through a probabilistic argument, namely by considering a random vector θ drawn uniformly on the sphere S^{d-1}.

Proposition 6.4 shows that $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$ is at least $(Ct)^{cn}$, where $C = 0.025$ and $c = 1/2$ are absolute constants; this bound writes $e^{-n\psi(t)}$, where $\psi(t) \asymp \log(1/t)$ as $t \to 0^+$. In the following section, we address the question of obtaining matching upper bounds on this lower tail.

6.3.2 Optimal control of the lower tail

In this section, we study conditions under which an upper bound matching the lower bound from Proposition 6.4 can be obtained. We start by noting that Assumption 6.1 is necessary to obtain such bounds:

Remark 6.4 (Necessity of small ball condition). Assume that there exists $c_1, c_2 > 0$ such that $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (c_1 t)^{c_2 n}$ for all $t \in (0,1)$. Then, Lemma 6.2 below implies that $p_t \leqslant (c_1 t^2)^{c_2}$ for all $t \in (0,1)$. This means that P_X satisfies Assumption 6.1 with $C = \sqrt{c_1}$ and $\alpha = 2c_2$.

Lemma 6.2. *For $t \in (0,1)$, let $p_t = \sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X\rangle| \leqslant t)$. Then, $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \geqslant p_{\sqrt{t}}^n$.*

Proof of Lemma 6.2. Let $p < p_{\sqrt{t}}$. By definition of $p_{\sqrt{t}}$, there exists $\theta \in S^{d-1}$ such that $\mathbb{P}(\langle \theta, X\rangle^2 \leqslant t) \geqslant p$. Hence, by independence, with probability at least p^n, $\langle \theta, X_i\rangle^2 \leqslant t$ for $i = 1, \dots, n$, so that $\lambda_{\min}(\widehat{\Sigma}_n) \leqslant \langle \widehat{\Sigma}_n \theta, \theta\rangle \leqslant t$. Taking $p \to p_{\sqrt{t}}$ concludes the proof. □

As Theorem 6.4 shows, Assumption 6.1 is also sufficient to obtain an optimal control on the lower tail.

Theorem 6.4. *Let X be a random vector in $\mathbf{R}^d$. Assume that $\mathbb{E}[XX^\top] = I_d$, and that X satisfies Assumption 6.1. Then, if $n \geqslant 6d/\alpha$, we have for every $t \in (0,1)$:*

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\big) \leqslant (C't)^{\alpha n/6} \tag{6.27}$$

where $C' = 3C^4 e^{1+9/\alpha}$.

Note that Theorem 6.4 can be stated in the non-isotropic case, where $\Sigma = \mathbb{E}[XX^\top]$ is arbitrary:

Corollary 6.3. *Let X be a random vector in $\mathbf{R}^d$ such that $\mathbb{E}[\|X\|^2] < +\infty$, and let $\Sigma = \mathbb{E}[XX^\top]$. Assume that X satisfies Assumption 6.1. Then, if $d/n \leqslant \alpha/6$, for every $t \in (0,1)$, the empirical covariance matrix $\widehat{\Sigma}_n$ formed with an i.i.d. sample of size n satisfies*

$$\widehat{\Sigma}_n \succcurlyeq t\Sigma \tag{6.28}$$

with probability at least $1 - (C't)^{\alpha n/6}$, where C' is as in Theorem 6.4.

Proof of Corollary 6.3. We may assume that Σ is invertible: otherwise, we can just consider the span of the support of X, which is a subspace of $\mathbf{R}^d$ of dimension $d' \leqslant d \leqslant \alpha n/6$. Now, let $\widetilde{X} = \Sigma^{-1/2} X$; by definition, $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = I_d$, and $\widetilde{X}$ satisfies Assumption 6.1 since X does. By Theorem 6.4, with probability at least $1 - (C't)^{\alpha n/6}$, $\lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}) \geqslant t$, which amounts to $\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} \succcurlyeq tI_d$, and thus $\widehat{\Sigma}_n \succcurlyeq t\Sigma$. □

It is worth noting that Theorem 6.4 does not require any condition on the upper tail of $XX^\top$, aside from the assumption $\mathbb{E}[XX^\top] = I_d$. Indeed, as noted in Remark 6.4, it only requires the necessary Assumption 6.1. In particular, it does not require any sub-Gaussian assumption on X, similarly to the results from Koltchinskii and Mendelson (2015); Oliveira (2016); van de Geer and Muro (2014); Yaskov (2014, 2015); this owes to the fact that a sum of independent positive random variables is naturally bounded away from 0.

Remark 6.5 (Extension to random quadratic forms). Theorem 6.4 extends (up to straightforward changes in notations) to random quadratic forms $v \mapsto \langle A_i v, v\rangle$ where $A_1, \dots, A_n$ are positive semi-definite and i.i.d., with $\mathbb{E}[A_i] = I_d$ (Theorem 6.4 corresponds to the rank 1 case where $A_i = X_i X_i^\top$). On the other hand, the lower bound of Proposition 6.4 is specific to rank 1 matrices, as can be seen by considering the counterexample where $A_i = I_d$ almost surely.

Idea of the proof. The proof of Theorem 6.4 is provided in Section 6.5. It builds on the analysis of Oliveira (2016), who obtains sub-Gaussian deviation bounds under fourth moment assumptions (Assumption 6.3), although some refinements are needed to handle our considered regime (with t arbitrarily small).

The proof starts with the representation of $\lambda_{\min}(\widehat{\Sigma}_n)$ as the infimum of an empirical process:

$$\lambda_{\min}(\widehat{\Sigma}_n) = \inf_{\theta \in S^{d-1}} \langle \widehat{\Sigma}_n \theta, \theta\rangle = \inf_{\theta \in S^{d-1}} \left\{ Z(\theta) := \frac{1}{n}\sum_{i=1}^n \langle \theta, X_i\rangle^2 \right\}. \tag{6.29}$$

In order to control this infimum, a natural approach is to first control $Z(\theta)$ on a suitable finite ε-covering of S^{d-1} using Assumption 6.1, independence, and a union bound, and then to extend this control to S^{d-1} by approximation. However, this approach (see e.g. Vershynin, 2012 for a use of this argument) fails here, since the control of the approximation term would require an exponential upper bound on $\|\widehat{\Sigma}_n\|_{\mathrm{op}}$, which requires a sub-Gaussian assumption on X. Instead, as in Oliveira (2016), we use the so-called PAC-Bayesian technique McAllester, 1999b,a; Langford and Shawe-Taylor, 2003; Catoni, 2007; Audibert and Catoni, 2011. This technique enables one to control a smoothed version of the process $Z(\theta)$, namely

$$Z(\rho) := \int_{\mathbf{R}^d} \left(\frac{1}{n}\sum_{i=1}^n \langle \theta, X_i\rangle^2 \right) \rho(\mathrm{d}\theta)\,,$$

uniformly over all smoothing distributions ρ on $\mathbf{R}^d$ whose relative entropy $\mathrm{KL}(\rho, \pi)$ with respect to a fixed "prior" distribution π on $\mathbf{R}^d$ is bounded. The proof then involves controlling (i) the Laplace transform of the process; (ii) the approximation term; and (iii) the entropy term. In order to control the last two, a careful choice of the smoothing distribution (and prior) is needed.

Remark 6.6 (PAC-Bayes vs. ε-net argument). As indicated before, the use of an ε-net argument would fail here, since it would lead to an approximation term depending on $\|\widehat{\Sigma}_n\|_{\mathrm{op}}$. On the other hand, the use of a smoothing distribution which is "isotropic" and centered at a point θ enables one to obtain an approximation term in terms of $\mathrm{Tr}(\widehat{\Sigma}_n)/d$, which can be bounded after proper truncation of X (in a way that does not overly degrade Assumption 6.1).

Remark 6.7 (Choice of prior and posteriors: entropy term). The PAC-Bayesian technique is classically employed in conjunction with Gaussian prior and smoothing distribution (Langford and Shawe-Taylor, 2003; Audibert and Catoni, 2011; Oliveira, 2016). This choice is convenient

here, since both the approximation and entropy term have closed-form expressions (in addition, a Gaussian distribution centered at θ yields the desired "isotropic" approximation term).

However, in order to obtain non-vacuous bounds for arbitrarily small t, we need the approximation term (and thus the "diameter" γ of the smoothing distribution) to be arbitrarily small. But as $\gamma \to 0$, the entropy term for Gaussian distributions grows too rapidly (as d/γ^2, instead of the $d\log(1/\gamma)$ rate suggested by covering numbers), which ultimately leads to vacuous bounds. In order to bypass this difficulty, we employ a more refined choice of prior and smoothing distributions, which leads to an optimal entropy term of $d\log(1/\gamma)$. In addition, as can be shown through symmetry arguments, this choice of smoothing also leads to an "isotropic" approximation term controlled by $\mathrm{Tr}(\widehat{\Sigma}_n)/d$ instead of $\|\widehat{\Sigma}_n\|_{\mathrm{op}}$.

Formulation in terms of moments. The statements of this section on the lower tail of $\lambda_{\min}(\widehat{\Sigma}_n)$ can equivalently be rephrased in terms of its negative moments. For $q \geqslant 1$, we denote $\|Z\|_{L^q} := \mathbb{E}[|Z|^q]^{1/q} \in [0, +\infty]$ the L^q norm of a real random variable Z.

Corollary 6.4. *Under the assumptions of Theorem 6.4 and for $n \geqslant 12/\alpha$, then for any $1 \leqslant q \leqslant \alpha n/12$,*

$$\|\max(1, \lambda_{\min}(\widehat{\Sigma}_n)^{-1})\|_{L^q} \leqslant 2^{1/q} \cdot C' \,. \tag{6.30}$$

Conversely, the previous inequality implies that $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (2C')^{\alpha n/12}$ for all $t \in (0,1)$.

Finally, for any random vector X in $\mathbf{R}^d$, $d \geqslant 2$, such that $\mathbb{E}[XX^\top] = I_d$, we have for any $q \geqslant n/2$:

$$\|\lambda_{\min}(\widehat{\Sigma}_n)^{-1}\|_{L^q} = +\infty \,.$$

The proof of Corollary 6.4 is provided in Section 6.6.2.

6.3.3 The small-ball condition for independent covariates

We now discuss conditions under which the small-ball condition (Assumption 6.1) holds in the case of independent coordinates. In this section, we assume that the coordinates X^j, $1 \leqslant j \leqslant d$, of $X = \widetilde{X}$ are independent. Note that the condition $\mathbb{E}[XX^\top] = I_d$ means that the X^j are centered and with unit variance.

Let us introduce the *Lévy concentration function* $Q_Z : \mathbf{R}^+ \to [0,1]$ of a real random variable Z defined by, for $t \geqslant 0$,

$$Q_Z(t) := \sup_{a \in \mathbf{R}} \mathbb{P}(|Z - a| \leqslant t) \,.$$

Anti-concentration (or small ball probabilities) Nguyen and Vu (2013) refers to nonvacuous upper bounds on this function. Here, in order to establish Assumption 6.1, it suffices to show that $Q_{\langle \theta, X\rangle}(t) \leqslant (Ct)^\alpha$ for all $t > 0$ and $\theta \in S^{d-1}$. We are thus lead to establish anti-concentration of linear combinations of independent variables $\langle \theta, X\rangle = \sum_{j=1}^d \theta^j X^j$, uniformly over $\theta \in S^{d-1}$, namely to provide upper bounds on:

$$Q_X(t) := \sup_{\theta \in S^{d-1}} Q_{\langle \theta, X\rangle}(t) \,.$$

Small-ball probabilities naturally appear in the study of the smallest singular value of a random matrix (see Rudelson and Vershynin, 2010). Tao and Vu (2009a,b); Rudelson and Vershynin (2008, 2009) studied anti-concentration for variables of the form $\langle \theta, X\rangle$, and deduced estimates

of the smallest singular value of random matrices. These bounds are however slightly different from the one we need: indeed, they hold for "unstructured" vectors θ (which do not have additive structure, see Rudelson and Vershynin, 2010), rather than uniformly over $\theta \in S^{d-1}$. Here, in order to show that Assumption 6.1 holds, we need bounds over Q_X, which requires some additional assumption on the distribution of the coordinates X^j.

Clearly, $Q_X \geqslant \max_{1\leqslant j\leqslant d} Q_{X^j}$, and in particular we need the coordinates X^j themselves to exhibit anti-concentration. Remarkably, a result of Rudelson and Vershynin (2014) (building on a reduction by Rogozin, 1987 to uniform variables) shows that, if the X^j have bounded densities, a reverse inequality holds:

Proposition 6.5 (Rudelson and Vershynin, 2014, Theorem 1.2)**.** *Assume that $X^1, \dots, X^d$ are independent and have density bounded by $C_0 > 0$. Then, for every $\theta \in \mathbf{R}^d$, $\sum_{j=1}^d \theta^j X^j$ has density bounded by $\sqrt{2}\, C_0$. In other words, $Q_X(t) \leqslant 2\sqrt{2}\, C_0 t$ for every $t > 0$, i.e., Assumption 6.1 holds with $\alpha = 1$ and $C = 2\sqrt{2}\, C_0$.*

Equivalently, if $\max_{1\leqslant j\leqslant d} Q_{X^j}(t) \leqslant Ct$ for all $t > 0$, then $Q_X(t) \leqslant \sqrt{2}Ct$ for all $t > 0$, and the constant $\sqrt{2}$ is optimal (Rudelson and Vershynin, 2014). Whether a general bound of Q_X in terms of $\max_{1\leqslant j\leqslant d} Q_{X_j}$ holds is unclear (for instance, the inequality $Q_X \leqslant \sqrt{2} \max_{1\leqslant j\leqslant d} Q_{X_j}$ does not hold, as shown by considering X^1, X^2 independent Bernoulli $1/2$ variables, and $\theta = (1/\sqrt{2}, 1/\sqrt{2})$: then $Q_{X^j}(3/8) = 1/2$ but $Q_{\langle \theta, X\rangle}(3/8) = 3/4$). While independence yields in general

$$Q_{\langle \theta, X\rangle}(t) \leqslant \min_{1\leqslant j\leqslant d} Q_{\theta^j X^j}(t) = \min_{1\leqslant j\leqslant d} Q_{X^j}(t/|\theta^j|) \leqslant \max_{1\leqslant j\leqslant d} Q_{X^j}(\sqrt{d}\cdot t)\,,$$

this bound exhibits an undesirable dependence on the dimension d.

Another way of expressing the "non-atomicity" of the distributions of coordinates X^j, which is stable through linear combinations of independent variables, is the rate of decay of their Fourier transform. Indeed, if X^j is atomic, then its characteristic function will not vanish at infinity. Proposition 6.6 below, which relies on an inequality by Esséen, provides uniform anti-concentration for one-dimensional marginals $\langle \theta, X\rangle$ in terms of the Fourier transform of the X^j, establishing Assumption 6.1 beyond bounded densities. In what follows, we denote Φ_Z the characteristic function of a real random variable Z, defined by $\Phi_Z(\xi) = \mathbb{E}[e^{i\xi Z}]$ for $\xi \in \mathbf{R}$.

Proposition 6.6. *Assume that $X^1, \dots, X^d$ are independent and that there exists constants $C_0 > 0$ and $\alpha \in (0, 1)$ such that, for every $1 \leqslant j \leqslant d$ and $\xi \in \mathbf{R}$,*

$$|\Phi_{X^j}(\xi)| \leqslant \frac{1}{(1 + |\xi|/C_0)^\alpha}\,. \tag{6.31}$$

Then, $X = (X^1, \dots, X^d)$ satisfies Assumption 6.1 with $C = 2^{1/\alpha}(2\pi)^{1/\alpha - 1}(1-\alpha)^{-1/\alpha} C_0$.

The proof relies on the following lemma.

Lemma 6.3. *Let $X^1, \dots, X^d$ be independent real random variables. Assume that there exists a sub-additive function $g : \mathbf{R}^+ \to \mathbf{R}$ such that, for every $j = 1, \dots, d$ and $\xi \in \mathbf{R}$,*

$$|\Phi_{X^j}(\xi)| \leqslant \exp(-g(\xi^2))\,.$$

Then, for every $t \in \mathbf{R}$,

$$Q_X(t) \leqslant t \cdot \int_{-2\pi/t}^{2\pi/t} \exp(-g(\xi^2))\, \mathrm{d}\xi\,. \tag{6.32}$$

Proof of Lemma 6.3. For every $\theta \in S^{d-1}$ and $\xi \in \mathbf{R}$, we have, by independence of the X^j,

$$|\Phi_{\langle\theta,X\rangle}(\xi)| = \left|\mathbb{E}\left[e^{i\xi(\theta_1 X^1 + \cdots + \theta_d X^d)}\right]\right| = \left|\mathbb{E}\left[e^{i\xi\theta_1 X^1}\right]\right| \cdots \left|\mathbb{E}\left[e^{i\xi\theta_d X^d}\right]\right| \leqslant \exp\left[-\left(g(\theta_1^2\xi^2) + \cdots + g(\theta_d^2\xi^2)\right)\right] \leqslant \exp(-g(\xi^2))\,,$$

where the last inequality uses the sub-additivity of g and the fact that $\theta_1^2 + \cdots + \theta_d^2 = \|\theta\|^2 = 1$. Lemma 6.3 then follows from Esséen's inequality Esseen (1966), which states that for any real random variable Z,

$$Q_Z(t) \leqslant t \cdot \int_{-2\pi/t}^{2\pi/t} |\Phi_Z(\xi)|\,\mathrm{d}\xi\,. \qquad \square$$

Proof of Proposition 6.6. The functions $g_1 : u \mapsto \alpha\log(1+u)$ and $g_2 : u \mapsto C_0^{-1}\sqrt{u}$ are concave functions on $\mathbf{R}^+$ taking the value 0 at 0, and therefore sub-additive. Since g_1 is also increasing, the function $g : u \mapsto g_1 \circ g_2(u) = \alpha\log(1 + C_0^{-1}\sqrt{u})$ is also sub-additive. Condition (6.31) simply writes $\Phi_{X^j}(\xi) \leqslant \exp(-g(\xi^2))$, so that by Lemma 6.3

$$Q_X(t) \leqslant t\int_{-2\pi/t}^{2\pi/t} \frac{1}{(1+|\xi|/C_0)^\alpha}\mathrm{d}\xi \leqslant 2t\int_0^{2\pi/t} \frac{\mathrm{d}\xi}{(\xi/C_0)^\alpha} = 2tC_0^\alpha\left(\frac{2\pi}{t}\right)^{1-\alpha}/(1-\alpha)\,,$$

which implies that $Q_X(t) \leqslant (Ct)^\alpha$, concluding the proof. $\square$

6.4 Proofs from Section 6.2

In this section, we gather the remaining proofs of results from Section 6.2 on least squares regression, namely those of Proposition 6.1, Theorem 6.1, Proposition 6.2, Theorem 6.3 and Proposition 6.3.

6.4.1 Preliminary: risk of Ridge and OLS estimators

We start with general expressions for the risk, which will be used several times in the proofs. Here, we assume that (X, Y) is as in Section 6.2, namely $\mathbb{E}[Y^2] < +\infty$, $\mathbb{E}[\|X\|^2] < +\infty$ and $\Sigma := \mathbb{E}[XX^\top]$ is invertible. Letting $\varepsilon := Y - \langle\beta^*, X\rangle$ denote the error, where $\beta^* := \Sigma^{-1}\mathbb{E}[YX]$ is the risk minimizer, we let $m(X) := \mathbb{E}[\varepsilon|X] = \mathbb{E}[Y|X] - \langle\beta^*, X\rangle$ denote the *misspecification* (or *approximation*) error of the linear model, and $\sigma^2(X) := \mathrm{Var}(\varepsilon|X) = \mathrm{Var}(Y|X)$ denote the conditional variance of the noise.

Lemma 6.4 (Risk of the Ridge estimator)**.** *Assume that (X, Y) is of the previous form. Let $\lambda \geqslant 0$, and assume that either $\lambda > 0$ or that P_X is non-degenerate and $n \geqslant d$. The risk of the Ridge estimator $\widehat{\beta}_{\lambda,n}$, defined by*

$$\widehat{\beta}_{\lambda,n} := \operatorname*{arg\,min}_{\beta\in\mathbf{R}^d}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - \langle\beta, X_i\rangle)^2 + \lambda\|\beta\|^2\right\} = \left(\widehat{\Sigma}_n + \lambda I_d\right)^{-1}\cdot\frac{1}{n}\sum_{i=1}^n Y_i X_i\,, \tag{6.33}$$

equals

$$\mathbb{E}\left[\mathcal{E}(\widehat{\beta}_{\lambda,n})\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n m(X_i)X_i - \lambda\beta^*\right\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\right] + \\ + \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sigma^2(X_i)\|X_i\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\right]. \tag{6.34}$$

Proof. Since $Y_i = \langle \beta^*, X_i \rangle + \varepsilon_i$ for $i = 1, \dots, n$, and since $\langle \beta^*, X_i \rangle X_i = X_i X_i^\top \beta^*$, we have

$$\frac{1}{n}\sum_{i=1}^n Y_i X_i = \widehat{\Sigma}_n \beta^* + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i \,. \tag{6.35}$$

As a result, the excess risk of $\widehat{\beta}_{\lambda,n}$ (which is well-defined by the assumptions) writes

$$\begin{aligned}
\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_{\lambda,n})\big] &= \mathbb{E}\left[\left\| (\widehat{\Sigma}_n + \lambda I_d)^{-1}\left(\widehat{\Sigma}_n \beta^* + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\right) - \beta^* \right\|_\Sigma^2\right] \\
&= \mathbb{E}\left[\left\| (\widehat{\Sigma}_n + \lambda I_d)^{-1} \cdot \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i - \lambda (\widehat{\Sigma}_n + \lambda I_d)^{-1} \beta^* \right\|_\Sigma^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i - \lambda \beta^* \right\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}} \middle| X_1, \dots, X_n\right]\right] \\
&= \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n m(X_i) X_i - \lambda \beta^* \right\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}}\right] + \\
&\quad + \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sigma^2(X_i) \|X_i\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}}\right]
\end{aligned} \tag{6.36}$$

where (6.36) is obtained by expanding and using the fact that, for $1 \leqslant i, j \leqslant n$ with $i \neq j$,

$$\begin{aligned}
\mathbb{E}\big[\varepsilon_i \varepsilon_j | X_1, \dots, X_n\big] &= m(X_i) m(X_j) \,, \\
\mathbb{E}\big[\varepsilon_i^2 | X_1, \dots, X_n\big] &= m(X_i)^2 + \sigma^2(X_i) \,.
\end{aligned}$$

□

In the special case where $\lambda = 0$, the previous risk decomposition becomes:

Lemma 6.5 (Risk of the OLS estimator)**.** *Assume that P_X is non-degenerate and $n \geqslant d$. Then,*

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] = \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n m(X_i) \widetilde{X}_i \right\|^2_{\widetilde{\Sigma}_n^{-2}}\right] + \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sigma^2(X_i) \|\widetilde{X}_i\|^2_{\widetilde{\Sigma}_n^{-2}}\right], \tag{6.37}$$

where we let $\widetilde{X}_i = \Sigma^{-1/2} X_i$ and $\widetilde{\Sigma}_n = \Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2}$.

Proof. This follows from Lemma 6.4 and the fact that, when $\lambda = 0$, for every $x \in \mathbf{R}^d$,

$$\|x\|_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}} = \|\Sigma^{-1/2} x\|_{\Sigma^{1/2}\widehat{\Sigma}_n^{-1}\Sigma\widehat{\Sigma}_n^{-1}\Sigma^{1/2}} = \|\Sigma^{-1/2} x\|_{\widetilde{\Sigma}_n^{-2}} \,.$$

□

6.4.2 Proof of Theorem 6.1 and Proposition 6.1

Upper bound on the minimax risk. First, we provide an upper bound on the maximum risk of the least-squares estimator over the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$. As in Theorem 6.1, we assume that $n \geqslant d$ and that P_X is non-degenerate. Let $(X, Y) \sim P \in \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$, so that $m(X) = 0$ and $\sigma^2(X) \leqslant \sigma^2$. It then follows from Lemma 6.5 that

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{\sigma^2}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sigma^2(X_i) \|\widetilde{X}_i\|^2_{\widetilde{\Sigma}_n^{-2}}\right] = \frac{\sigma^2}{n^2}\mathbb{E}\left[\mathrm{Tr}\left(\widetilde{\Sigma}_n^{-2} \sum_{i=1}^n \widetilde{X}_i \widetilde{X}_i^\top\right)\right] = \frac{\sigma^2}{n}\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) \,.$$

Hence, the maximum excess risk of the OLS estimator $\widehat{\beta}_n^{\mathrm{LS}}$ over the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ (and consequently, the minimax risk over this class) is at most $\sigma^2 \mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]/n$.

Lower bound on the minimax risk. We will now provide a lower bound on the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. We will in fact establish the lower bound both in the setting of Theorem 6.1 (namely, P_X is non-degenerate and $n \geqslant d$) and that of Proposition 6.1 (the remaining cases). In particular, we do not assume for now that P_X is non-degenerate or that $n \geqslant d$.

For $\beta^* \in \mathbf{R}^d$, let P_{β^*} denote the joint distribution of (X, Y) where $X \sim P_X$ and $Y = \langle \beta^*, X \rangle + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X. Now, consider the decision problem with model $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) = \{P_{\beta^*} : \beta^* \in \mathbf{R}^d\}$, decision space $\mathbf{R}^d$ and loss function $\mathcal{L}(\beta^*, \beta) = \mathcal{E}_{P_{\beta^*}}(\beta) = \|\beta - \beta^*\|_\Sigma^2$. Let $R(\beta^*, \widehat{\beta}_n) = \mathbb{E}_{\beta^*}[\mathcal{L}(\beta^*, \widehat{\beta}_n)]$ denote the risk under P_{β^*} of a decision rule $\widehat{\beta}_n$ (that is, an estimator of β^* using an i.i.d. sample of size n from P_{β^*}), namely its expected excess risk. Consider the prior $\Pi_\lambda = \mathcal{N}(0, \sigma^2/(\lambda n) I_d)$ on $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. A standard computation (see, e.g., Gelman et al., 2013) shows that the posterior $\Pi_\lambda(\cdot | (X_1, Y_1), \ldots, (X_n, Y_n))$ is $\mathcal{N}(\widehat{\beta}_{\lambda,n}, (\sigma^2/n) \cdot (\widehat{\Sigma}_n + \lambda I_d)^{-1})$. Since the loss function $\mathcal{L}$ is quadratic, the Bayes estimator under Π_λ is the expectation of the posterior, which is $\widehat{\beta}_{\lambda,n}$. Hence, using the comparison between minimax and Bayes risks:

$$\inf_{\widehat{\beta}_n} \sup_{P_{\beta^*} \in \mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)} R(\beta^*, \widehat{\beta}_n) \geqslant \inf_{\widehat{\beta}_n} \mathbb{E}_{\beta^* \sim \Pi_\lambda}\big[R(\beta^*, \widehat{\beta}_n)\big] = \mathbb{E}_{\beta^* \sim \Pi_\lambda}\big[R(\beta^*, \widehat{\beta}_{\lambda,n})\big]\,, \tag{6.38}$$

where the infimum is over all estimators $\widehat{\beta}_n$. Note that the left-hand side of (6.38) is simply the minimax excess risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. On the other hand, applying Lemma 6.4 with $m(X) = 0$ and $\sigma^2(X) = \sigma^2$ and noting that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n \|X_i\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}}\right] &= \mathbb{E}\left[\mathrm{Tr}\left\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\sum_{i=1}^n X_i X_i^\top\right\}\right] \\ &= n\,\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\big\}\big]\,,\end{aligned}$$

we obtain

$$R(\beta^*, \widehat{\beta}_{\lambda,n}) = \lambda^2\,\mathbb{E}\Big[\|\beta^*\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}}\Big] + \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\big\}\big]\,.$$

This implies that

$$\begin{aligned}\mathbb{E}_{\beta^* \sim \Pi_\lambda}\big[R(\beta^*, \widehat{\beta}_{\lambda,n})\big] = \mathbb{E}_{\beta^* \sim \Pi_\lambda}\Big[\lambda^2\,\mathbb{E}\Big[\|\beta^*\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}}\Big]\Big] + \\ + \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\big\}\big]\end{aligned} \tag{6.39}$$

where $\mathbb{E}$ simply denotes the expectation with respect to $(X_1, \ldots, X_n) \sim P_X^n$. Now, by Fubini's theorem, and since $\mathbb{E}_{\beta^* \sim \Pi_\lambda}[\beta^*(\beta^*)^\top] = \sigma^2/(\lambda n) I_d$, we have

$$\begin{aligned}&\mathbb{E}_{\beta^* \sim \Pi_\lambda}\Big[\lambda^2\,\mathbb{E}\Big[\|\beta^*\|^2_{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}}\Big]\Big] \\ &= \lambda^2 \cdot \mathbb{E}\Big[\mathbb{E}_{\beta^* \sim \Pi_\lambda}\Big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\beta^*(\beta^*)^\top\big\}\Big]\Big] \\ &= \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\lambda I_d\big\}\big]\,.\end{aligned} \tag{6.40}$$

Plugging (6.40) into (6.39) shows that the Bayes risk under Π_λ equals

$$\frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}(\widehat{\Sigma}_n + \lambda I_d)\big\}\big] = \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\big\}\big]\,. \tag{6.41}$$

Hence, by (6.38) the minimax risk is larger than $(\sigma^2/n) \cdot \mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\}]$ for every $\lambda > 0$. We will now distinguish the settings of Theorem 6.1 and Proposition 6.1.

Degenerate case. First, we assume that P_X is degenerate or that $n < d$. By Definition 6.1, with probability $p > 0$, the matrix $\widehat{\Sigma}_n$ is non-invertible. When this occurs, let $\theta \in \mathbf{R}^d$ be such that $\|\theta\| = 1$ and $\widehat{\Sigma}_n(\Sigma^{-1/2}\theta) = 0$. We then have, for every $\lambda > 0$,

$$\langle \Sigma^{-1/2}(\widehat{\Sigma}_n + \lambda I_d)\Sigma^{-1/2}\theta, \theta\rangle = 0 + \lambda\|\Sigma^{-1/2}\theta\|^2 \leqslant \lambda \cdot \lambda_{\min}^{-1},$$

where $\lambda_{\min} = \lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ. This implies that

$$\mathrm{Tr}\{\Sigma^{1/2}(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma^{1/2}\} \geqslant \lambda_{\max}(\Sigma^{1/2}(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma^{1/2}) = \lambda_{\min}^{-1}(\Sigma^{-1/2}(\widehat{\Sigma}_n+\lambda I_d)\Sigma^{-1/2}) \geqslant \frac{\lambda_{\min}}{\lambda}$$

so that

$$\frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\big\}\big] \geqslant \frac{\sigma^2}{n} \cdot p \cdot \frac{\lambda_{\min}}{\lambda}\,. \tag{6.42}$$

Recalling that the left-hand side of equation (6.42) is a lower bound on the minimax risk for every $\lambda > 0$, and noting that the right-hand side tends to $+\infty$ as $\lambda \to 0$, shows that the minimax risk is infinite as claimed in Proposition 6.1.

Non-degenerate case. Now, assume that P_X is non-degenerate and that $n \geqslant d$. By Definition 6.1, $\widehat{\Sigma}_n$ is invertible almost surely. In addition, $\mathrm{Tr}\{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma\} = \mathrm{Tr}\{(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}+\lambda\Sigma^{-1})^{-1}\}$ is decreasing in λ (since $\lambda \mapsto \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} + \lambda\Sigma^{-1}$ is increasing in λ), positive, and converges as $\lambda \to 0^+$ to $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})$. By the monotone convergence theorem, it follows that

$$\lim_{\lambda\to 0^+} \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\big\}\big] = \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big]\,, \tag{6.43}$$

where the limit in the right-hand side belongs to $(0, +\infty]$. Since the left-hand side is a lower bound on the minimax risk, the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ is larger than $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$.

Conclusion of the proof. Since $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) \subset \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$, the minimax risk over $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ is at least as large as that over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. In the case when P_X is degenerate or $n < d$, we showed that the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ is infinite, establishing Proposition 6.1. In the case when P_X is non-degenerate and $n \geqslant d$, we showed that the minimax risk over $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ is smaller than $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$ and that the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ is larger than the same quantity, so that both minimax risks agree and are equal to $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$, as claimed in Theorem 6.1.

6.4.3 Proof of Theorem 6.3

The proof starts with the following lemma.

Lemma 6.6. *For any positive symmetric $d \times d$ matrix A and $p \in [1, 2]$,*

$$\mathrm{Tr}(A^{-1}) + \mathrm{Tr}(A) - 2d \leqslant \max(1, \lambda_{\min}(A)^{-1}) \cdot \mathrm{Tr}\big(|A - I_d|^{2/p}\big)\,. \tag{6.44}$$

Proof of Lemma 6.6. Let us start by showing that, for every $a > 0$,

$$a^{-1} + a - 2 \leqslant \max(1, a^{-1}) \cdot |a - 1|^{2/p}\,. \tag{6.45}$$

Multiplying both sides of (6.45) by $a > 0$, it amounts to

$$(a-1)^2 = 1 + a^2 - 2a \leqslant \max(a,1) \cdot |a-1|^{2/p},$$

namely to $|a-1|^{2-2/p} \leqslant \max(a,1)$. For $a \in (0,2]$, this inequality holds since $|a-1| \leqslant 1$ and $2 - 2/p \geqslant 0$, so that $|a-1|^{2-2/p} \leqslant 1 \leqslant \max(a,1)$. For $a \geqslant 2$, the inequalities $|a-1| \geqslant 2$ and $2 - 2/p \leqslant 1$ imply that $|a-1|^{2-2/p} \leqslant |a-1| \leqslant a \leqslant \max(a,1)$. This establishes (6.45).

Now, let $a_1, \dots, a_d > 0$ be the eigenvalues of A. Without loss of generality, assume that $a_d = \min_j(a_j) = \lambda_{\min}(A)$. Then, by inequality (6.45) and the bound $\max(1, a_j^{-1}) \leqslant \max(1, a_d^{-1})$, we have

$$\mathrm{Tr}(A^{-1}) + \mathrm{Tr}(A) - 2d = \sum_{j=1}^{d} (a_j^{-1} + a_j - 2) \leqslant \max(1, a_d^{-1}) \sum_{j=1}^{d} |a_j - 1|^{2/p},$$

which is precisely the desired inequality (6.44). □

Proof of Theorem 6.3. Let $p \in (1,2]$ which will be determined later, and denote $q := p/(p-1)$ its complement. Applying Lemma 6.6 to $A = \widetilde{\Sigma}_n$ yields:

$$\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) + \mathrm{Tr}(\widetilde{\Sigma}_n) - 2d \leqslant \max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-1}) \cdot \mathrm{Tr}\big(|\widetilde{\Sigma}_n - I_d|^{2/p}\big).$$

Since $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n)] = d$, taking the expectation in the above bound and dividing by d yields:

$$\begin{aligned}
\frac{1}{d} \cdot \mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] - 1 &\leqslant \mathbb{E}\Big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-1}) \cdot \frac{1}{d}\mathrm{Tr}\big(|\widetilde{\Sigma}_n - I_d|^{2/p}\big)\Big] \\
&\leqslant \mathbb{E}\big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-1})^q\big]^{1/q} \cdot \mathbb{E}\Big[\Big(\frac{1}{d}\mathrm{Tr}\big(|\widetilde{\Sigma}_n - I_d|^{2/p}\big)\Big)^p\Big]^{1/p} \qquad (6.46) \\
&\leqslant \mathbb{E}\big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-q})\big]^{1/q} \cdot \mathbb{E}\Big[\frac{1}{d}\mathrm{Tr}\big((\widetilde{\Sigma}_n - I_d)^2\big)\Big]^{1/p} \qquad (6.47)
\end{aligned}$$

where (6.46) comes from Hölder's inequality, while (6.47) is obtained by noting that $x \mapsto x^p$ is convex and that $(1/d)\mathrm{Tr}(A)$ is the average of the eigenvalues of the symmetric matrix A. Next,

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{d}\mathrm{Tr}\big((\widetilde{\Sigma}_n - I_d)^2\big)\Big] &= \frac{1}{d}\mathrm{Tr}\Big\{\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}(\widetilde{X}_i\widetilde{X}_i^\top - I_d)\Big)^2\Big]\Big\} \\
&= \frac{1}{n^2 d}\mathrm{Tr}\Big\{\sum_{1\leqslant i,j\leqslant n}\mathbb{E}\big[(\widetilde{X}_i\widetilde{X}_i^\top - I_d)(\widetilde{X}_j\widetilde{X}_j^\top - I_d)\big]\Big\} \\
&= \frac{1}{nd}\mathrm{Tr}\Big\{\mathbb{E}\big[(\widetilde{X}\widetilde{X}^\top - I_d)^2\big]\Big\}, \qquad (6.48)
\end{aligned}$$

where we used in (6.48) the fact that, for $i \neq j$, $\mathbb{E}\big[(\widetilde{X}_i\widetilde{X}_i^\top - I_d)(\widetilde{X}_j\widetilde{X}_j^\top - I_d)\big] = \mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top - I_d]\mathbb{E}[\widetilde{X}_j\widetilde{X}_j^\top - I_d] = 0$. Now, observe that for $x \in \mathbf{R}^d$, $xx^\top - I_d$ has eigenvalue $\|x\|^2 - 1$ in the direction of x, and eigenvalue -1 in any orthogonal direction. It follows that

$$\mathrm{Tr}\{(xx^\top - I_d)^2\} = \big(\|x\|^2 - 1\big)^2 + (d-1)\cdot(-1)^2 = \|x\|^4 - 2\|x\|^2 + d,$$

so that (6.48) becomes, recalling that $\mathbb{E}[\|\widetilde{X}\|^2] = d$ and $\mathbb{E}[\|\widetilde{X}\|^4] \leqslant \kappa d^2$ (Assumption 6.2),

$$\mathbb{E}\Big[\frac{1}{d}\mathrm{Tr}\big((\widetilde{\Sigma}_n - I_d)^2\big)\Big] = \frac{1}{nd}\Big(\mathbb{E}[\|\widetilde{X}\|^4] - 2\mathbb{E}[\|\widetilde{X}\|^2] + d\Big) = \frac{1}{n}\Big(\frac{1}{d}\mathbb{E}[\|\widetilde{X}\|^4] - 1\Big) \leqslant \frac{\kappa d}{n}. \qquad (6.49)$$

On the other, recall that $\widetilde{X}$ satisfies Assumption 6.1 and that $n \geqslant \max(6d/\alpha, 12/\alpha)$. Hence, letting $C' \geqslant 1$ be the constant in Theorem 6.4, we have by Corollary 6.4:

$$\mathbb{E}\big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-q})\big] \leqslant 2C'^q\,. \tag{6.50}$$

Finally, plugging the bounds (6.49) and (6.50) into (6.47) and recalling that $1/p = 1 - 1/q = 1 - 2/(\alpha' n)$, we obtain

$$\frac{1}{d} \cdot \mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] - 1 \leqslant (2C'^q)^{1/q} \cdot \Big(\frac{\kappa d}{n}\Big)^{1/p} \leqslant 2C' \cdot \frac{\kappa d}{n} \cdot \Big(\frac{n}{\kappa d}\Big)^{2/(\alpha' n)}\,. \tag{6.51}$$

Now, since $\kappa = \mathbb{E}[\|\widetilde{X}\|^4]/\mathbb{E}[\|\widetilde{X}\|^2]^2 \geqslant 1$ and $d \geqslant 1$,

$$\Big(\frac{n}{\kappa d}\Big)^{2/(\alpha' n)} \leqslant n^{2/(\alpha' n)} = \exp\Big(\frac{2\log n}{\alpha' n}\Big)\,.$$

An elementary analysis shows that the function $g : x \mapsto \log x/x$ is increasing on $(0, e]$ and decreasing on $[e, +\infty)$. Hence, if $x, y > 1$ satisfy $x \geqslant y \log y \geqslant e$, then

$$\frac{\log x}{x} \leqslant \frac{\log y + \log\log y}{y \log y} \leqslant \frac{1 + e^{-1}}{y}$$

where we used $\log\log y/\log y \leqslant g(e) = e^{-1}$. Here by assumption $n \geqslant 12\alpha^{-1}\log(12\alpha^{-1}) = 2\alpha'^{-1}\log(2\alpha'^{-1})$, and thus $\log n/n \leqslant (1 + e^{-1})/(2/\alpha')$, so that

$$\Big(\frac{n}{\kappa d}\Big)^{2/(\alpha' n)} \leqslant \exp\Big(\frac{2}{\alpha'} \cdot \frac{1 + e^{-1}}{2/\alpha'}\Big) = \exp\big(1 + e^{-1}\big) \leqslant 4\,.$$

Plugging this inequality into (6.51) yields the desired bound (6.21). Equation (6.22) then follows by Theorem 6.1. □

6.4.4 Proof of Proposition 6.3

Recall that, by Lemma 6.5, we have

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] = \mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n m(X_i)\Sigma^{-1/2}X_i\Big\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg] + \frac{1}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \sigma^2(X_i)\big\|\Sigma^{-1/2}X_i\big\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg]\,. \tag{6.52}$$

Now, since $\widetilde{\Sigma}_n^{-2} \leqslant \lambda_{\min}(\widetilde{\Sigma}_n)^{-2} I_d$, we have for every random variable V_n:

$$\begin{aligned}\mathbb{E}\big[\|V_n\|^2_{\widetilde{\Sigma}_n^{-2}}\big] &\leqslant \mathbb{E}\big[\|V_n\|^2\big] + \mathbb{E}\big[\big\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\big\}_+ \cdot \|V_n\|^2\big] \\ &\leqslant \mathbb{E}\big[\|V_n\|^2\big] + \mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\}_+^2\big]^{1/2} \cdot \mathbb{E}\big[\|V_n\|^4\big]^{1/2}\,, \end{aligned} \tag{6.53}$$

where (6.53) follows from the Cauchy-Schwarz inequality. Letting $V_n = \sigma(X_i)\Sigma^{-1/2}X_i$, we obtain from (6.53)

$$\begin{aligned}&\frac{1}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \sigma^2(X_i)\big\|\Sigma^{-1/2}X_i\big\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg] \\ &\leqslant \frac{1}{n}\mathbb{E}\big[\sigma^2(X)\|\Sigma^{-1/2}X\|^2\big] + \frac{1}{n}\mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\}_+^2\big]^{1/2}\mathbb{E}\big[\sigma^4(X)\|\Sigma^{-1/2}X\|^4\big]^{1/2}\,. \end{aligned} \tag{6.54}$$

On the other hand, let $V_n = n^{-1}\sum_{i=1}^n m(X_i)\Sigma^{-1/2}X_i$; we have, since $\mathbb{E}[m(X_i)X_i] = \mathbb{E}[\varepsilon_i X_i] = 0$,

$$\begin{aligned}
\mathbb{E}\big[\|V_n\|^2\big] &= \mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n m(X_i)X_i\Big\|_{\Sigma^{-1}}^2\bigg] \\
&= \frac{1}{n^2}\sum_{1\leqslant i,j\leqslant n}\mathbb{E}\big[\,\langle m(X_i)X_i, m(X_j)X_j\rangle_{\Sigma^{-1}}\,\big] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\big[m(X_i)^2\|\Sigma^{-1/2}X_i\|^2\big] + \frac{1}{n^2}\sum_{i\neq j}\big\langle\mathbb{E}[m(X_i)X_i], \mathbb{E}[m(X_j)X_j]\big\rangle_{\Sigma^{-1}} \\
&= \frac{1}{n}\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big]\,. \qquad (6.55)
\end{aligned}$$

In addition,

$$\mathbb{E}\big[\|V_n\|^4\big] = \frac{1}{n^4}\sum_{1\leqslant i,j,k,l\leqslant n}\mathbb{E}\big[\,\langle m(X_i)X_i, m(X_j)X_j\rangle_{\Sigma^{-1}}\,\langle m(X_k)X_k, m(X_l)X_l\rangle_{\Sigma^{-1}}\,\big]\,.$$

Now, by independence and since $\mathbb{E}[m(X)X] = 0$, each term in the sum above where one index among i, j, k, l is distinct from the others cancels. We therefore have

$$\begin{aligned}
\mathbb{E}\big[\big\|V_n\big\|^4\big] &= \frac{1}{n^4}\sum_{i=1}^n\mathbb{E}\big[\|m(X_i)X_i\|_{\Sigma^{-1}}^4\big] + \frac{2}{n^4}\sum_{1\leqslant i<j\leqslant n}\mathbb{E}\big[\|m(X_i)X_i\|_{\Sigma^{-1}}^2\|m(X_j)X_j\|_{\Sigma^{-1}}^2\big] + \\
&\quad + \frac{4}{n^4}\sum_{1\leqslant i<j\leqslant n}\mathbb{E}\big[\langle m(X_i)X_i, m(X_j)X_j\rangle_{\Sigma^{-1}}^2\big] \\
&\leqslant \frac{1}{n^4}\sum_{i=1}^n\mathbb{E}\big[\|m(X_i)X_i\|_{\Sigma^{-1}}^4\big] + \frac{6}{n^4}\sum_{1\leqslant i<j\leqslant n}\mathbb{E}\big[\|m(X_i)X_i\|_{\Sigma^{-1}}^2\|m(X_j)X_j\|_{\Sigma^{-1}}^2\big] \qquad (6.56) \\
&= \frac{1}{n^3}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big] + \frac{6}{n^4}\cdot\frac{n(n-1)}{2}\cdot\mathbb{E}\big[m(X)^2\|X\|_{\Sigma^{-1}}^2\big]^2 \\
&\leqslant \frac{1}{n^3}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big] + \frac{3}{n^2}\cdot\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big]^2 \\
&\leqslant \frac{4}{n^2}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big] \qquad (6.57)
\end{aligned}$$

where (6.56) and (6.57) rely on the Cauchy-Schwarz inequality. Hence, it follows from (6.53), (6.55) and (6.57) that

$$\begin{aligned}
&\mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n m(X_i)\Sigma^{-1/2}X_i\Big\|_{\widetilde{\Sigma}_n^{-2}}^2\bigg] \\
&\leqslant \frac{1}{n}\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big] + \mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\}_+^2\big]^{1/2}\cdot\Big(\frac{4}{n^2}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big]\Big)^{1/2} \\
&\leqslant \frac{1}{n}\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big] + \frac{2}{n}\mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\}_+^2\big]^{1/2}\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big]^{1/2}\,. \qquad (6.58)
\end{aligned}$$

Plugging (6.54) and (6.58) into the decomposition (6.52) yields:

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{1}{n}\mathbb{E}\big[(m(X)^2+\sigma^2(X))\|\Sigma^{-1/2}X\|^2\big] + \frac{1}{n}\mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2}\times \\ \times\Big(\mathbb{E}\big[\sigma^4(X)\|\Sigma^{-1/2}X\|^4\big]^{1/2} + 2\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big]^{1/2}\Big) \tag{6.59}$$

Oliveira's subgaussian bound. Oliveira (2016) showed that, under Assumption 6.3, we have

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \geqslant 1-\varepsilon\big) \geqslant 1-\delta$$

provided that

$$n \geqslant \frac{81\kappa(d+2\log(2/\delta))}{\varepsilon^2}\,.$$

This can be rewritten as:

$$\mathbb{P}\Bigg(\lambda_{\min}(\widehat{\Sigma}_n) < 1 - 9\kappa^{1/2}\sqrt{\frac{(d+2\log(2/\delta))}{n}}\Bigg) \leqslant \delta\,. \tag{6.60}$$

Bound on the remaining term. Since the function $x \mapsto x^2$ is 2-Lipschitz on $[0,1]$, we have $(x^{-2}-1)_+ = (1-x^2)_+/x^2 \leqslant 2(1-x)_+/x^2$ for $x>0$, so that by Cauchy-Schwarz,

$$\mathbb{E}\big[\{\lambda_{\min}(\widehat{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2} \leqslant \mathbb{E}\Big[\frac{4\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^2}{\lambda_{\min}(\widehat{\Sigma}_n)^4}\Big]^{1/2} \\ \leqslant 2\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big]^{1/4}\mathbb{E}\big[\lambda_{\min}(\widehat{\Sigma}_n)^{-8}\big]^{1/4}\,. \tag{6.61}$$

First, note that

$$\begin{aligned}\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big] &= \int_0^\infty \mathbb{P}\big(\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4 \geqslant u\big)\mathrm{d}u \\ &= \int_0^1 \mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant 1-u^{1/4}\big)\mathrm{d}u \\ &= \int_0^1 \mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant 1-v^{1/2}\big)2v\mathrm{d}v\,. \end{aligned} \tag{6.62}$$

Now, let $v^{1/2} = 9\kappa^{1/2}\sqrt{[d+2\log(2/\delta)]/n}$, so that the bound (6.60) yields $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant 1-v^{1/2}) \leqslant \delta$. We have, equivalently,

$$\delta = 2\exp\Big(-\frac{n}{162\kappa}\Big(v - \frac{81\kappa d}{n}\Big)\Big) \leqslant 2\exp\Big(-\frac{n}{324\kappa}v\Big)$$

as long as $v \geqslant 162\kappa d/n$. Plugging this inequality into (6.62) yields

$$\begin{aligned}\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big] &\leqslant \int_0^{\min(162\kappa d/n,1)} 2v\mathrm{d}v + \int_{\min(162\kappa d/n,1)}^1 2\exp\Big(-\frac{n}{324\kappa}v\Big)2v\mathrm{d}v \\ &\leqslant \Big(\frac{162\kappa d}{n}\Big)^2 + \Big(\frac{324\kappa}{n}\Big)^2\int_0^\infty 4\exp(-w)w\mathrm{d}w \\ &= \Big(\frac{162\kappa d}{n}\Big)^2 + 4\Big(\frac{324\kappa}{n}\Big)^2 \end{aligned}$$

so that, using the inequality $(x+y)^{1/4} \leqslant x^{1/4} + y^{1/4}$,

$$\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big]^{1/4} \leqslant 9\sqrt{\frac{2\kappa d}{n}} + 18\sqrt{\frac{2\kappa}{n}} \leqslant 27\sqrt{\frac{2\kappa d}{n}}\,. \tag{6.63}$$

Also, by Corollary 6.4 and the fact that $\alpha n/12 \geqslant 8$, $\mathbb{E}[\lambda_{\min}(\widehat{\Sigma}_n)^{-8}] \leqslant 2C'^8$, so that inequality (6.61) becomes

$$\mathbb{E}\big[\{\lambda_{\min}(\widehat{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2} \leqslant 2\times 27\sqrt{\frac{2\kappa d}{n}} \times 2^{1/4}C'^2 \leqslant 92C'^2\sqrt{\frac{\kappa d}{n}}\,. \tag{6.64}$$

Final bound. Now, let $\chi > 0$ as in Proposition 6.3. Since

$$\mathbb{E}[\varepsilon^2|X] = m(X)^2 + \sigma^2(X) \geqslant \max(m(X)^2, \sigma^2(X))\,,$$

we have

$$\max\Big(\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big], \mathbb{E}\big[\sigma^4(X)\|\Sigma^{-1/2}X\|^4\big]\Big) \leqslant \mathbb{E}[\mathbb{E}[\varepsilon^2|X]^2\|\Sigma^{-1/2}X\|^4] = \chi d^2\,. \tag{6.65}$$

Putting the bounds (6.64) and (6.65) inside (6.59) yields

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] &\leqslant \frac{1}{n}\mathbb{E}\big[\big(m(X)^2+\sigma^2(X)\big)\|\Sigma^{-1/2}X\|^2\big] + \frac{1}{n}\cdot 92C'^2\sqrt{\frac{\kappa d}{n}}\cdot 3\sqrt{\chi}d \\ &= \frac{1}{n}\mathbb{E}\big[(Y-\langle\beta^*,X\rangle)^2\|\Sigma^{-1/2}X\|^2\big] + 276C'^2\sqrt{\kappa\chi}\Big(\frac{d}{n}\Big)^{3/2}\,,\end{aligned} \tag{6.66}$$

where we used the fact that $\mathbb{E}[(Y-\langle\beta^*,X\rangle)^2|X] = m(X)^2+\sigma^2(X)$. This establishes (6.23). Finally, if $P \in \mathcal{P}_{\mathrm{mis}}(P_X,\sigma^2)$, then $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$, so that

$$\chi = \mathbb{E}[\mathbb{E}[\varepsilon^2|X]^2\|\Sigma^{-1/2}X\|^4]/d^2 \leqslant \sigma^4\mathbb{E}[\|\Sigma^{-1/2}X\|^4]/d^2 \leqslant \sigma^4\kappa\,,$$

where we used the fact that $\mathbb{E}[\|\Sigma^{-1/2}X\|^4] \leqslant \kappa d^2$ by Assumption 6.3 (see Remark 6.3). Plugging this inequality, together with $\mathbb{E}[(Y-\langle\beta^*,X\rangle)^2\|\Sigma^{-1/2}X\|^2] \leqslant \sigma^2 d$, inside (6.66), yields the upper bound (6.24). This concludes the proof.

6.5 Proof of Theorem 6.4

6.5.1 Truncation and small-ball condition

The first step of the proof is to replace X by the truncated vector $X' := \big(1\wedge\frac{\sqrt{d}}{\|X\|}\big)X$; likewise, let $X_i' = \big(1\wedge\frac{\sqrt{d}}{\|X_i\|}\big)X_i$ for $1\leqslant i\leqslant n$, and $\widehat{\Sigma}_n' := n^{-1}\sum_{i=1}^n X_i'(X_i')^\top$. Note that $X'(X')^\top \preccurlyeq XX^\top$ and $\|X'\| = \sqrt{d}\wedge\|X\|$, so that $\widehat{\Sigma}_n' \preccurlyeq \widehat{\Sigma}_n$ and $\mathbb{E}[\|X'\|^2] \leqslant \mathbb{E}[\|X\|^2] = d$. It follows that $\lambda_{\min}(\widehat{\Sigma}_n') \leqslant \lambda_{\min}(\widehat{\Sigma}_n)$, hence it suffices to establish a lower bound for $\lambda_{\min}(\widehat{\Sigma}_n')$.

In addition, for every $\theta\in S^{d-1}$, $t\in(0,C^{-1})$ and $a\geqslant 1$,

$$\begin{aligned}\mathbb{P}(|\langle X',\theta\rangle|\leqslant t) &\leqslant \mathbb{P}\left(|\langle X,\theta\rangle|\leqslant at\right) + \mathbb{P}\left(\frac{\sqrt{d}}{\|X\|}\leqslant\frac{1}{a}\right) \\ &\leqslant (Cat)^\alpha + \mathbb{P}(\|X\|\geqslant a\sqrt{d}) \\ &\leqslant (Cat)^\alpha + \frac{\mathbb{E}[\|X\|^2]}{a^2 d} \qquad (6.67)\\ &= (Ct)^\alpha a^\alpha + \frac{1}{a^2} \qquad (6.68)\end{aligned}$$

where we applied Markov's inequality in (6.67). In particular, letting $a = (Ct)^{-\alpha/(2+\alpha)}$, inequality (6.68) becomes

$$\mathbb{P}(|\langle X', \theta\rangle| \leqslant t) \leqslant 2(Ct)^{2\alpha/(2+\alpha)}\,. \tag{6.69}$$

6.5.2 Concentration and PAC-Bayesian inequalities

The smallest eigenvalue $\lambda_{\min}(\widehat{\Sigma}'_n)$ of $\widehat{\Sigma}'_n$ may be written as the infimum of an empirical process indexed by the unit sphere $S^{d-1} = \{v \in \mathbf{R}^d : \|v\| = 1\}$:

$$\lambda_{\min}(\widehat{\Sigma}'_n) = \inf_{v \in S^{d-1}} \langle \widehat{\Sigma}'_n v, v\rangle = \inf_{v \in S^{d-1}} \frac{1}{n}\sum_{i=1}^{n} \langle X'_i, v\rangle^2\,.$$

Now, recall that the variables $\langle X'_i, \theta\rangle^2$ are i.i.d. and distributed as $\langle X', \theta\rangle^2$ for every $\theta \in S^{d-1}$. The inequality (6.69) on the left tail of this variable can be expressed in terms of its Laplace transform, through the following lemma:

Lemma 6.7. *Let Z be a nonnegative random variable. Assume that there exists $\alpha \in (0, 1]$ and $C > 0$ such that, for every $t \geqslant 0$, $\mathbb{P}(Z \leqslant t) \leqslant (Ct)^\alpha$. Then, for every $\lambda > 0$,*

$$\mathbb{E}[\exp(-\lambda Z)] \leqslant (C/\lambda)^\alpha\,. \tag{6.70}$$

Proof of Lemma 6.7. Since $0 \leqslant \exp(-\lambda Z) \leqslant 1$, we have

$$\mathbb{E}[\exp(-\lambda Z)] = \int_0^1 \mathbb{P}(\exp(-\lambda Z) \geqslant t)\mathrm{d}t = \int_0^1 \mathbb{P}\left(Z \leqslant \frac{\log(1/t)}{\lambda}\right)\mathrm{d}t \leqslant \int_0^1 \left(C\frac{\log(1/t)}{\lambda}\right)^\alpha \mathrm{d}t\,.$$

Now, for $u > 0$, the map $\alpha \mapsto u^\alpha = e^{\alpha \log u}$ is convex on $\mathbf{R}$, so that $u^\alpha \leqslant \alpha u + (1-\alpha)$ for $0 \leqslant \alpha \leqslant 1$. It follows that

$$\int_0^1 \log^\alpha(1/t)\mathrm{d}t \leqslant \alpha \int_0^1 (-\log t)\mathrm{d}t + (1-\alpha) = \alpha\big[-t\log t + t\big]_0^1 + (1-\alpha) = 1\,,$$

which establishes inequality (6.70). □

Here, inequality (6.69) implies that, for every $\theta \in S^{d-1}$,

$$\mathbb{P}(\langle X', \theta\rangle^2 \leqslant t) = \mathbb{P}(|\langle X', \theta\rangle| \leqslant \sqrt{t}) \leqslant 2(C\sqrt{t})^{2\alpha/(2+\alpha)} = 2(C^2 t)^{\alpha/(2+\alpha)}\,.$$

Hence, Lemma 6.7 with $Z = \langle X', \theta\rangle^2$ implies that, for every $\lambda > 0$,

$$\mathbb{E}[\exp(-\lambda\langle X', \theta\rangle^2)] \leqslant 2(C^2/\lambda)^{\alpha/(2+\alpha)}\,.$$

In other words, for $i = 1, \ldots, n$, $\mathbb{E}[\exp(Z_i(\theta))] \leqslant 1$, where, letting $\alpha' = \alpha/(2+\alpha)$, we define

$$Z_i(\theta) = -\lambda\langle X'_i, \theta\rangle^2 + \alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2$$

with $\lambda > 0$ a fixed parameter that will be optimized later. In particular, letting

$$Z(\theta) = Z_1(\theta) + \cdots + Z_n(\theta) = n\left[-\lambda\langle \widehat{\Sigma}'_n\theta, \theta\rangle + \alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2\right]\,,$$

the independence of $Z_1(\theta), \dots, Z_n(\theta)$ implies that, for every $\theta \in S^{d-1}$,

$$\mathbb{E}[\exp(Z(\theta))] = \mathbb{E}[\exp(Z_1(\theta))] \cdots \mathbb{E}[\exp(Z_n(\theta))] \leqslant 1\,. \tag{6.71}$$

The bound (6.71) controls the upper tail of $Z(\theta)$ for fixed $\theta \in \Theta$. In order to obtain a uniform control over θ, similarly to Audibert and Catoni (2011); Oliveira (2016) we will use the PAC-Bayesian technique for bounding empirical processes (McAllester, 1999b,a; Catoni, 2007). For completeness, we include a proof of Lemma 6.8 (which is a standard bound) below.

Lemma 6.8 (PAC-Bayesian deviation bound)**.** *Let Θ be a measurable space, and $Z(\theta)$, $\theta \in \Theta$, be a real-valued measurable process. Assume that $\mathbb{E}[\exp Z(\theta)] \leqslant 1$ for every $\theta \in \Theta$. Let π be a probability distribution on Θ. Then,*

$$\mathbb{P}\left(\forall \rho, \int_\Theta Z(\theta)\rho(\mathrm{d}\theta) \leqslant \mathrm{KL}(\rho, \pi) + t\right) \geqslant 1 - e^{-t}\,, \tag{6.72}$$

where ρ spans all probability measures on Θ, and $\mathrm{KL}(\rho, \pi) := \int_\Theta \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\rho \in [0, +\infty]$ is the Kullback-Leibler divergence between ρ and π, and where we define the integral in (6.72) to be $-\infty$ when the negative part is not integrable.

Proof of Lemma 6.8. By integrating the inequality $\mathbb{E}[\exp Z(\theta)] \leqslant 1$ with respect to π and using the Fubini-Tonelli theorem, we obtain

$$\mathbb{E}\left[\int_\Theta \exp Z(\theta)\pi(\mathrm{d}\theta)\right] \leqslant 1\,. \tag{6.73}$$

In addition, using the duality between the log-Laplace transform and the Kullback-Leibler divergence (see, e.g., Catoni, 2004, p. 159):

$$\log \int_\Theta \exp(Z(\theta))\pi(\mathrm{d}\theta) = \sup_\rho \left\{\int_\Theta Z(\theta)\rho(\mathrm{d}\theta) - \mathrm{KL}(\rho, \pi)\right\}$$

where the supremum spans over all probability distributions ρ over Θ, the inequality (6.73) writes

$$\mathbb{E}\left[\exp \sup_\rho \left\{\int_\Theta Z(\theta)\rho(\mathrm{d}\theta) - \mathrm{KL}(\rho, \pi)\right\}\right] \leqslant 1\,. \tag{6.74}$$

Applying Markov's inequality to (6.74) yields the desired bound (6.72). □

Here, we let $\Theta = S^{d-1}$ and $Z(\theta)$ as defined above. In addition, we take π to be the uniform distribution on S^{d-1}, and for $v \in S^{d-1}$ and $\gamma > 0$ we define $\Theta(v, \gamma) := \{\theta \in S^{d-1} : \|\theta - v\| \leqslant \gamma\}$ and let $\rho_{v,\gamma} = \pi(\Theta(v, \gamma))^{-1}\mathbf{1}(\Theta(v, \gamma)) \cdot \pi$ be the uniform distribution over $\Theta(v, \gamma)$. In this case, the PAC-Bayesian bound of Lemma 6.8 writes: for every $t > 0$, with probability at least $1 - e^{-t}$, for every $v \in S^{d-1}$ and $\gamma > 0$,

$$n\left[-\lambda F_{v,\gamma}(\widehat{\Sigma}'_n) + \alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2\right] \leqslant \mathrm{KL}(\rho_{v,\gamma}, \pi) + t\,, \tag{6.75}$$

where we define for every symmetric matrix Σ:

$$F_{v,\gamma}(\Sigma) := \int_\Theta \langle \Sigma\theta, \theta\rangle \rho_{v,\gamma}(\mathrm{d}\theta)\,. \tag{6.76}$$

6.5.3 Control of the approximation term

Now, using the symmetries of the smoothing distributions $\rho_{v,\gamma}$, we will show that, for every $\gamma > 0$, $v \in S^{d-1}$ and symmetric matrix Σ,

$$F_{v,\gamma}(\Sigma) = \big(1 - \phi(\gamma)\big)\langle \Sigma v, v\rangle + \phi(\gamma) \cdot \frac{1}{d}\mathrm{Tr}(\Sigma)\,, \tag{6.77}$$

where for $\gamma > 0$,

$$\phi(\gamma) := \frac{d}{d-1}\int_\Theta \big(1 - \langle \theta, v\rangle^2\big)\rho_{v,\gamma}(\mathrm{d}\theta) \in [0, d/(d-1)\gamma^2]\,. \tag{6.78}$$

First, note that

$$F_{v,\gamma}(\Sigma) = \mathrm{Tr}(\Sigma A_{v,\gamma})\,, \qquad \text{where} \qquad A_{v,\gamma} := \int_\Theta \theta\theta^\top \rho_{v,\gamma}(\mathrm{d}\theta)\,.$$

In addition, for every isometry $U \in O(d)$ of $\mathbf{R}^d$ and $v \in S^{d-1}$, $\gamma > 0$, the image measure $U_*\rho_{v,\gamma}$ of $\rho_{v,\gamma}$ under U is $\rho_{Uv,\gamma}$ (since U sends $\Theta(v,\gamma)$ to $\Theta(Uv,\gamma)$ and preserves the uniform distribution π on S^{d-1}). It follows that

$$UA_{v,\gamma}U^{-1} = \int_\Theta (U\theta)(U\theta)^\top \rho_{v,\gamma}(\mathrm{d}\theta) = \int_\Theta \theta\theta^\top \rho_{Uv,\gamma}(\mathrm{d}\theta) = A_{Uv,\gamma}\,. \tag{6.79}$$

In particular, $A_{v,\gamma}$ commutes with every isometry $U \in O(d)$ such that $Uv = v$. Taking U to be the orthogonal reflection with respect to $H_v := (\mathbf{R}v)^\perp$, $A_{v,\gamma}$ preserves $\ker(U - I_d) = \mathbf{R}v$ and is therefore of the form $\phi_1(v,\gamma)vv^\top + C_{v,\gamma}$ where $\phi_1(v,\gamma) \in \mathbf{R}$ and $C_{v,\gamma}$ is a symmetric operator with $C_{v,\gamma}H_v \subset H_v$ and $C_{v,\gamma}v = v$. Next, taking $U = vv^\top + U_v$ where U_v is an arbitrary isometry of H_v, it follows that $C_{v,\gamma}$ commutes on H_v with all isometries U_v, and is therefore of the form $\phi_2(v,\gamma)P_v$, where $P_v = I_d - vv^\top$ is the orthogonal projection on H_v and $\phi_2(v,\gamma) \in \mathbf{R}$. To summarize, we have:

$$A_{v,\gamma} = \phi_1(v,\gamma)vv^\top + \phi_2(v,\gamma)(I_d - vv^\top)\,.$$

Now, the identity (6.79) shows that, for every $U \in O(d)$ and v, γ, $\phi_1(Uv,\gamma) = \phi_1(v,\gamma)$ and $\phi_2(Uv,\gamma) = \phi_2(v,\gamma)$; hence, these constants do not depend on v and are simply denoted $\phi_1(\gamma), \phi_2(\gamma)$. Defining $\phi(\gamma) := d \cdot \phi_2(\gamma)$ and $\widetilde{\phi}(\gamma) := \phi_1(\gamma) - \phi_2(\gamma)$, we therefore have:

$$A_{v,\gamma} = \widetilde{\phi}(\gamma)vv^\top + \phi(\gamma) \cdot \frac{1}{d}I_d\,. \tag{6.80}$$

Next, observe that

$$\int_{S^{d-1}} \rho_{v,\gamma}\pi(\mathrm{d}v) = \pi\,; \tag{6.81}$$

this follows from the fact that the measure π' on the left-hand side of (6.81) is a probability distribution on S^{d-1} invariant under any $U \in O(d)$, since

$$U_*\pi' = \int_{S^{d-1}} U_*\rho_{v,\gamma}\pi(\mathrm{d}v) = \int_{S^{d-1}} \rho_{Uv,\gamma}\pi(\mathrm{d}v) = \int_{S^{d-1}} \rho_{v,\gamma}(U_*\pi)(\mathrm{d}v) = \int_{S^{d-1}} \rho_{v,\gamma}\pi(\mathrm{d}v) = \pi'\,.$$

Equation (6.81), together with Fubini's theorem, implies that

$$\int_{S^{d-1}} A_{v,\gamma}\pi(\mathrm{d}v) = \int_{S^{d-1}}\int_{S^{d-1}} \theta\theta^\top \rho_{v,\gamma}(\mathrm{d}\theta)\pi(\mathrm{d}v) = \int_{S^{d-1}} \theta\theta^\top \pi(\mathrm{d}\theta) =: A\,.$$

Since A commutes with isometries (by invariance of π), it is of the form cI_d with $c = \mathrm{Tr}(A)/d = (1/d)\int_{S^{d-1}} \|\theta\|^2 \pi(\mathrm{d}\theta) = 1/d$. Plugging (6.80) into the previous equality, we obtain

$$\frac{1}{d} I_d = \int_{S^{d-1}} \Big[\widetilde{\phi}(\gamma) v v^\top + \phi(\gamma) \cdot \frac{1}{d} I_d\Big] \pi(\mathrm{d}v) = \frac{1}{d}\widetilde{\phi}(\gamma) I_d + \frac{1}{d}\phi(\gamma) I_d \,,$$

so that $\widetilde{\phi}(\gamma) = 1 - \phi(\gamma)$. The decomposition (6.80) then writes:

$$A_{v,\gamma} = \big(1 - \phi(\gamma)\big) v v^\top + \phi(\gamma) \cdot \frac{1}{d} I_d \,.$$

Recalling that $F_{v,\gamma}(\Sigma) = \mathrm{Tr}(\Sigma A_{v,\gamma})$, we obtain the desired expression (6.77) for $F_{v,\gamma}$.

Finally, note that on the one hand,

$$\langle A_{v,\gamma} v, v\rangle = (1 - \phi(\gamma))\|v\|^2 + \phi(\gamma) \cdot \frac{1}{d}\|v\|^2 = 1 - \frac{d-1}{d}\phi(\gamma) \,,$$

while on the other hand:

$$\langle A_{v,\gamma} v, v\rangle = \int_{S^{d-1}} \langle \theta, v\rangle^2 \rho_{v,\gamma}(\mathrm{d}\theta) \,,$$

so that

$$\phi(\gamma) = \frac{d}{d-1} \int_{S^{d-1}} \big(1 - \langle \theta, v\rangle^2\big) \rho_{v,\gamma}(\mathrm{d}\theta) \geqslant 0 \,,$$

where we used that $\langle \theta, v\rangle^2 \leqslant 1$ by the Cauchy-Schwarz inequality.

Now, let α denote the angle between θ and v. We have $\langle \theta, v\rangle = \cos\alpha$ and $\|\theta - v\|^2 = (1 - \cos\alpha)^2 + \sin^2\alpha = 2(1 - \cos\alpha)$, so that $\langle \theta, v\rangle = 1 - \frac{1}{2}\|\theta - v\|^2$. Since $\rho_{v,\gamma}(\mathrm{d}\theta)$-almost surely, $\|\theta - v\| \leqslant \gamma$, this implies

$$1 - \langle \theta, v\rangle^2 = 1 - \Big(1 - \frac{1}{2}\|\theta - v\|^2\Big)^2 = \|\theta - v\|^2 - \frac{1}{4}\|\theta - v\|^4 \leqslant \gamma^2 \,.$$

Integrating this inequality over $\rho_{v,\gamma}$ yields $\phi(\gamma) \leqslant d/(d-1)\gamma^2$; this establishes (6.78).

6.5.4 Control of the entropy term

We now turn to the control of the entropy term in (6.75). Specifically, we will show that, for every $v \in S^{d-1}$ and $\gamma > 0$,

$$\mathrm{KL}(\rho_{v,\gamma}, \pi) \leqslant d \log\left(1 + \frac{2}{\gamma}\right) . \tag{6.82}$$

First, since $\mathrm{d}\rho_{v,\gamma}/\mathrm{d}\pi = \pi[\Theta(v,\gamma)]^{-1}$ $\rho_{v,\gamma}$-almost surely, $\mathrm{KL}(\rho_{v,\gamma}, \pi) = \log \pi[\Theta(v,\gamma)]^{-1}$. Now, let $N = N_c(\gamma, S^{d-1})$ denote the γ-covering number of S^{d-1}, namely the smallest $N \geqslant 1$ such that there exists $\theta_1, \dots, \theta_N \in S^{d-1}$ with

$$S^{d-1} = \bigcup_{i=1}^{N} \Theta(\theta_i, \gamma) \,. \tag{6.83}$$

Applying a union bound to (6.83) and using the fact that $\pi[\Theta(\theta_i,\gamma)] = \pi[\Theta(v,\gamma)]$ yields $1 \leqslant N\pi[\Theta(v,\gamma)]$, namely

$$\mathrm{KL}(\rho_{v,\gamma}, \pi) \leqslant \log N \,. \tag{6.84}$$

Now, let $N_p(\gamma, S^{d-1})$ denote the γ-packing number of S^{d-1}, which is the largest number of points in S^{d-1} with pairwise distances at least γ. We have, denoting $B^d = \{x \in \mathbf{R}^d : \|x\| \leqslant 1\}$,

$$N \leqslant N_p(\gamma, S^{d-1}) \leqslant N_p(\gamma, B^d) \leqslant \left(1 + \frac{2}{\gamma}\right)^d, \tag{6.85}$$

where the first inequality follows from a comparison of covering and packing numbers (Vershynin, 2018, Lemma 4.2.8), the second one from the inclusion $S^{d-1} \subset B^d$ and the last one from a volumetric argument (Vershynin, 2018, Lemma 4.2.13). Combining (6.84) and (6.85) establishes (6.82).

6.5.5 Conclusion of the proof

First note that, since $\|X_i'\|^2 = \|X_i\|^2 \wedge d \leqslant d$ for $1 \leqslant i \leqslant n$,

$$\mathrm{Tr}(\widehat{\Sigma}_n') = \frac{1}{n}\sum_{i=1}^n \|X_i'\|^2 \leqslant d\,. \tag{6.86}$$

Putting together the previous bounds (6.75), (6.77), (6.82) and (6.86), we get with probability $1 - e^{-nu}$, for every $v \in S^{d-1}$, $\gamma \in (0, 1/2]$,

$$\begin{aligned}\alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2 - \frac{d}{n}\log\left(1 + \frac{2}{\gamma}\right) - u &\leqslant \lambda F_{v,\gamma}(\widehat{\Sigma}_n') \\ &= \lambda\Big((1-\phi(\gamma))\langle\widehat{\Sigma}_n' v, v\rangle + \phi(\gamma)\cdot\frac{1}{d}\mathrm{Tr}(\widehat{\Sigma}_n')\Big) \\ &\leqslant \lambda\left[(1-\phi(\gamma))\langle\widehat{\Sigma}_n' v, v\rangle + \phi(\gamma)\right]\end{aligned}$$

In particular, rearranging, and using the fact that $\phi(\gamma) \leqslant 1/2$ for $\gamma \leqslant 1/2$, as well as $\phi(\gamma) \leqslant \gamma^2$ and $\lambda_{\min}(\widehat{\Sigma}_n') = \inf_v \langle\widehat{\Sigma}_n' v, v\rangle$, we get with probability $1 - e^{-nu}$,

$$\lambda_{\min}(\widehat{\Sigma}_n') \geqslant \frac{2}{\lambda}\left[\alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2 - \frac{d}{n}\log\left(1 + \frac{2}{\gamma}\right) - u\right] - 2\gamma^2 \tag{6.87}$$

We first approximately maximize the above lower bound in γ, given λ. Since $\gamma \leqslant 1/2$, $1 + 2/\gamma \leqslant 1 + 1/\gamma^2 \leqslant 5/(4\gamma^2)$. We are therefore led to minimize

$$\frac{2d}{\lambda n}\log\left(\frac{5}{4\gamma^2}\right) + 2\gamma^2$$

over $\gamma^2 \leqslant 1/4$. Now, let $\gamma^2 = d/(2\lambda n)$, which belongs to the prescribed range if

$$\lambda \geqslant \frac{2d}{n}\,. \tag{6.88}$$

For this choice of γ, the lower bound (6.87) becomes

$$\begin{aligned}\lambda_{\min}(\widehat{\Sigma}_n') &\geqslant \frac{2}{\lambda}\left[\alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2 - \frac{d}{n}\log\left(\frac{5\lambda n}{2d}\right) - u\right] - \frac{d}{\lambda n} \\ &= \frac{2}{\lambda}\left[\left(\alpha' - \frac{d}{n}\right)\log\lambda - \alpha'\log C^2 - \left\{\log 2 + \frac{d}{n}\log\left(\frac{5n}{2d}\right) + \frac{d}{2n}\right\} - u\right]\end{aligned}$$

Now, recall that by assumption, $d/n \leqslant \alpha/6 \leqslant 1/6$, so that (by monotonicity of $x \mapsto -x\log x$ on $(0, e^{-1}]$, replacing d/n by $1/6$) the term inside the braces is smaller than $c_0 = 1.3$. In addition, assume that $\lambda \geqslant C^4$, so that $\log(\lambda/C^4) \geqslant 0$; in this case, condition (6.88) is automatically satisfied, since $2d/n \leqslant 1/3 \leqslant C^4$. Finally, since $\alpha' = \alpha/(2+\alpha) \geqslant \alpha/3$ and $d/n \leqslant \alpha/6$, $\alpha' \leqslant 2(\alpha' - d/n)$ and $\alpha' - d/n \geqslant \alpha/6$, so that

$$\left(\alpha' - \frac{d}{n}\right)\log\lambda - \alpha'\log C^2 \geqslant \left(\alpha' - \frac{d}{n}\right)\log\left(\frac{\lambda}{C^4}\right) \geqslant \frac{\alpha}{6}\log\left(\frac{\lambda}{C^4}\right),$$

the previous inequalities implies that, for every $\lambda \geqslant C^4$ and $u > 0$, with probability at least $1 - e^{-nu}$,

$$\lambda_{\min}(\widehat{\Sigma}'_n) \geqslant \frac{2}{\lambda}\left[\frac{\alpha}{6}\log\left(\frac{\lambda}{C^4}\right) - c_0 - u\right] = \frac{\alpha}{3C^4}\frac{\log\lambda' - 6\alpha^{-1}(c_0+u)}{\lambda'}$$

where $\lambda' = \lambda/C^4 \geqslant 1$. A simple analysis shows that for $c \in \mathbf{R}$, the function $\lambda' \mapsto (\log\lambda' - c)/\lambda'$ admits a maximum on $(0, +\infty)$ of e^{-c-1}, reached at $\lambda' = e^{c+1}$. Here $c = 6\alpha^{-1}(c_0+u) > 0$, so that $\lambda' > e > 1$. Hence, for every $u > 0$, with probability at least $1 - e^{-nu}$,

$$\lambda_{\min}(\widehat{\Sigma}'_n) \geqslant \frac{\alpha}{3C^4}\exp\left(-1 - \frac{6(c_0+u)}{\alpha}\right) \geqslant C'^{-1}e^{-6u/\alpha} =: t\,, \tag{6.89}$$

where we let $C' := 3C^4 e^{1+9/\alpha}$ (using the fact that $6c_0 \leqslant 8$ and $1/\alpha \leqslant e^{1/\alpha}$). Inverting the bound (6.89), we obtain that for every $t < C'^{-1}$,

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}'_n) \leqslant t\big) \leqslant (C't)^{\alpha n/6}\,.$$

Since $\lambda_{\min}(\widehat{\Sigma}_n) \geqslant \lambda_{\min}(\widehat{\Sigma}'_n)$, and since the bound trivially holds for $t \geqslant C'^{-1}$, this concludes the proof.

6.6 Remaining proofs from Section 6.3

In this Section, we gather the proofs of remaining results from Section 6.3, namely Proposition 6.4 and Corollary 6.4.

6.6.1 Proof of Proposition 6.4

Proof. Let Θ be a random variable distributed uniformly on the unit sphere S^{d-1} and independent of X. We have

$$\sup_{\theta\in S^{d-1}} \mathbb{P}(|\langle\theta, X\rangle| \leqslant t) \geqslant \mathbb{E}\big[\mathbb{P}(|\langle\Theta, X\rangle| \leqslant t|\Theta)\big] = \mathbb{E}\big[\mathbb{P}(|\langle\Theta, X\rangle| \leqslant t|X)\big]\,.$$

Next, note that for every $x \in \mathbf{R}^d$, $\langle\Theta, x\rangle$ is distributed as $\|x\|\cdot\Theta_1$, where Θ_1 denotes the first coordinate of Θ. Since X is independent of Θ, the above inequality becomes

$$\sup_{\theta\in S^{d-1}} \mathbb{P}(|\langle\theta, X\rangle| \leqslant t) \geqslant \mathbb{E}\left[\mathbb{P}\left(|\Theta_1| \leqslant \frac{t}{\|X\|}\Big|X\right)\right]. \tag{6.90}$$

Now, since $\mathbb{E}[\|X\|^2] = \mathrm{Tr}(\mathbb{E}[XX^\top]) = d$, Markov's inequality implies that $\mathbb{P}(\|X\| \geqslant 2\sqrt{d}) \leqslant \mathbb{E}[\|X\|^2]/(4d) \leqslant 1/4$. Since $r \mapsto \mathbb{P}_\theta(|\theta_1| \leqslant t/r)$ is non-increasing, plugging this into (6.90) yields

$$\sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X\rangle| \leqslant t) \geqslant \frac{3}{4} \cdot \mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big)\,. \tag{6.91}$$

Let us now derive the distribution of $|\Theta_1|$. Let $\phi : S^{d-1} \to \mathbf{R}$ be the projection on the first coordinate: $\phi(\theta) = \theta_1$ for $\theta \in S^{d-1}$. Note that for $u \in [-1,1]$, $\phi^{-1}(u) = \{u\} \times (\sqrt{1-u^2} \cdot S^{d-2})$ which is isometric to $\sqrt{1-u^2} \cdot S^{d-2}$ and hence has $(d-2)$-dimensional Hausdorff measure $C_d(1-u^2)^{(d-2)/2}$ for some constant C_d. In addition, since $\phi(\theta) = \langle e_1, \theta\rangle$ (where $e_1 = (1,0,\dots,0)$), $\nabla\phi(\theta) \in (\mathbf{R}\theta)^\perp$ is the orthogonal projection of e_1 on $(\mathbf{R}\theta)^\perp$, namely $e_1 - \theta_1\theta$, with norm $\|\nabla\phi(\theta)\| = \sqrt{1-\theta_1^2}$. Fix $t \in (0,1]$ and define $g(\theta) = \mathbf{1}(|\theta_1| \leqslant t)/\sqrt{1-\theta_1^2}$, which equals $\mathbf{1}(|u| \leqslant t)/\sqrt{1-u^2}$ on $\phi^{-1}(u)$ (for $u \in (-1,1)$), and such that $g(\theta) \cdot \|\nabla\phi(\theta)\| = \mathbf{1}(|\theta_1| \leqslant t)$. Hence, the coarea formula (Federer, 1996, Theorem 3.2.2) implies that, for every $t \in (0,1]$,

$$\begin{aligned}\mathbb{P}(|\Theta_1| \leqslant t) &= \int_{S^{d-1}} g(\theta)\|\nabla\phi(\theta)\|\pi(\mathrm{d}\theta) = \int_{-1}^{1} \frac{\mathbf{1}(|u| \leqslant t)}{\sqrt{1-u^2}} \times C_d(1-u^2)^{(d-2)/2}\mathrm{d}u \\ &= 2C_d \int_0^t (1-u^2)^{(d-3)/2}\mathrm{d}u\,. \end{aligned} \tag{6.92}$$

If $d = 2$, (6.92) implies that $|\Theta_1|$ has density $(2/\pi)/\sqrt{1-t^2} \geqslant 2/\pi$ on $[0,1]$, and hence for $t \in [0,1]$:

$$\mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big) \geqslant \frac{2}{\pi} \times \frac{t}{2\sqrt{2}}\,. \tag{6.93}$$

If $d = 3$, (6.92) implies that $|\Theta_1|$ is uniformly distributed on $[0,1]$, so that for $t \in [0,1]$

$$\mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big) = \frac{t}{2\sqrt{3}}\,. \tag{6.94}$$

Now, assume that $d \geqslant 4$. Letting $t = 1$ in (6.92) yields the value of the constant C_d, which normalizes the right-hand side: since $1 - u^2 \leqslant e^{-u^2}$,

$$\begin{aligned}(2C_d)^{-1} &= \int_0^1 (1-u^2)^{(d-3)/2}\mathrm{d}u \leqslant \int_0^1 e^{-(d-3)u^2/2}\mathrm{d}u \\ &\leqslant \frac{1}{\sqrt{d-3}} \int_0^{\sqrt{d-3}} e^{-u^2/2}\mathrm{d}u \leqslant \frac{1}{\sqrt{d-3}} \times \sqrt{\frac{\pi}{2}}\,,\end{aligned}$$

so that $2C_d \geqslant \sqrt{2(d-3)/\pi}$. Finally, if $u \leqslant 1/(2\sqrt{d})$, then

$$\big(1-u^2\big)^{(d-3)/2} \geqslant \Big(1 - \frac{1}{4d}\Big)^{d/2} \geqslant \Big(1 - \frac{1}{16}\Big)^2\,,$$

using the fact that $4d \geqslant 16$ and that the function $x \mapsto (1-1/x)^{x/8}$ is increasing on $(1, +\infty)$. Plugging the above lower bounds in (6.92) shows that, for $t \leqslant 1$,

$$\mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big) = 2C_d \int_0^{t/(2\sqrt{d})} (1-u^2)^{(d-3)/2}\mathrm{d}u \geqslant \sqrt{\frac{2(d-3)}{\pi}} \times \Big(\frac{15}{16}\Big)^2 \frac{t}{2\sqrt{d}} \geqslant \frac{t}{3} \tag{6.95}$$

where the last inequality is obtained by noting that $(d-3)/d \geqslant 1/4$ for $d \geqslant 4$ and lower bounding the resulting constant. The bounds (6.93), (6.94) and (6.95) imply that, for every $d \geqslant 2$ and $t \leqslant 1$,

$$\mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big) \geqslant \frac{t}{\pi\sqrt{2}}\,. \tag{6.96}$$

The first inequality of Proposition 6.4 follows by combining inequalities (6.91) and (6.96). The second inequality (6.26) is a consequence of the first by Lemma 6.2. □

6.6.2 Proof of Corollary 6.4

Corollary 6.4 directly follows from Theorem 6.4, Proposition 6.4 and Lemma 6.9 below.

Lemma 6.9. *Let Z be a nonnegative real variable.*

1. *If there exist some constants $C \geqslant 1$ and $a \geqslant 2$ such that $\mathbb{P}(Z \leqslant t) \leqslant (Ct)^a$ for all $t > 0$, then $\|Z^{-1}\|_{L^q} \leqslant \|\max(1, Z^{-1})\|_{L^q} \leqslant 2^{1/q}C \leqslant 2C$ for all $1 \leqslant q \leqslant a/2$.*
2. *Conversely, if $\|Z^{-1}\|_{L^q} \leqslant C$ for some constants $q \geqslant 1$ and $C > 0$, then $\mathbb{P}(Z \leqslant t) \leqslant (Ct)^q$ for all $t > 0$.*
3. *Finally, if there exist constants $c, a > 0$ such that $\mathbb{P}(Z \leqslant t) \geqslant (ct)^a$ for all $t \in (0, 1)$, then $\|Z^{-1}\|_{L^q} = +\infty$ for $q \geqslant a$.*

Proof. For the first point, since $\max(1, Z^{-q})$ is nonnegative, we have

$$\mathbb{E}[\max(1, Z^{-q})] = \int_0^\infty \mathbb{P}(\max(1, Z^{-q}) \geqslant u)\,\mathrm{d}u = \int_0^\infty \mathbb{P}(\min(1, Z) \leqslant u^{-1/q})\,\mathrm{d}u\,.$$

For $u \leqslant C^q$, we bound $\mathbb{P}(\min(1, Z) \leqslant u^{-1/q}) \leqslant 1$, while for $u \geqslant C^q$ (so that $u^{-1/q} \leqslant C^{-1} \leqslant 1$), we bound $\mathbb{P}(\min(1, Z) \leqslant u^{-1/q}) = \mathbb{P}(Z \leqslant u^{-1/q}) \leqslant (Cu^{-1/q})^a$. We then conclude that

$$\|\max(1, Z^{-1})\|_{L^q}^q \leqslant C^q + \int_{C^q}^\infty (C^{-q}u)^{-a/q}\mathrm{d}u = C^q\Big[1 + \int_1^\infty v^{-a/q}\mathrm{d}v\Big] \leqslant 2C^q\,,$$

where we let $v = C^{-q}u$ and used the fact that $\int_1^\infty v^{-a/q}\mathrm{d}v \leqslant \int_1^\infty v^{-2}\mathrm{d}v = 1$ since $q \leqslant a/2$. The second point follows from Markov's inequality: for every $t > 0$,

$$\mathbb{P}(Z \leqslant t) = \mathbb{P}(Z^{-q} \geqslant t^{-q}) \leqslant t^q \cdot \mathbb{E}[Z^{-q}] \leqslant (Ct)^q\,.$$

Finally, for the third point, since $\mathbb{P}(Z \leqslant u^{-1/q}) \geqslant (cu^{-1/q})^a$ for $u > 1$, we have for $q \geqslant a$:

$$\mathbb{E}[Z^{-q}] = \int_0^\infty \mathbb{P}(Z \leqslant u^{-1/q})\mathrm{d}u \geqslant \int_1^\infty c^a u^{-a/q}\mathrm{d}u \geqslant c^a \int_1^\infty u^{-1}\mathrm{d}u = +\infty\,.$$ □

6.7 Conclusion

In this work, we conducted a detailed decision-theoretic analysis of random-design linear prediction, by providing matching upper and lower bounds on the minimax risk under weak conditions. In particular, we showed that the minimax risk is determined by the distribution of statistical leverage scores, and is approximately minimized in high dimension by centered Gaussian covariates. We also obtained the first upper bounds on the expected risk of the

ordinary least squares estimator in the random design setting with non-Gaussian covariates. Those bounds scale as most as $(1 + o(1))\sigma^2 d/n$ as $d = o(n)$ with noise level σ^2, under some mild conditions on the distribution of covariates.

The previous upper bounds relied on a study of the lower tail and negative moments of empirical covariance matrices. We showed a general lower bound on this lower tail in dimension $d \geqslant 2$, and established a matching upper bound under a necessary "small-ball" regularity condition on the design. The proof of this result relied on the use of PAC-Bayesian smoothing of empirical processes, with refined non-Gaussian smoothing distributions.

It is worth noting that our upper bound on the lower tail of $\lambda_{\min}(\widehat{\Sigma}_n)$ (Theorem 6.4) requires $n \geqslant 6d$; while we did not attempt to optimize the constant factor 6, the approach used here is not sufficient to obtain meaningful bounds for square (or nearly square) design matrices, whose aspect ratio d/n is close to 1. It would be interesting to see if the bound of Theorem 6.4 can be extended to this case (for instance with centered, variance 1 independent coordinates with bounded density, as in Section 6.3.3, or more generally under Assumption 6.1 with $\alpha = 1$), by leveraging the techniques from Rudelson and Vershynin (2008, 2009); Tao and Vu (2009b,a).

Chapter 7

An improper estimator with optimal excess risk in misspecified density estimation and logistic regression

Abstract. We introduce a procedure for predictive conditional density estimation under logarithmic loss, which we call SMP (Sample Minmax Predictor). This estimator minimizes a new general excess risk bound for supervised statistical learning. On standard examples, this bound scales as d/n with d the model dimension and n the sample size, and critically remains valid under model misspecification. Being an improper (out-of-model) procedure, SMP improves over within-model estimators such as the maximum likelihood estimator, whose excess risk degrades under misspecification. Compared to approaches reducing to the sequential problem, our bounds remove suboptimal $\log n$ factors, addressing an open problem from Grünwald and Kotłowski (2011) for the considered models, and can handle unbounded classes. For the Gaussian linear model, the predictions and risk bound of SMP are governed by leverage scores of covariates, nearly matching the optimal risk in the well-specified case without conditions on the noise variance or approximation error of the linear model. For logistic regression, SMP provides a non-Bayesian approach to calibration of probabilistic predictions relying on virtual samples, and can be computed by solving two logistic regressions. It achieves a non-asymptotic excess risk of $O((d + B^2R^2)/n)$, where R bounds the norm of features and B that of the comparison parameter; by contrast, no within-model estimator can achieve better rate than $\min(BR/\sqrt{n}, de^{BR}/n)$ in general (Hazan et al., 2014). This provides a computationally more efficient alternative to Bayesian approaches, which require approximate posterior sampling, thereby partly answering a question by Foster et al. (2018).

Contents

7.1 Introduction

Consider the standard problem of density estimation: given an i.i.d. sample $Z_1, \dots, Z_n$ from an unknown distribution P on some measurable space $\mathcal{Z}$, the goal is to produce a good approximation $\widehat{P}_n$ of P. One way to measure the quality of an estimate $\widehat{P}_n$ is through its predictive risk: given a base measure μ on $\mathcal{Z}$, the risk of a density g on $\mathcal{Z}$ with respect to μ is given by

$$R(g) = \mathbb{E}[\ell(g, Z)]\,, \qquad \text{where} \qquad \ell(g, z) = -\log g(z) \tag{7.1}$$

for $z \in \mathcal{Z}$ and where Z is a random variable with distribution P. Letting $\mathcal{G}$ denote the set of all probability densities on $\mathcal{Z}$ with respect to μ, the loss function $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbf{R}$ defined by (7.1), called *logarithmic* (or *negative log-likelihood*, *entropy* or *logistic*) loss, measures the error of the density $g \in \mathcal{G}$ (which can be interpreted as a probabilistic prediction of the outcome) given outcome $z \in \mathcal{Z}$. This loss function is standard in the information theory literature, due to its link with coding (Cover and Thomas, 2006). The risk of a density g can be interpreted in relation to the joint probability assigned by g to a large i.i.d. test sample $Z'_1, \dots, Z'_m$ from P: by the law of large numbers, as m tends to infinity, almost surely

$$\prod_{j=1}^{m} g(Z'_j) = \exp\Big(-\sum_{j=1}^{m} \ell(g, Z'_j)\Big) = \exp\Big(-m[R(g) + o(1)]\Big)\,.$$

In addition, assume that P of Z has a density $p \in \mathcal{G}$; we then have, for every $g \in \mathcal{G}$,

$$R(g) - R(p) = \mathbb{E}\Big[\log\Big(\frac{p(Z)}{g(Z)}\Big)\Big] = \int_{\mathcal{Z}} \log\Big(\frac{p}{g}\Big) p \,\mathrm{d}\mu = \mathrm{KL}(p \cdot \mu, g \cdot \mu) \geqslant 0\,,$$

where $\mathrm{KL}(P, Q) := \int_{\mathcal{Z}} \log\big(\frac{\mathrm{d}P}{\mathrm{d}Q}\big)\mathrm{d}P$ denotes the *Kullback-Leibler divergence* (or *relative entropy*) between distributions P and Q. In particular, the risk is minimized by the true density p (if it exists), and prediction under logarithmic loss is equivalent to density estimation under Kullback-Leibler risk.

Our aim is to find *estimators*, which associate to any sample $Z_1, \dots, Z_n$ a density $\widehat{g}_n \in \mathcal{G}$, whose risk is controlled in some general setting. While it is typically impossible to obtain finite-sample guarantees without any assumption on the underlying distribution P (see e.g. Devroye et al., 1996), oftentimes one expects this distribution to possess some structure. In such cases, it is natural to introduce inductive bias in the procedure; one standard way to do so is to select a suitable class of densities $\mathcal{F} \subset \mathcal{G}$ (often called a *statistical model*) that is susceptible to capture at least part of the structure of P, and thus provide a non-trivial approximation thereof.

A classical approach is then to assume that the model $\mathcal{F}$ is *well-specified*, in the sense that it contains the true density p. In this case, the problem of estimating P falls within the classical framework of parametric statistics (Ibragimov and Has'minskii, 1981; van der Vaart, 1998; Lehmann and Casella, 1998). This theory provides strong support for the maximum likelihood estimator (MLE), which arises as an asymptotically optimal estimator for regular models as the sample size n grows (Hájek, 1972; Le Cam, 1986; Ibragimov and Has'minskii, 1981). The same problem can also be treated for a fixed sample size, through the lens of statistical decision theory (Wald, 1949; Lehmann and Casella, 1998), which emphasizes optimal estimators in the average (Bayesian) and minimax senses. Generally speaking, these approaches offer precise descriptions of achievable rates of convergence (up to correct leading constants) and of *efficient*

estimators that make the best use of available data. A major limitation of this approach, however, is that these results rely on the unrealistic assumption that the true distribution belongs to the selected model. Such an assumption is generally unlikely to hold, since the model usually involves a simplified representation of the phenomenon under study: it comes from a choice of the statistician, who has no control over the true underlying mechanism.

A more realistic situation occurs when the underlying model captures some aspects of the true distribution, such as its most salient properties, but not all of them. In other words, the statistical model provides some non-trivial approximation of the true distribution, and is thus "wrong but useful". In such a case, a meaningful objective is to approximate the true distribution (namely, to predict its realizations) almost as well as the best distribution in the model. This task can naturally be cast in the framework of Statistical Learning Theory (Vapnik, 1998), where one constrains the comparison class $\mathcal{F}$ while making few modeling assumptions about the true distribution. Given a class $\mathcal{F}$ of densities, the performance of an estimator $\widehat{g}_n$ is evaluated in terms of its *excess risk* with respect to the class $\mathcal{F}$, namely

$$\mathcal{E}(\widehat{g}_n) := R(\widehat{g}_n) - \inf_{f\in\mathcal{F}} R(f)\,.$$

We say that the estimator $\widehat{g}_n$ is *proper* (or a *plug-in* estimator) when it takes value inside the class $\mathcal{F}$, otherwise $\widehat{g}_n$ will be referred to as an *improper* procedure. Below, we discuss two established approaches to this problem.

Maximum Likelihood Estimation. Arguably the simplest and most standard procedure is the *Maximum Likelihood Estimator* (MLE), or *Empirical Risk Minimizer* (ERM) with logarithmic loss, given by

$$\widehat{f}_n := \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) = \arg\max_{f\in\mathcal{F}} \prod_{i=1}^n f(Z_i)\,. \tag{7.2}$$

Assume now that $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is some parametric model indexed by an open subset $\Theta \subset \mathbf{R}^d$, such that the density $f_\theta(z)$ depends smoothly on θ, and denote $\widehat{f}_n = f_{\widehat{\theta}_n}$ the MLE. First, consider the well-specified case where the true distribution P belongs to the model, say $P = f_{\theta^*}\cdot\mu$, and denote $I(\theta^*) := \mathbb{E}[-\nabla^2 \log f_\theta(Z)]|_{\theta=\theta^*}$ the Fisher information matrix, assumed invertible. Then, under standard regularity and moment conditions (van der Vaart, 1998; Ibragimov and Has'minskii, 1981), we have as $n\to\infty$,

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \overset{(d)}{\to} \mathcal{N}(0, I(\theta^*)^{-1}) \quad \text{while} \quad \mathcal{E}(f_\theta) = \frac{1}{2}\|\theta-\theta^*\|^2_{I(\theta^*)} + o(\|\theta-\theta^*\|^2)\,,$$

where we denote $\|u\|_A := \langle Au, u\rangle^{1/2}$ for any $u\in\mathbf{R}^d$ and symmetric positive matrix A. This implies that $2n\mathcal{E}(f_{\widehat{\theta}_n})$ converges in distribution to a χ^2_d distribution; hence, under suitable domination conditions, the asymptotic excess risk of the MLE satisfies $\mathbb{E}[\mathcal{E}(\widehat{f}_n)] = d/(2n) + o(n^{-1})$. This asymptotic performance turns out to be unimprovable in the well-specified case: for instance, MLE is locally asymptotically minimax optimal (Hájek, 1972; Le Cam and Yang, 2000).

In contrast to its optimality in the well-specified case, the performance of MLE can degrade in the general misspecified case, where it depends on the true distribution P. Indeed, let $\theta^* = \arg\min_{\theta\in\Theta} R(f_{\theta^*})$ be the optimal parameter, and $G = \mathbb{E}[\nabla\ell(f_\theta, Z)\nabla\ell(f_\theta, Z)^\top]|_{\theta=\theta^*}$, $H = \mathbb{E}[\nabla^2\ell(f_\theta, Z)]|_{\theta=\theta^*}$; when P belongs to the model, $G = H = I(\theta^*)$, but in general those

matrices are distinct. In this case, under suitable conditions, it follows from general results on the asymptotic behavior of M-estimators (van der Vaart, 1998; White, 1982) that

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \overset{(d)}{\to} \mathcal{N}(0, H^{-1}GH^{-1}) \quad \text{and} \quad \mathcal{E}(f_\theta) = \frac{1}{2}\|\theta - \theta^*\|_H^2 + o(\|\theta - \theta^*\|^2)\,.$$

Again under suitable domination conditions, this implies that, as $n \to \infty$,

$$\mathbb{E}[\mathcal{E}(\widehat{f}_n)] = \frac{\mathrm{Tr}(H^{-1/2}GH^{-1/2})}{2n} + o\Big(\frac{1}{n}\Big) = \frac{d_{\mathrm{eff}}}{2n} + o\Big(\frac{1}{n}\Big)\,; \tag{7.3}$$

here, the constant $d_{\mathrm{eff}} := \mathrm{Tr}(H^{-1/2}GH^{-1/2})$ depends on the distribution P, and can typically be arbitrarily large, as will be seen below in the case of logistic regression. In fact, degradation under model misspecification is not specific to MLE, and is typically a limitation shared by any *proper* (or *plug-in*) estimator that returns a distribution within the class $\mathcal{F}$, such as penalized MLE. Finally, let us note that, while we adopted an asymptotic viewpoint in this discussion for the sake of clarity, our focus will be on explicit finite sample bounds.

Sequential prediction and online-to-offline conversion. In contrast, distribution-free excess risk bounds have been obtained in the literature (Barron, 1987; Catoni, 2004; Yang, 2000; Juditsky et al., 2008; Audibert, 2009) through a reduction to the comparatively much better understood setting of sequential prediction under logarithmic loss (Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006; Shtarkov, 1987; Grünwald, 2007). In this problem, which is connected to coding (Cover and Thomas, 2006) and the minimum description length (MDL) principle (Rissanen, 1985; Grünwald, 2007), one seeks to control *cumulative* criteria such as the cumulative excess risk, or the regret

$$\sum_{i=1}^n \ell(\widehat{g}_{i-1}, Z_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^n \ell(f, Z_i)$$

over all sequences $Z_1, \dots, Z_n \in \mathcal{Z}$, where $\widehat{g}_{i-1}$ is selected based on $Z_1, \dots, Z_{i-1}$. The control of such cumulative quantities is significantly simplified by the observation that

$$\sum_{i=1}^n \ell(\widehat{g}_{i-1}, Z_i) - \inf_{f\in\mathcal{F}} \sum_{i=1}^n \ell(f, Z_i) = -\log\left(\frac{\prod_{i=1}^n \widehat{g}_{i-1}(Z_i)}{\sup_{f\in\mathcal{F}} \prod_{i=1}^n f(Z_i)}\right),$$

where the ratio inside the logarithm can be interpreted as a ratio of joint densities over $Z_1, \dots, Z_n$. This enables one to determine the minimax regret (Shtarkov, 1987), as well as to control the regret of specific sequential prediction strategies $\widehat{g}_0, \dots, \widehat{g}_{n-1}$. Among those, arguably the most standard are Bayesian mixture strategies (Vovk, 1998; Littlestone and Warmuth, 1994; Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006) with near-optimal guarantees (Clarke and Barron, 1994; Xie and Barron, 2000; Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006), where given a prior distribution π on the parameter space Θ, $\widehat{g}_i$ is the *Bayesian predictive posterior*:

$$\widehat{g}_i(z) = \frac{\int_\Theta f_\theta(Z_1)\cdots f_\theta(Z_i) f_\theta(z)\pi(\mathrm{d}\theta)}{\int_\Theta f_\theta(Z_1)\cdots f_\theta(Z_i)\pi(\mathrm{d}\theta)} = \int_\Theta f_\theta(z)\,\pi(\mathrm{d}\theta|Z_1, \dots, Z_i)\,. \tag{7.4}$$

For smooth, bounded parametric families of dimension d, the minimax cumulative excess risk and regret are known to scale as $(d\log n)/2 + C(\mathcal{F})$ for some constant $C(\mathcal{F})$ depending

on the model, see Clarke and Barron (1994); Merhav and Feder (1998). Note that regret bounds hold for any sequence, and in particular do not require the sequence of observations to be sampled from a distribution in the model. A generic procedure called *online to batch conversion* (Littlestone, 1989; Cesa-Bianchi et al., 2004) enables one to convert any guarantee on the cumulative excess risk into one on the *non-cumulative* excess risk for the average of the successive densities output by the sequential procedure, namely

$$\bar{g}_n = \frac{1}{n+1} \sum_{i=0}^{n} \widehat{g}_i \,. \tag{7.5}$$

When applied to Bayes mixture rules, this yields the so-called *progressive mixture* or *mirror averaging* procedure (Yang and Barron, 1999; Catoni, 1997, 2004; Juditsky et al., 2008; Audibert, 2009), with excess risk bounded by $O((d \log n)/n + C(\mathcal{F})/n)$.

While appropriate for sequential prediction, this approach is not fully satisfactory in the statistical learning setting considered here, for the following reasons. First, the obtained $O(d \log n/n)$ rate features a suboptimal $\log n$ factor, when compared to the $O(d/n)$ rate of MLE in the well-specified case; this highlights the inefficiency of the averaged estimator $\bar{g}_n$, which mixes estimators $\widehat{g}_i$ computed with only a fraction of the sample. Obtaining bounds of $O(d/n)$ for the individual risk was posed as an open problem by Grünwald and Kotłowski (2011). Second, the minimax regret (and in particular the model-dependent constant $C(\mathcal{F})$) is typically *infinite* (Shtarkov, 1987; Clarke and Barron, 1994; Rissanen, 1996; Grünwald, 2007) for unbounded "infinite-volume" classes $\mathcal{F}$ including Gaussian models, so that no uniform guarantee can be obtained over such classes through regret minimization and online-to-offline conversion, reflecting the poor localization of such bounds. These first two limitations are shared by any approach reducing to the sequential problem, which takes into account early rounds where few observations are available. A third limitation lies in the computational requirements of such procedures: in particular, Bayesian mixture approaches involve — absent a conjugate prior allowing exact computations — approximate posterior computations, which are often significantly more expensive than maximum likelihood optimization, inhibiting practical use of such methods.

7.1.1 Our contributions

Let us now summarize our main contributions. Note that, while the previous discussion dealt with density estimation, most of this work in fact deals with *conditional* density estimation, where one seeks to estimate the conditional distribution of a response Y to an input variable X, under logarithmic loss $\ell(f, (X, Y)) = -\log f(Y|X)$ (see Section 7.2.2).

SMP: a general procedure for conditional density estimation. In the present work, we introduce a general procedure for predictive density estimation under entropy risk. This estimator, which we call *Sample Minmax Predictor* (SMP), is obtained by minimizing a new general excess risk bound for supervised statistical learning (Theorem 7.1), and in particular conditional density estimation (Theorem 7.2). In short, SMP is the solution of some minmax problem obtained by considering *virtual samples*. SMP satisfies an excess risk bound valid under model misspecification, and unlike previous approaches does not rely on a reduction to the sequential problem, thereby improving rates for parametric classes from $O(d \log n/n)$ to $O(d/n)$ for our considered models, addressing an open problem from Grünwald and Kotłowski (2011) in this case.

SMP for the Gaussian linear model. We apply SMP to the *Gaussian linear model* $\mathcal{F} = \{f_\theta(\cdot|x) = \mathcal{N}(\langle\theta, x\rangle, \sigma^2) : \theta \in \mathbf{R}^d\}$ for some $\sigma^2 > 0$, a classical conditional model for a scalar response $y \in \mathbf{R}$ to covariates $x \in \mathbf{R}^d$. SMP then smoothes predictions in terms of *leverage scores*, and for every distribution of covariates, its expected excess risk in the general misspecified case is at most twice the minimax excess risk in the *well-specified* case, but without any condition on the approximation error of the linear model or noise variance (Theorem 7.4). This yields an excess risk bound of $d/n+O((d/n)^2)$ over the class $\mathcal{F}$ under some regularity assumptions on covariates (Corollary 7.1); such a guarantee cannot be obtained for a within-model estimator, or through a regret minimization approach.

We also consider a Ridge-regularized variant of SMP, and study its performance on balls of the form $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ for $B > 0$. For covariates X bounded by $R > 0$, we establish two guarantees: a "finite-dimensional" bound of $O(d \log(BR/\sqrt{d})/n)$ (Proposition 7.3), removing an extra $\log n$ term from results of Kakade and Ng (2005) in the sequential case, and a dimension-free "nonparametric" bound (Theorem 7.5), where explicit dependence on d is replaced by a dependence on the covariance structure of covariates, matching well-specified minimax rates over such balls in infinite dimension (Caponnetto and De Vito, 2007).

SMP for logistic regression. We then turn to logistic regression, arguably the most standard model for a binary response $y \in \{-1, 1\}$ to covariates $x \in \mathbf{R}^d$, given by $\mathcal{F} = \{f_\theta(1|x) = \sigma(\langle\theta, x\rangle) : \theta \in \mathbf{R}^d\}$, where $\sigma(u) = e^u/(1+e^u)$. In this case, SMP admits a simple form, and its prediction can be computed by solving two logistic regressions. Assuming that $\|X\| \leqslant R$, we show that a Ridge-penalized variant of SMP achieves excess risk $O((d+B^2R^2)/n)$ with respect to the ball $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ for all $B > 0$ (Corollary 7.2), together with dimension-free bounds (Theorem 7.6). In contrast, results of Hazan et al. (2014) show that no within-model estimator can achieve better rate than $\min(BR/\sqrt{n}, de^{BR}/n)$ without further assumptions. Compared to approaches obtaining fast rates through Bayesian mixtures (Kakade and Ng, 2005; Foster et al., 2018), computation of SMP replaces posterior sampling by optimization. SMP thus provides a natural non-Bayesian approach to uncertainty quantification and calibration of probabilistic estimates, relying on virtual samples.

7.1.2 Related work

Well-specified density estimation. There is a rich statistical literature on predictive density estimation under entropy risk in the well-specified case (where the true distribution is assumed to belong to the model), see Harris (1989); Komaki (1996); Hartigan (1998); Aslan (2006); Liang and Barron (2004); George et al. (2006); Sweeting et al. (2006); Brown et al. (2008) and references therein. First, as mentioned above, MLE is known to be asymptotically normal and efficient (van der Vaart, 1998; Ibragimov and Has'minskii, 1981; Le Cam and Yang, 2000) in this case; its asymptotic optimality can be formalized precisely by Hájek's local asymptotic minimax theorem (Hájek, 1972; Le Cam and Yang, 2000). Beyond this optimality result, a number of refinements have been explored: improvement of Bayes predictive distributions over the MLE for finite samples (Aitchison, 1975), higher-order risk asymptotics (Hartigan, 1998; Ghosh, 1994; Aslan, 2006) and second-order minimax procedures (Aslan, 2006), exact minimax procedures for location and scale families (Liang and Barron, 2004), as well as admissibility and shrinkage for the Gaussian model (Brown et al., 2008). While related to this line of work, our approach differs from it by relaxing the (restrictive) assumption that

the distribution of interest belongs to the specified model; another difference with some of the aforementioned references is our non-asymptotic focus.

Non-asymptotic analyses of estimators under misspecification. The asymptotic behavior of MLE (including consistency and asymptotic normality) in the misspecified case is also well-understood (White, 1982; van der Vaart, 1998). Beyond the asymptotic setting, non-asymptotic analyses of MLE and related procedures have been carried by van de Geer (1999); Birgé and Massart (1993, 1998); Yang and Barron (1998); Wong and Shen (1995); Spokoiny (2012), by using techniques from empirical process theory (van der Vaart and Wellner, 1996; Talagrand, 2014; Massart, 2007; Boucheron et al., 2013). In addition to these classical references, we mention two approaches that circumvent in different ways reliance on the machinery of empirical process theory. First, Zhang (2006a) relies on information-theoretic inequalities to analyze Bayesian and penalized estimators; this approach is considerably expanded by Grünwald and Mehta (2019), who obtain bounds in terms of refined complexity measures. Our guarantees have notable commonalities with those of Grünwald and Mehta (2019), in that excess risk is controlled in terms of some min-max quantity for logarithmic loss, yet they are of a different nature. Indeed, the bounds from Grünwald and Mehta (2019) apply to many estimators such as MLE (while ours are tailored to SMP); the price to pay is that such guarantees depend on the true distribution and can degrade under model misspecification, reflecting the behavior of the estimators they apply to. Another difference is that, while the guarantees of Grünwald and Mehta (2019) do not rely on online-to-offline conversion and iterate averaging, the risk is controlled in terms of the same quantity that appears in the sequential case, with the same shortcomings for parametric or unbounded models (this reflects the focus of this paper on bounded nonparametric models). Second, Ostrovskii and Bach (2018) developed an analysis relying on self-concordance, which applies in particular to logistic regression. Overall, this literature differs from ours in that it studies estimators such as (penalized) MLE, which inevitably degrade for some misspecified distributions.

Sequential prediction. As mentioned previously, the sequential variant of prediction under logarithmic loss is well-studied (Shtarkov, 1987; Clarke and Barron, 1994; Merhav and Feder, 1998; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006; Grünwald, 2007). These guarantees on cumulative criteria have been transported to the individual excess risk considered here (Barron, 1987; Catoni, 2004; Yang and Barron, 1999; Juditsky et al., 2008; Audibert, 2009). To the best of our knowledge, prior to the present work, this online-to-offline conversion was the only approach to obtaining distribution-free excess risk guarantees. As mentioned above, reduction to the sequential case is suboptimal, in that it leads to extra logarithmic factors in the rate and cannot provide uniform guarantees over unbounded models. Our general guarantee for SMP provides a more "localized" risk bound adapted to such situations.

Stability. Our general bound on the excess risk is related to the approach in terms of stability of the loss of the predictor under sample changes (Bousquet and Elisseeff, 2002; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010; Koren and Levy, 2015), in particular in its use of exchangeability. While close in spirit, our bounds involve a different quantity; the difference between the two is particularly apparent in the context of logistic regression, where it enables us to remove some exponential constants.

Logistic regression. An important motivation for this work was recent progress and questions on logistic regression, arguably the most common model for conditional density estimation with binary response (Berkson, 1944; McCullagh and Nelder, 1989; van der Vaart, 1998). Under boundedness assumptions, it can be seen as a special convex and Lipschitz stochastic optimization problem, for which slow rates of convergence are available (Zinkevich, 2003; Nemirovski et al., 2009; Bubeck, 2015). In addition, logistic regression is also an exp-concave problem, which enables fast rates (Hazan et al., 2007; Koren and Levy, 2015; Mehta, 2017), but with an exponential dependence on the domain radius. It is shown by Hazan et al. (2014) that such rates are unimprovable without further assumptions. To obtain improved results, one thread of work proceeds under additional assumptions, and performs a refined analysis using (generalized) *self-concordance* of the logistic loss (Bach, 2010, 2014; Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019); this leads to distribution-dependent guarantees which improve for favorable distributions, but exhibit exponential dependence in the worst case. Another approach consists in using out-of-model procedures, for which the lower bound of Hazan et al. (2014) does not apply. By using Bayes mixtures strategies and reducing to the sequential problem, Kakade and Ng (2005); Foster et al. (2018) establish fast risk rates without exponential dependence on the norm, bypassing the previous lower bound; the question of finding a practical procedure enjoying such guarantees without expensive posterior sampling is left open by Foster et al. (2018). Our work is cast in the same setting under weak distributional assumptions, and provides a practical approach with fast rates guarantees in this case. We also note that our analysis of SMP does rely on self-concordance, albeit applied to a different estimator.

7.1.3 Outline and notations

This chapter is organized as follows. In Section 7.2, we introduce the setting and state a general excess risk bound for supervised learning (Theorem 7.1) and its instantiation to conditional density estimation (Theorem 7.2), minimized by SMP, which will be used throughout. Section 7.3 provides direct consequences of the previous bounds in the context of (unconditional) density estimation with multinomial and Gaussian models. In Section 7.4, we study SMP and its guarantees for conditional density estimation with the Gaussian linear model. We finally turn to logistic regression in Section 7.5. The proofs are gathered in Section 7.7, while Section 7.6 concludes.

Notations. Throughout this chapter, we denote $\langle x, y\rangle := x^\top y$ the canonical scalar product of $x, y \in \mathbf{R}^d$, and $\|x\| := \langle x, x\rangle^{1/2}$ the associated Euclidean norm. Likewise, for any symmetric positive semi-definite $d \times d$ matrix Σ, we let $\langle x, y\rangle := \langle \Sigma x, y\rangle$ and $\|x\|_\Sigma = \langle x, x\rangle_\Sigma^{1/2}$.

7.2 General excess risk bounds

7.2.1 A general excess risk bound for statistical learning

In this section, we let $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$ be three measurable spaces, corresponding respectively to the feature, label and prediction spaces, and let $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$ be a loss function. Denote by $\widehat{\mathcal{F}}$ the space of all measurable functions $\mathcal{X} \to \widehat{\mathcal{Y}}$ (also called predictors), and let $\mathcal{F} \subset \widehat{\mathcal{F}}$ be a class of predictors. We also consider a penalization function $\phi : \mathcal{F} \to \mathbf{R}$. Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

and let

$$\ell_\phi(f, z) = \ell(f(x), y) + \phi(f)$$

for any $z = (x, y) \in \mathcal{Z}$ and $f \in \mathcal{F}$. When no penalization is used ($\phi \equiv 0$) we simply write $\ell = \ell_0$. Let P be some probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The quality of a predictor $g \in \widehat{\mathcal{F}}$ is measured through its *risk*

$$R(g) = \mathbb{E}[\ell(g, Z)] = \mathbb{E}[\ell(g(X), Y)] \tag{7.6}$$

where $Z = (X, Y) \sim P$, whenever this expectation is well-defined and belongs to $\mathbf{R} \cup \{+\infty\}$, which we assume from now on. Also, define the *excess risk* (with respect to $\mathcal{F}$) of g as

$$\mathcal{E}(g) = R(g) - \inf_{f \in \mathcal{F}} R(f)\,. \tag{7.7}$$

We define similarly $R_\phi(f) = \mathbb{E}[\ell_\phi(f, Z)]$ for $f \in \mathcal{F}$ and $\mathcal{E}_\phi(g) = R(g) - \inf_{f \in \mathcal{F}} R_\phi(f)$.

In this setting, the distribution P is unknown, and we will avoid making strong assumptions on it. The aim is to produce, given an i.i.d. sample $Z_1^n = (Z_1, \ldots, Z_n)$ from P, a predictor $\widehat{g}_n : \mathcal{X} \to \widehat{\mathcal{Y}}$ whose expected excess risk $\mathbb{E}[\mathcal{E}(\widehat{g}_n)]$ (where the expectation holds over the random sample) is small. In other words, $\widehat{g}_n$ should predict almost as well as the best element in $\mathcal{F}$, up to a controlled small additional term. Given a sample $Z_1^n = (Z_1, \ldots, Z_n)$, we denote

$$\widehat{f}_{\phi,n} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell_\phi(f, Z_i) \tag{7.8}$$

a (penalized) *empirical risk minimizer* (ERM); when $\phi \equiv 0$, we simply denote the ERM as $\widehat{f}_n$. Throughout this chapter, we assume to simplify that this minimum is attained. This holds in virtually all the examples considered below; in addition, the arguments naturally extend to approximate minimizers. By convention, all minimizers of the empirical risk will be chosen symmetrically in the sample points $Z_1, \ldots, Z_n$. We also introduce

$$\widehat{f}^{\,z}_{\phi,n} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \ell_\phi(f, Z_i) + \ell_\phi(f, z) \right\} \tag{7.9}$$

for any $z \in \mathcal{Z}$. Theorem 7.1 below introduces a new bound on the excess risk of any prediction rule, together with a predictor that minimizes it. It holds for a general loss ℓ, but in the following sections we apply it to the logarithmic loss only, for which the predictor can be made explicit.

Theorem 7.1 (Main excess risk bound and Sample Minmax Predictor)**.** *For any predictor $\widehat{g}_n$ depending on Z_1^n, we have*

$$\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{g}_n(X), y) - \ell_\phi(\widehat{f}^{\,(X,y)}_{\phi,n}(X), y) \Big\} \Big] \tag{7.10}$$

where $\widehat{f}^{\,z}_{\phi,n}$ is defined by (7.9) for $z \in \mathcal{Z}$ and $Z = (X, Y) \sim P$ is independent of Z_1^n. In addition, the right-hand side of (7.10) is minimized by the predictor

$$\widetilde{f}_{\phi,n}(x) = \operatorname*{arg\,min}_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell_\phi(\widehat{f}^{\,(x,y)}_{\phi,n}(x), y) \Big\}\,, \tag{7.11}$$

which we call SMP (Sample Minmax Predictor) whenever it exists, in which case (7.10) becomes

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\inf_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell_\phi(\widehat{f}^{\,(X,y)}_{\phi,n}(X), y) \Big\} \Big]\,. \tag{7.12}$$

The proof of Theorem 7.1 is given in Section 7.7.1. The excess risk bound of Theorem 7.1 is related to the stability of the (regularized) empirical risk minimizer. Indeed, if the ERM $\widehat{f}_{\phi,n}^{(X,y)}$ obtained by adding a new sample (X, y) does not depend too much on the label y, *i.e.* if the set $\{\widehat{f}_{\phi,n}^{(X,y)} : y \in \mathcal{Y}\}$ is small in expectation, then the min-max quantity in the bound (7.12) will also be small.

The use of stability to establish guarantees for learning algorithms such as ERM or approximate ERM was pioneered by Bousquet and Elisseeff (2002). Stability arguments were used by Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010) to prove fast rates of order $O(1/n)$ for ERM in strongly convex stochastic optimization problems and more recently by Koren and Levy (2015) for exp-concave problems. However, while related in spirit to the notion of stability, the excess risk bound of Theorem 7.1 differs from standard stability bounds. Indeed, approaches based on stability control the risk in terms of variations of the loss of the output hypothesis (such as ERM) under changes of the sample (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Srebro et al., 2010; Koren and Levy, 2015). By contrast, Theorem 7.1 controls the risk in terms of some min-max quantity, which measures the size of the set of empirical risk minimizers obtained by adding one sample. The difference between the two is most apparent in the context of logistic regression (see Section 7.5 below), where it is critical to obtain improved guarantees that could not be derived from loss stability of regularized risk minimizers.

It is worth noting that the SMP (7.11) whose risk is controlled in (7.12) is *not* the regularized ERM, that is, the algorithm whose "stability" is controlled. In fact, $\widetilde{f}_{\phi,n}$ is in general an *improper* predictor, which does not belong to the class $\mathcal{F}$; it may be seen as a "center" of the set of risk minimizers obtained by adding one sample, in a sense related to the loss function. In fact, we will show in what follows that SMP enjoys guarantees which are not achievable by proper predictors such as regularized ERM.

7.2.2 Conditional density estimation with the logarithmic loss

We now turn to conditional density estimation, which is the focus of this work, by considering the logarithmic loss. Let μ be a measure on $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ be the set of probability densities on $\mathcal{Y}$ with respect to μ, namely the set of measurable functions $f : \mathcal{Y} \to \mathbf{R}^+$ such that $\int_{\mathcal{Y}} f \mathrm{d}\mu = 1$. The logarithmic loss is defined as $\ell(f, y) = -\log f(y)$ for $f \in \widehat{\mathcal{Y}}$ and $y \in \mathcal{Y}$. In this setting, a predictor $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ corresponds to a conditional density. We denote $f(y|x) = f(x)(y)$ and as before $\ell(f, z) = \ell(f(x), y)$ for $z = (x, y)$. Note that, in this case, the ERM (7.8) corresponds to the (conditional) maximum likelihood estimator (MLE). The risk of any conditional density f is

$$R(f) = -\mathbb{E}\big[\log f(Y|X)\big]$$

whenever this expectation is defined. Note that

$$R(g) - R(f) = \mathbb{E}\left[\log \frac{f(Y|X)}{g(Y|X)}\right] \tag{7.13}$$

for any conditional densities f, g with respect to μ, which only depends on the conditional distributions $f\mu, g\mu$, and not on the measure μ which dominates them. In particular, we may choose μ such that the risk $R(f)$ is well-defined and finite for some $f \in \mathcal{F}$, and identify f and g with the corresponding conditional distributions. There exists a best predictor $f^* \in \mathcal{F}$ whenever the excess risk $\mathcal{E}(f) = \mathbb{E}[\ell(f, Z) - \ell(f^*, Z)]$ is defined and belongs to $[0, +\infty]$ for

every $f \in \mathcal{F}$. Following what we did in Section 7.2.1, given a penalization function $\phi : \mathcal{F} \to \mathbf{R}$, we define the penalized risk R_ϕ and the penalized excess risk $\mathcal{E}_\phi$.

Theorem 7.2 below shows that both SMP defined in Theorem 7.1 and its excess risk bound (7.12) can be described explicitly in this case.

Theorem 7.2 (Excess risk bound for conditional density estimation)**.** *In the case of the logarithmic loss, the SMP $\widetilde{f}_{\phi,n}$ defined in (7.11) writes*

$$\widetilde{f}_{\phi,n}(y|x) = \frac{\widehat{f}_{\phi,n}^{(x,y)}(y|x)e^{-\phi(\widehat{f}_{\phi,n}^{(x,y)})}}{\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(x,y')}(y'|x)e^{-\phi(\widehat{f}_{\phi,n}^{(x,y')})}\mu(\mathrm{d}y')}, \tag{7.14}$$

whenever the integral $\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X)e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})}\mu(\mathrm{d}y)$ is finite almost surely (over Z_1^n, X). In addition, its excess risk bound (7.12) writes

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] \leqslant \mathbb{E}_{Z_1^n,X}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X)e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})}\mu(\mathrm{d}y)\Big)\Big]. \tag{7.15}$$

Remark 7.1. In the non-regularized case where $\phi \equiv 0$, SMP simply writes

$$\widetilde{f}_n(y|x) = \frac{\widehat{f}_n^{(x,y)}(y|x)}{\int_{\mathcal{Y}} \widehat{f}_n^{(x,y')}(y'|x)\mu(\mathrm{d}y')},$$

while its excess risk bound (7.15) takes the form:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}_{Z_1^n,X}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_n^{(X,y)}(y|X)\mu(\mathrm{d}y)\Big)\Big].$$

Theorem 7.2 is proved in Section 7.7.1. The SMP (7.14) minimizes, for every value of x, the worst-case (over $y \in \mathcal{Y}$) excess loss $\ell(\widetilde{f}_{\phi,n}(x), y) - \ell_\phi(\widehat{f}_{\phi,n}^{(x,y)}(x), y)$ with respect to the ERM on the sample $Z_1^n, (X, y)$. As explained above, the right-hand side of (7.15) corresponds to (the expectation of) a measure of complexity of the class $\{\widehat{f}_{\phi,n}^{(X,y)}, y \in \mathcal{Y}\}$ associated to the log-loss. We will see below, in particular cases for $\mathcal{F}$, that despite being derived from a general bound for statistical learning, the excess risk bound of the SMP is remarkably tight and close to the optimal risk in the well-specified case. In fact, we will see in the case of the Gaussian linear model (Section 7.4.2) that the bound of the SMP is intrinsic to the hardness of the problem.

In the unconditional case, the prediction of the estimator (7.14) closely resembles that of a sequential prediction strategy called Sequential Normalized Maximum Likelihood (SNML), introduced by Roos and Rissanen (2008) and related to the Last Step Minimax algorithm (which restricts to proper predictions) from Takimoto and Warmuth (2000)[1]. Interestingly, the motivation is completely different: the SNML algorithm was introduced as a computationally efficient relaxation of the minimax algorithm (in terms of cumulative regret) for sequential prediction under log-loss; its worst-case regret was shown to be almost minimax (Kotłowski and Grünwald, 2011), and in fact minimax for some specific families (Bartlett et al., 2013). By contrast, in our case the SMP estimator naturally arises as the minimizer of a novel upper bound on the *non-cumulative* excess risk.

[1]Specifically, the prediction of SMP coincides with that of the SNML-1 algorithm from Roos and Rissanen (2008) at step $n+1$, while SNML-2 from Roos and Rissanen (2008) (simply called SNML in subsequent work Kotłowski and Grünwald (2011); Bartlett et al. (2013)) is slightly different: it minimizes worst-case regret with respect to next step ERM on the whole sequence, instead of just the last sample.

7.3 Some consequences for density estimation

In this section, we consider the problem of (unconditional) density estimation: the space $\mathcal{X}$ is assumed to be trivial (with a single element) and is thus omitted[2], and no penalization is used ($\phi \equiv 0$). In other words, given access to an i.i.d. sample $(Y_1, \dots, Y_n)$ from a distribution P on $\mathcal{Y}$, and given a family $\mathcal{F}$ of probability densities on $\mathcal{Y}$ with respect to μ (namely, a statistical model $\mathcal{F}$), the aim is to find a predictive distribution $\widehat{g}_n$ on $\mathcal{F}$ whose excess risk with respect to $\mathcal{F}$ is as small as possible. Note that the model may be *misspecified*, in the sense that $P \notin \mathcal{F}$. Introduce the Kullback-Leibler (KL) divergence

$$\mathrm{KL}(P, Q) = \mathbb{E}_{Z\sim P}\Big[\log \frac{\mathrm{d}P}{\mathrm{d}Q}(Z)\Big]$$

between distributions P and Q (which is infinite whenever P is not absolutely continuous with respect to Q). If $\mathrm{KL}(P, f^*) < +\infty$ then $f^* = \arg\min_{f\in\mathcal{F}} \mathrm{KL}(P, f)$ and the excess risk (7.7) writes $\mathcal{E}(f) = \mathrm{KL}(P, f) - \mathrm{KL}(P, f^*)$ for any $f \in \mathcal{F}$. For this reason, the risk R is also called *KL risk*.

In the next sections, we apply Theorem 7.2 to misspecified density estimation on standard families. In each case, the SMP is explicit and the excess risk bound scales as d/n irrespective of the true distribution P. These bounds are tight, since they are *within a factor of* 2 of the optimal asymptotic rate in the well-specified case. Also, we compare it with MLE and online to batch conversion of sequential prediction strategies. In all considered examples, SMP improves these estimators.

7.3.1 Finite alphabet: the multinomial model

In this section, we assume that $\mathcal{Y}$ is a finite set with d elements, μ is the counting measure and $\mathcal{F} = \{(p(y))_{y\in\mathcal{Y}} \in \mathbf{R}_+^{\mathcal{Y}} : \sum_{y\in\mathcal{Y}} p(y) = 1\}$ is the multinomial model (which is always well-specified). For any $y \in \mathcal{Y}$, we let $N_n(y) = \sum_{i=1}^n \mathbf{1}(Y_i = y)$.

Proposition 7.1. *If $\mathcal{Y}$ is a finite set with d elements, then SMP corresponds to the* Laplace estimator

$$\widetilde{f}_n(y) = \frac{N_n(y) + 1}{n + d}\,. \tag{7.16}$$

In addition, the bound (7.15) writes in this case

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \log\Big(\frac{n+d}{n+1}\Big) \leqslant \frac{d-1}{n}\,. \tag{7.17}$$

Proposition 7.1 is proved in Section 7.7.2. In this case, the SMP corresponds to the Laplace estimator, which is the Bayes predictive distribution under an uniform prior on $\mathcal{F}$. The first bound in (7.17) is tight: it is an equality when Y is constant almost surely.

On MLE. The MLE is given by $\widehat{f}_n(y) = N_n(y)/n$. Its expected risk is infinite unless P is concentrated on a single point. Indeed, let $y_0, y_1 \in \mathcal{Y}$ be distinct elements such that $\mathbb{P}(Y = y_0), \mathbb{P}(Y = y_1) > 0$; with positive probability, $Y_1 = \cdots = Y_n = y_0$, so that $\widehat{f}_n(y) = \mathbf{1}(y = y_0)$,

[2]While conditional density estimation can be cast as a special case of density estimation, we adopt the opposite perspective since SMP exploits the conditional structure.

$\ell(\widehat{f}_n, y_1) = +\infty$ and thus $R(\widehat{f}_n) = +\infty$. Hence, $\mathbb{E}[R(\widehat{f}_n)] = +\infty$. In order to obtain non-vacuous expected risk for MLE in this case, one may restrict to $\mathcal{F}_\delta = \{p \in \mathcal{F} : \forall y \in \mathcal{Y},\ p(y) \geqslant \delta\}$ for some $\delta \in (0,1)$, so that log ratios of densities are bounded. In this case, whenever $p \in \mathcal{F}_\delta$, the excess risk of MLE has asymptotically efficient rate $(d-1)/(2n) + o(n^{-1})$. This reflects the fact that the model is well-specified.

On online to batch conversion. The minimax cumulative regret with respect to the class $\mathcal{F}$ scales as $(d-1)(\log n)/2 + O(1)$ (Cesa-Bianchi and Lugosi, 2006). Hence, any upper bound based on online-to-batch conversion (Cesa-Bianchi et al., 2004) can be no better than $(d-1)(\log n)/(2n) + O(1/n)$.

7.3.2 The Gaussian location model

We now let $\mathcal{Y} = \mathbf{R}^d$ and consider the Gaussian location model, namely the family $\mathcal{F} = \{\mathcal{N}(\theta, \Sigma) : \theta \in \mathbf{R}^d\}$ of Gaussian distributions with fixed positive covariance matrix Σ. We let $\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i$.

Proposition 7.2. *A risk minimizer $f^* = \mathcal{N}(\theta^*, \Sigma) \in \mathcal{F}$ exists if and only if $\mathbb{E}\|Y\| < +\infty$, in which case $\theta^* = \mathbb{E}[Y]$. For $n \geqslant 1$, the SMP is given by $\widetilde{f}_n = \mathcal{N}(\bar{Y}_n, (1+1/n)^2\Sigma)$, and whenever $\mathbb{E}\|Y\| < +\infty$ the bound (7.15) writes*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant d \log\left(1 + \frac{1}{n}\right) \leqslant \frac{d}{n}\,. \tag{7.18}$$

In addition, when the model is well-specified, we have

$$\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] = d\log\left(1+\frac{1}{n}\right) - \frac{d}{2n} < \frac{d}{2n}\,.$$

The proof of Proposition 7.2 is given in Section 7.7.2 below. It provides an excess risk bound valid under misspecification, under the minimal hypothesis necessary to define excess risk. In addition, this bound does not depend on the distribution of Y, and is essentially a factor of 2 above the optimal asymptotic risk $d/(2n)$ even for a worst-case distribution. In particular, this implies that finding a predictive distribution with small excess risk is feasible even when identifying the best parameter in the family is not: indeed, estimating the parameter θ^* with an accuracy independent of the true distribution of Y is not possible.

On MLE and proper estimators. Assume that $\mathbb{E}\|Y\|^2 < +\infty$ and define $\Sigma_Y = \mathbb{E}[(Y - \mathbb{E}Y)(Y-\mathbb{E}Y)^\top]$. The excess risk of the MLE $\widehat{f}_n = \mathcal{N}(\bar{Y}_n, \Sigma)$ is given by

$$\mathcal{E}(\widehat{f}_n) = \frac{1}{2}\mathbb{E}\big\|\bar{Y}_n - \mathbb{E}[Y]\big\|_{\Sigma^{-1}}^2 = \frac{1}{2n}\mathrm{Tr}(\Sigma^{-1}\Sigma_Y).$$

In the misspecified case where $\Sigma_Y \neq \Sigma$, this quantity depends on the true distribution of Y and can be arbitrarily large depending on Σ_Y. This limitation is in fact shared by any proper estimator of the form $f_{\widehat{\theta}_n} = \mathcal{N}(\widehat{\theta}_n, \Sigma)$ for some $\widehat{\theta}_n$, as explained next. Consider the family of distributions $\{P_{\theta^*} = \mathcal{N}(\theta^*, \Sigma_Y) : \theta^* \in \mathbf{R}^d\}$ for some arbitrary symmetric positive matrix Σ_Y, and the loss function $L(\theta^*, \theta) = \|\theta - \theta^*\|_{\Sigma^{-1}}^2/2$. It is a standard result in decision theory (see e.g. Lehmann and Casella, 1998) that the empirical mean $\bar{Y}_n$ is minimax optimal for this problem and has constant risk $\mathrm{Tr}(\Sigma^{-1}\Sigma_Y)/(2n)$. Therefore, for any proper estimator $f_{\widehat{\theta}_n}$,

$$\sup_{\theta^* \in \mathbf{R}^d} \mathbb{E}_{Y\sim P_{\theta^*}}\big[\mathcal{E}(f_{\widehat{\theta}_n})\big] = \frac{1}{2}\sup_{\theta^*\in\mathbf{R}^d} \mathbb{E}_{\theta^*}\big\|\widehat{\theta}_n - \mathbb{E}[Y]\big\|_{\Sigma^{-1}}^2 \geqslant \frac{\mathrm{Tr}(\Sigma^{-1}\Sigma_Y)}{2n}\,.$$

On online to batch conversion. The minimax cumulative regret with respect to the full Gaussian family $\mathcal{F}$ is infinite (see, e.g., Grünwald, 2007): this comes from the fact that regret after the first step (the first prediction being made before seeing any sample) is unbounded. This difficulty does not appear in the batch setting, where one can predict conditionally on the sample, in a translation-invariant fashion. One can guarantee finite minimax regret by considering a restricted model $\{\mathcal{N}(\theta, \Sigma) : \theta \in K\}$ for some compact set $K \subset \mathbf{R}^d$ (Grünwald, 2007), in which case minimax regret scales as $d(\log n)/2 + C_K + o(1)$ (for some constant C_K depending on K) so that online to batch conversion yields an excess risk bound of $d(\log n)/(2n) + C_K/n + o(1/n)$, which again exhibits an extra $\log n$ factor.

Exact minimax rate in the misspecified case. In fact, for the Gaussian location family, the minimax excess risk in the general misspecified case, namely

$$\inf_{\widehat{g}_n} \sup_{P} \mathbb{E}_{Y\sim P}\big[\mathcal{E}(\widehat{g}_n)\big] \tag{7.19}$$

where the supremum spans over all probability distributions P on $\mathbf{R}^d$ such that $\mathbb{E}\|Y\|^2 < +\infty$, the infimum over density estimators $\widehat{g}_n$ and where the excess risk is under the true distribution P, can be determined exactly, together with a minimax estimator, as shown below.

Theorem 7.3. *For the Gaussian location model, the minimax excess risk* (7.19) *in the misspecified case* (*namely, over all distributions with finite second moment*) *is equal to*

$$\inf_{\widehat{g}_n} \sup_{P} \mathbb{E}_{Y\sim P}\big[\mathcal{E}(\widehat{g}_n)\big] = \frac{d}{2}\log\left(1+\frac{1}{n}\right).$$

In addition, this minimax excess risk is achieved by the estimator $\widehat{g}_n = \mathcal{N}(\bar{Y}_n, (1+1/n)\Sigma)$, *which satisfies* $\mathbb{E}[\mathcal{E}(\widehat{g}_n)] = (d/2)\log(1+1/n)$ *for any distribution* P *of* Y *such that* $\mathbb{E}[\|Y\|^2] < +\infty$.

Theorem 7.3 is proven in Section 7.7.2 below. Note that $\widehat{g}_n$ corresponds to the Bayes predictive posterior under uniform prior, which is known to achieve the minimax risk in the *well-specified* case (Ng, 1980; Murray, 1977), see also George et al. (2006). Remarkably, both the minimax excess risk and the minimax estimator remain the same in the misspecified case. This holds even though the posterior itself (a distribution on $\mathcal{F}$) does not concentrate on a neighborhood of the best parameter $\theta^* = \mathbb{E}[Y]$ in the misspecified case (contrary to the well-specified case), when the true variance is large. An explanation for this phenomenon is that the out-of-model correction of the Bayes predictive posterior (critically due to averaging over the posterior) brings it closer to distributions with high variance, thereby compensating the high variability for such distributions. As a result, the Bayes predictive posterior equalizes the excess risk across all distributions. This suggests that posterior concentration rates alone, which do not take into account the latter effect (and degrade under model misspecification when the true variance is large), fail to accurately characterize the excess risk of predictive posteriors under model misspecification.

Finally, Theorem 7.3 shows that the worst-case excess risk bound (7.18) of SMP is exactly twice the minimax excess risk for distributions with finite variance.

7.4 Gaussian linear conditional density estimation

In this section, we turn to conditional density estimation, starting with arguably the most standard family, namely the linear Gaussian model. After introducing the setting, notations and basic assumptions (Section 7.4.1), we consider the non-penalized SMP and its excess risk bounds with respect to the full unrestricted model (Section 7.4.2). Next, we consider in Section 7.4.3 the Ridge-regularized SMP and its performance, both in the finite-dimensional context and in the nonparametric one where d may be larger than n. In the latter case, the bounds only depend on the covariance structure of X and on the norm of the comparison parameter.

7.4.1 Setting: the Gaussian linear model

Consider the spaces $\mathcal{X} = \mathbf{R}^d$ and $\mathcal{Y} = \mathbf{R}$ and the family of conditional distributions

$$\mathcal{F} = \left\{ f_\theta(\cdot|x) = \mathcal{N}(\langle \theta, x \rangle, \sigma^2) : \theta \in \mathbf{R}^d \right\} \tag{7.20}$$

for some $\sigma^2 > 0$; up to the change of variables $y' = y/\sigma$, we will assume without loss of generality that $\sigma^2 = 1$. Throughout this section, we consider log-loss with respect to the base measure $\mu = (2\pi)^{-1/2} \mathrm{d}y$ on $\mathbf{R}$, so that for $\theta \in \mathbf{R}^d$ and $(x, y) \in \mathbf{R}^d \times \mathbf{R}$:

$$\ell(f_\theta, (x, y)) = -\log f_\theta(y|x) = \frac{1}{2}(y - \langle \theta, x \rangle)^2 \,, \tag{7.21}$$

and hence the risk of f_θ writes

$$R(f_\theta) = \frac{1}{2}\mathbb{E}\left[(Y - \langle \theta, X \rangle)^2\right] .$$

The problem of conditional density estimation in the Gaussian linear model is intimately linked (but not equivalent) to that of linear least-squares regression, namely statistical learning with the square loss and a comparison class formed by linear predictors. Let us discuss the connection and differences between the two problems:

- In the least-squares problem, one is interested in a *point prediction* of the response y given the covariates x, or equivalently in an estimate of the *conditional expectation* $\mathbb{E}[Y|X]$ of Y given X. By contrast, in the setting of density estimation one seeks a *probabilistic prediction* of y given x, or equivalently an estimate of the *conditional distribution* of Y given X, which includes a quantification of the uncertainty of Y given X.
- When one restricts to proper, within-model estimators (taking values in $\mathcal{F}$), the two problems are equivalent, as shown by the expression of the loss (7.21).
- On the other hand, in the context of conditional density estimation, the possibility of using improper (out-of-model) estimators provides more flexibility. As we will see, this additional flexibility is essential to bypass lower bounds for proper estimators in the misspecified case.

Let us emphasize that in the context of conditional density estimation, well-specification refers to the fact that the conditional distribution of Y given X belongs to the model. As in the unconditional case, we are interested in bounds that do not degrade under model misspecification, and hence require only weak assumptions on this conditional distribution. Assumption 7.1 below will be made throughout this section, while further assumptions will be made in Sections 7.4.2 and 7.4.3 respectively.

Assumption 7.1 (Finite second moments)**.** We assume that both X and Y are square integrable, namely

$$\mathbb{E}\|X\|^2 < +\infty \quad \text{and} \quad \sigma_Y^2 := \mathbb{E}[Y^2] < +\infty.$$

We will denote $\Sigma = \Sigma_X = \mathbb{E}[XX^\top]$ the second-order moment matrix, which we will call (following a common abuse of terminology) the *covariance matrix* of X, even when X is not centered. Assumption 7.1 implies that YX is integrable (by the Cauchy-Schwarz inequality) and that $\mathbb{E}[\langle \theta, X\rangle^2] = \langle \Sigma\theta, \theta\rangle$, so that the risk $R(f_\theta)$ is finite[3] and equals:

$$R(f_\theta) = \frac{1}{2}\langle \Sigma\theta, \theta\rangle - \langle \theta, \mathbb{E}[YX]\rangle + \frac{1}{2}\mathbb{E}[Y^2]\,,$$

with gradient $\nabla R(f_\theta) = \Sigma\theta - \mathbb{E}[YX]$. In particular, whenever Σ is invertible, the population risk minimizer $f^* \in \mathcal{F}$ is given by $f^* = f_{\theta^*}$ with $\theta^* = \Sigma^{-1}\mathbb{E}[YX]$, while the excess risk of $f_\theta \in \mathcal{F}$ writes $\mathcal{E}(f_\theta) = \frac{1}{2}\|\theta - \theta^*\|_\Sigma^2$. Likewise, whenever the empirical covariance matrix

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \tag{7.22}$$

is invertible, there exists a unique empirical risk minimizer given by

$$\widehat{\theta}_n = \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d} \sum_{i=1}^n (Y_i - \langle \theta, X_i\rangle)^2 = \widehat{\Sigma}_n^{-1}\widehat{S}_n \tag{7.23}$$

where $\widehat{S}_n = n^{-1}\sum_{i=1}^n Y_i X_i$. Hence, whenever $\widehat{\Sigma}_n$ is invertible (almost surely), the MLE is uniquely defined, and equals the *ordinary least squares* estimator given by (7.23).

7.4.2 The unregularized SMP

In this section, we consider uniform excess risk bounds for unpenalized SMP ($\phi \equiv 0$) with respect to the linear Gaussian class $\mathcal{F}$ given by (7.20). This setting is relevant when $n \gg d$, especially when little is known or assumed on the optimal parameter θ^*. We will work under the following

Assumption 7.2 (Non-degenerate design)**.** The covariance matrix Σ is invertible and the empirical covariance matrix $\widehat{\Sigma}_n$ is invertible almost surely.

The fact that Σ is invertible amounts to assuming that X is not supported in any hyperplane of $\mathbf{R}^d$. This assumption is not restrictive, since otherwise one can simply restrict to the span of the support of X, a subspace of $\mathbf{R}^d$; we make it merely for convenience in statements and notations. In addition, a simple induction (see Definition 6.1 in Chapter 6) shows that Assumption 7.2 amounts to assuming that $n \geqslant d$ and that $\mathbb{P}(X \in H) = 0$ for any hyperplane $H \subset \mathbf{R}^d$. Note that the latter is granted whenever X admits a density with respect to the Lebesgue measure. Moreover, as explained in Section 7.4.1, Assumption 7.2 amounts to say that MLE in the model (7.20) is uniquely determined almost surely.

[3]The assumption $\mathbb{E}[Y^2] < +\infty$ is not strictly necessary to ensure that $R(f_\theta)$ is finite for some base measure μ. Indeed, taking $\mu = \mathcal{N}(0,1)$, log-loss writes $\ell(f_\theta, (x,y)) = \langle \theta, x\rangle^2/2 - y\langle \theta, x\rangle$, and the slightly weaker assumption that YX is integrable suffices. We nonetheless take a uniform dominating measure μ and make Assumption 7.1, in order to make the connection with the least-squares problem more explicit.

Once again in this case, SMP leads to an improper estimator, which can be made explicit and satisfies a sharp excess risk bound. Let us introduce the rescaled empirical covariance matrix

$$\widetilde{\Sigma}_n = \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} = \frac{1}{n}\sum_{i=1}^n \widetilde{X}_i\widetilde{X}_i^\top \quad \text{where} \quad \widetilde{X}_i = \Sigma^{-1/2}X_i. \tag{7.24}$$

Note that the rescaled design $\widetilde{X}_i$ is such that $\mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top] = I_d$ for $i = 1, \dots, n$. As explained in Theorem 7.4 below, the excess risk of SMP is connected to the fluctuations of $\widetilde{\Sigma}_n$.

Theorem 7.4. *Assume that Assumptions 7.1 and 7.2 are fulfilled. For the Gaussian linear family $\mathcal{F}$ given by (7.20), SMP is given by*

$$\widetilde{f}_n(\cdot|x) = \mathcal{N}\Big(\langle\widehat{\theta}_n, x\rangle, \big(1 + \big\langle(n\widehat{\Sigma}_n)^{-1}x, x\big\rangle\big)^2\Big). \tag{7.25}$$

In addition, it satisfies the following excess risk bound:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\Big[-\log\Big(1 - \big\langle(n\widehat{\Sigma}_n + XX^\top)^{-1}X, X\big\rangle\Big)\Big] \leqslant \log\Big(1 + \frac{1}{n}\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big]\Big), \tag{7.26}$$

where $\widetilde{\Sigma}_n$ is the rescaled empirical covariance given by (7.24).

The proof of Theorem 7.4 is given in Section 7.7.3 below. The upper bound on the excess risk depends on the distribution of the design through the term $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$, namely through lower relative fluctuations of the empirical covariance matrix $\widehat{\Sigma}_n$ with respect to its population counterpart Σ. Note that this quantity is invariant under linear transformation of $X, X_1, \dots, X_n$.

A key feature of the excess risk bound (7.26) on the SMP is that it only depends on the distribution of X, and *not* on the conditional distribution of Y given X. The expected risk of the SMP is therefore not affected by model misspecification, similarly to what was observed in Section 7.3 for unconditional densities. This is once again a strong departure from the behavior of the MLE, as explained below.

Comparison with MLE and proper estimators. As explained above, MLE is given by $f_{\widehat{\theta}_n}$, where $\widehat{\theta}_n$ is the ordinary least-squares estimator (7.23). In the *well-specified case*, the minimax risk among *proper* estimators is achieved by MLE and equals $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]/(2n)$ (Theorem 6.1 in Chapter 6); hence, the excess risk of SMP is only within a factor 2 of the minimax risk for proper estimators in the well-specified case, despite the fact that the model can be misspecified. In the *misspecified case*, the risk of MLE scales as $\mathbb{E}_{(X,Y)\sim P}[(Y - \langle\theta^*, X\rangle)^2\|\Sigma^{-1/2}X\|^2]/n$ up to lower-order terms, and this dependence is unavoidable for any proper estimator (Proposition 1.6 in Chapter 6). This means that the risk of proper estimators deteriorates under misspecification, and that the minimax risk among proper estimators is infinite, since the previous quantity can be arbitrarily large.

Comparison with the well-specified case. One can in fact show that the first bound in (7.26) on the risk of SMP in the general misspecified case is exactly *twice* the minimax excess risk in the well-specified case. This shows that the general excess risk bound for SMP is intrinsic to the complexity of the problem in this case. Another consequence worth pointing is that the minimax excess risk in the misspecified case is at most twice that of the well-specified case.

Comparison with online algorithms. The minimax regret with respect to the full linear model is infinite, since regret after the first observation is unbounded. Hence, it is not possible to obtain any uniform excess risk bound from online-to-batch conversion of sequential procedures. We discuss non-uniform guarantees in Section 7.4.3.

Link with leverage scores. It is worth noting that the first part of the upper bound (7.26) has a natural interpretation. Indeed, the quantity $\langle (n\widehat{\Sigma}_n + XX^\top)^{-1} X, X\rangle$ is the *leverage score* of X in the sample $X_1, \dots, X_n, X$. This means that the excess risk of SMP can be upper bounded as

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\big[-\log(1-\widehat{\ell}_{n+1})\big], \quad \text{where} \quad \widehat{\ell}_{n+1} = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle$$

is the leverage score of one sample distributed as P_X among $n+1$. Intuitively, the more uneven the leverage scores are, the harder the prediction task will be, since the optimal parameter in the model will effectively be determined by smaller number of points and hence have larger variance.

Upper bounds. A first upper bound on the risk of the SMP can be obtained from (7.26) in the case of Gaussian covariates: when $X \sim \mathcal{N}(0, \Sigma)$, so that $\widetilde{X} \sim \mathcal{N}(0, I_d)$, we have $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})] = nd/(n-d-1)$ (Anderson, 2003; Breiman and Freedman, 1983), giving an upper bound of $\log(1 + d/(n-d-1))$ for SMP.

We now discuss extensions to more general distributions P_X of covariates. By the law of large numbers, one has $\widetilde{\Sigma}_n \to I_d$ as $n \to \infty$ and thus $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) \to d$ almost surely. Hence, one can expect that the excess risk bound (7.26) of the SMP scales as $d/n + o(1/n)$. In order to turn this into an explicit, non-asymptotic bound, we need to control the lower tail of $\widetilde{\Sigma}_n$. This requires some conditions on the distribution of X, in order to ensure even finiteness of $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$:

Assumption 7.3 (Small ball)**.** There exist constants $C \geqslant 1$ and $\alpha \in (0,1)$ such that, for any hyperplane $H \subset \mathbf{R}^d$ and $t > 0$,

$$\mathbb{P}(\mathrm{dist}(\Sigma^{-1/2} X, H) \leqslant t) \leqslant (Ct)^\alpha \,. \tag{7.27}$$

Assumption 7.3 quantifies Assumption 7.2, which states that $\mathbb{P}(X \in H) = 0$ for any hyperplane $H \subset \mathbf{R}^d$. It is equivalent to $\mathbb{P}(|\langle \theta, X\rangle| \leqslant t\|\theta\|_\Sigma) \leqslant (Ct)^\alpha$ for every $\theta \in \mathbf{R}^d$ and $t \in (0,1)$. This condition is a strengthened version of the *small-ball condition* considered by Koltchinskii and Mendelson (2015); Mendelson (2015); Lecué and Mendelson (2016), which amounts to requiring this for a single $t < C^{-1}$. A matching lower bound to (7.27) holds with $\alpha = 1$ and $C = 0.025$ for any distribution of X when $d \geqslant 2$ (Proposition 6.4 in Chapter 6).

Assumption 7.4 (Kurtosis)**.** $\mathbb{E}\|\Sigma^{-1/2} X\|^4 \leqslant \kappa d^2$ for some $\kappa \geqslant 1$.

Assumption 7.4 is a bound on the kurtosis of $\|\Sigma^{-1/2} X\|$, since $\mathbb{E}\|\Sigma^{-1/2} X\|^2 = d$. It is weaker than the following L^2–L^4 equivalence for one-dimensional marginals of X: $(\mathbb{E}\langle X, \theta\rangle^4)^{1/4} \leqslant \kappa^{1/4} (\mathbb{E}\langle X, \theta\rangle^2)^{1/2}$ for all $\theta \in \mathbf{R}^d$ (Oliveira, 2016), and a significantly weaker requirement on X than a sub-Gaussian assumption (Vershynin, 2012).

Corollary 7.1. *Suppose that Assumptions 7.1, 7.2, 7.3 and 7.4 hold, and let $\widetilde{f}_n$ be the SMP given by (7.25). Then, denoting $C' = 28C^4e^{1+9/\alpha}$, for $n \geqslant \min(6d/\alpha, 12\log(12/\alpha)/\alpha)$ we have*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \frac{d}{n}\Big(1 + C'\frac{\kappa d}{n}\Big)\,. \tag{7.28}$$

The proof of Corollary 7.1 is given in Section 7.7. It is a direct consequence of Theorem 7.4, together with an upper bound from Chapter 6 on the excess risk of the ordinary least-squares estimator in the well-specified case. The bound (7.28) deduced from Theorem 7.4 scales as $d/n + O((d/n)^2)$ as $d = o(n)$, with exact first-order constant and order-optimal second-order term $O((d/n)^2)$. The most technical argument is provided in Chapter 6, where a tight control on the smallest eigenvalue of $\widetilde{\Sigma}_n$ and on $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$ is obtained under Assumptions 7.3 and 7.4.

7.4.3 Ridge-regularized SMP

In the previous section, we considered uniform excess risk bounds with respect to the full Gaussian linear model $\mathcal{F}$. We now turn to non-uniform bounds over $\mathcal{F}$, where some dependence on the comparison parameter $\theta \in \mathbf{R}^d$ is allowed. Such guarantees are relevant when uniform bounds over $\mathcal{F}$ are not possible, which occurs either when $d > n$, or when the distribution of covariates X does not satisfy the regularity condition (Assumption 7.2 or 7.3) ensuring finite minimax risk.

Specifically, we investigate excess risk bounds with respect to balls of the form $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ for some $B > 0$. For this purpose, we will consider SMP with Ridge regularization $\phi(\theta) = \lambda\|\theta\|^2/2$ for some $\lambda > 0$. One advantage of the bounds obtained in this setting is that they remain meaningful in the *nonparametric* setting where d may be larger than n.

The upper bound from Theorem 7.5 below does not explicitly depend on the dimension d, but only on the covariance matrix Σ and on $\|\theta\|$. It extends readily to the case where $\mathbf{R}^d$ is replaced by a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, but we keep $\mathbf{R}^d$ in order to keep the setting and notations consistent with those of Section 7.4.2. We work in this section under the following assumption.

Assumption 7.5 (Bounded covariates)**.** $\|X\| \leqslant R$ almost surely for some constant $R > 0$.

Assumption 7.5 is automatically satisfied for instance in the Reproducing Kernel Hilbert Space (RKHS) setting, where the features x are of the form $x = \Phi(x')$ where $x' \in \mathcal{X}'$ is an input variable in some measurable space $\mathcal{X}'$ and $\Phi : \mathcal{X}' \to \mathbf{R}^d$ a measurable map such that the kernel $K : \mathcal{X}' \times \mathcal{X}' \to \mathbf{R}$ given by $K(x', x'') = \langle \Phi(x'), \Phi(x'')\rangle$ is bounded: $K \leqslant R^2$.

Recall that we consider the family $\mathcal{F} = \{f_\theta(\cdot|x) = \mathcal{N}(\langle\theta, x\rangle, 1) : \theta \in \mathbf{R}^d\}$, together with the Ridge penalization $\phi(\theta) = \lambda\|\theta\|^2/2$ for some $\lambda > 0$. Let

$$\widehat{\theta}_{\lambda,n} := \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d}\left\{\frac{1}{n}\sum_{i=1}^n \ell(f_\theta, (X_i, Y_i)) + \frac{\lambda}{2}\|\theta\|^2\right\} = (\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{S}_n$$

denote the Ridge estimator, where we recall that $\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^n X_iX_i^\top$ and $\widehat{S}_n = n^{-1}\sum_{i=1}^n Y_iX_i$, and let us also define

$$\widehat{\Sigma}_\lambda^x = n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d, \quad \widehat{K}_\lambda^x = (\widehat{\Sigma}_\lambda^x)^{-1} \quad \text{and} \quad \lambda' = \frac{n+1}{n}\lambda\,.$$

We also introduce the *degrees of freedom* of the Ridge estimator (Wahba, 1990; Friedman et al., 2001; Wasserman, 2006), given by

$$\mathsf{df}_\lambda(\Sigma) = \mathrm{Tr}[(\Sigma + \lambda I_d)^{-1}\Sigma]\,, \tag{7.29}$$

and note that

$$\mathsf{df}_\lambda(\Sigma) \leqslant \mathrm{Tr}[(\Sigma + \lambda I_d)^{-1}(\Sigma + \lambda I_d)] = d\,. \tag{7.30}$$

Theorem 7.5. *Let $\lambda > 0$. The penalized SMP (7.14) with penalty $\phi(\theta) = \frac{\lambda}{2}\|\theta\|^2$ is well-defined and writes $\widetilde{f}_{\lambda,n}(\cdot|x) = \mathcal{N}(\widetilde{\mu}_\lambda(x), \widetilde{\sigma}^2_\lambda(x))$, where*

$$\widetilde{\sigma}_\lambda(x)^2 = \big((1 - \|x\|^2_{\widehat{K}^x_\lambda})^2 + \lambda\|x\|^2_{(\widehat{K}^x_\lambda)^2}\big)^{-1} \tag{7.31}$$

and

$$\widetilde{\mu}_\lambda(x) = \langle\widehat{\theta}_{\lambda',n}, x\rangle - \lambda\widetilde{\sigma}_\lambda(x)^2\langle\widehat{\theta}_{\lambda',n}, x\rangle_{\widehat{K}^x_\lambda}. \tag{7.32}$$

In addition, under Assumptions 7.1 and 7.5, we have

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta\in\mathbf{R}^d}\Big\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\Big\} \leqslant 1.25\cdot\frac{\mathsf{df}_\lambda(\Sigma)}{n+1} \tag{7.33}$$

for every $\lambda \geqslant 2R^2/(n+1)$, where $\mathsf{df}_\lambda(\Sigma)$ is given by (7.29).

Although the space of parameters is finite dimensional (of dimension d), the bound (7.33) is "non-parametric" in the sense that it does not feature any explicit dependence on d; rather, it only depends on the spectral properties of Σ through $\mathsf{df}_\lambda(\Sigma)$. In particular, it remains nonvacuous even when $d \gg n$; in fact, as mentioned above, Theorem 7.5 remains valid (with the same proof, up to minor changes in terminology and notations) in the case of an infinite-dimensional RKHS.

Let us now discuss some consequences of Theorem 7.5.

- *Finite-dimensional case.* Since $\mathsf{df}_\lambda(\Sigma) \leqslant d$ (see (7.30)), Theorem 7.5 entails, for $\lambda = 2R^2/(n+1)$, that
$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - \inf_{\|\theta\|\leqslant B} R(f_\theta) \leqslant \frac{1.25d + B^2R^2}{n+1} \tag{7.34}$$
for every $B > 0$. This gives an excess risk bound of $O((d + B^2R^2)/n)$. Proposition 7.3 below further refines this finite-dimensional bound.

- *Slow, dimension-free rate.* Since $\mathsf{df}_\lambda(\Sigma) \leqslant \mathrm{Tr}(\Sigma)/\lambda \leqslant R^2/\lambda$ for $\lambda > 0$, Theorem 7.5 yields, for every $\lambda \geqslant 2R^2/(n+1)$ and $B > 0$,
$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - \inf_{\|\theta\|\leqslant B} R(f_\theta) \leqslant \frac{1.25R^2}{\lambda(n+1)} + \frac{\lambda B^2}{2} \leqslant \frac{2BR}{\sqrt{n}} + \frac{B^2R^2}{n}, \tag{7.35}$$
where the second inequality is obtained with $\lambda = \max(2R^2/(n+1), 2R/(B\sqrt{n+1}))$. This corresponds to the standard nonparametric slow rate for regression, except that it does not depend on the range of Y. This requires no assumption on the covariance Σ, aside from the inequality $\mathrm{Tr}(\Sigma) \leqslant R^2$ implied by the assumption $\|X\| \leqslant R$.

- *Nonparametric case.* More precise results can be obtained in terms of spectral properties of Σ. Let b be the rate of decay of the eigenvalues of Σ, such that $\mathsf{df}_\lambda(\Sigma) = O(\lambda^{-1/b})$. Then, Theorem 7.5 yields

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\|\theta\|\leqslant B} R(f_\theta) \leqslant O\Big(\frac{\lambda^{-1/b}}{n} + \lambda B^2\Big) = O(B^{2/(b+1)} n^{-b/(b+1)}) \tag{7.36}$$

  for $\lambda \asymp (B^2 n)^{-b/(b+1)}$. This matches the minimax rate for regression with unit noise over balls of RKHSs in the well-specified case, without additional assumptions on θ (Caponnetto and De Vito, 2007).

In the finite-dimensional case where $n \gg d$, one can improve the quadratic dependence on the norm $B = \|\theta\|$. This yields bounds that are appropriate when the covariate distribution is possibly degenerate, in the sense that Assumption 7.2 does not hold, so that excess risk bounds uniform in θ are no longer achievable.

Proposition 7.3. *Grant Assumptions 7.1 and 7.5. Then, for any $B > 0$, the Ridge-SMP $\widetilde{f}_{\lambda,n}$ of Theorem 7.5 with $\lambda = d/(B^2(n+1))$ satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta\in\mathbf{R}^d\,:\,\|\theta\|\leqslant B} R(f_\theta) \leqslant \frac{5d\log\big(2 + BR/\sqrt{d}\big)}{n+1}\,. \tag{7.37}$$

This bound is of order $O(d\log(BR/\sqrt{d})/n)$. This improves a bound obtained by Kakade and Ng (2005) (with optimized parameters, and after online-to-batch conversion) of $O(d\log(B^2R^2n/d)/n)$ from the sequential setting through Bayesian mixture strategies, by removing an extra $O(\log n)$ term.

Remark 7.2 (Parameter scaling). The previous results are valid for arbitrary parameters BR, d, n. In order to make these bounds more concrete, we now discuss some natural scaling for the norm BR. Consider the finite-dimensional case where $n \gg d$, and assume that Σ is well-conditioned, in the sense that $c := \|\Sigma\|_{\mathrm{op}} \cdot \|\Sigma^{-1}\|_{\mathrm{op}} = O(1)$. This means that X is approximately isotropic, or equivalently that the chosen norm on $\mathbf{R}^d$ does not favor specific directions, but rather controls signal strength $\|\theta\|_\Sigma \asymp \|\theta\|$; this can be ensured in practice by rescaling covariates. Also, assume that $\|\Sigma^{-1/2}X\| \leqslant \rho\sqrt{d}$ for some $\rho \geqslant 1$, a bounded leverage condition (Hsu et al., 2014), and let $\psi := \|\theta\|_\Sigma = \mathbb{E}[\langle\theta, X\rangle^2]^{1/2}$ denote signal strength. Then,

$$\|\theta\| \cdot \|X\| \leqslant \|\Sigma^{-1/2}\|_{\mathrm{op}} \cdot \|\Sigma^{1/2}\theta\| \cdot \|\Sigma^{1/2}\|_{\mathrm{op}} \cdot \|\Sigma^{-1/2}X\| \leqslant c^{1/2}\rho\psi\sqrt{d}\,,$$

so that $BR \leqslant c^{1/2}\rho\psi\sqrt{d} = O(\sqrt{d})$.

On the other hand, one can have $BR \ll \sqrt{d}$: this occurs in the "nonparametric" case where Σ has eigenvalue decay, and θ lies close to the space spanned by the leading eigenvectors of Σ; in this case, $\mathsf{df}_\lambda(\Sigma) \ll d$, and it is beneficial to replace d by $\mathsf{df}_\lambda(\Sigma)$ as in Theorem 7.5.

We close this section by pointing out that, in the well-conditioned finite-dimensional regime where $BR = O(\sqrt{d})$, the bounds (7.34) and (7.37) both yield a $O(d/n)$ guarantee, while the latter has an improved dependence on signal strength.

7.5 Logistic regression

In this section, we consider conditional density estimation with a binary response, using the logistic model. Section 7.5.1 introduces the setting. We consider the unpenalized SMP ($\phi \equiv 0$) in Section 7.5.2 and contrast its predictions with those of MLE. In Section 7.5.3 we introduce the Logistic SMP procedure with Ridge penalization, and establish a non-asymptotic bound on its excess risk.

7.5.1 Setting

We consider binary labels in $\mathcal{Y} = \{-1, 1\}$, with counting measure $\mu = \delta_0 + \delta_1$, while $\mathcal{X} = \mathbf{R}^d$. The *logistic model* is the family of conditional distributions given by

$$\mathcal{F} = \{f_\theta : \theta \in \mathbf{R}^d\}, \quad \text{where} \quad f_\theta(1|x) := 1 - f_\theta(-1|x) = \sigma(\langle \theta, x \rangle) \tag{7.38}$$

for any $x \in \mathbf{R}^d$, with $\sigma(u) = e^u/(1+e^u)$ for $u \in \mathbf{R}$ the *sigmoid* function. Since $\sigma(-u) = 1 - \sigma(u)$, one simply has $f_\theta(y|x) = \sigma(y\langle \theta, x \rangle)$ for $x \in \mathbf{R}^d$ and $y \in \{-1, 1\}$. The log-loss of $f_\theta \in \mathcal{F}$ at a sample $(x, y) \in \mathbf{R}^d \times \{-1, 1\}$ writes

$$\ell(f_\theta, (x, y)) = -\log f_\theta(y|x) = \log(1 + e^{-y\langle \theta, x \rangle}) = \ell(-y\langle \theta, x \rangle)\,, \tag{7.39}$$

where we introduced the *logistic* loss $\ell(u) = \log(1+e^u)$ for $u \in \mathbf{R}$. Let (X, Y) have distribution P on $\mathbf{R}^d \times \{-1, 1\}$, such that $\mathbb{E}\|X\| < +\infty$. Since $\ell'(u) = \sigma(u) \in [0, 1]$ for any $u \in \mathbf{R}$, we have $0 \leqslant \ell(u) \leqslant \log 2 + |u|$ so that $\ell(-Y\langle \theta, X \rangle) \leqslant \log 2 + \|\theta\|\|X\|$, and the risk of f_θ, namely

$$R(f_\theta) = \mathbb{E}[\ell(-Y\langle \theta, X \rangle)]\,, \tag{7.40}$$

is well-defined. Given a sample (X_i, Y_i), $1 \leqslant i \leqslant n$, a MLE $\widehat{\theta}_n$ is given by

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(-Y_i \langle \theta, X_i \rangle)\,, \tag{7.41}$$

A MLE (7.41) does not always exist, and may not be unique. Indeed, it is well-known (see Candès and Sur, 2018 for recent results on this topic in the high-dimensional regime) that there is no MLE (7.41) whenever the sets $\{X_i : Y_i = 1\}$ and $\{X_i : Y_i = -1\}$ are strictly *linearly separated* by a hyperplane, namely when one can find $\theta \in \mathbf{R}^d$ such that $Y_i\langle \theta, X_i \rangle > 0$ for all $i = 1, \dots, n$ (indeed, in this case the empirical risk of $t\theta$ converges to 0 as $t \to +\infty$, while the empirical risk is positive on $\mathbf{R}^d$). In addition, when a MLE exists in $\mathbf{R}^d$, one can see that it is unique if and only if $V = \operatorname{span}(X_1, \dots, X_n) = \mathbf{R}^d$: in this case, the empirical risk is strictly convex on $\mathbf{R}^d$ since $\ell : \mathbf{R} \to \mathbf{R}$ is.

It is convenient to enrich the class $\mathcal{F}$ given by (7.38) to ensure existence (though not uniqueness) of MLE in the separated case. Specifically, define the model $\overline{\mathcal{F}}$ obtained by adding to $\mathcal{F}$ the conditional densities $f_{\infty,\theta}$ for $\theta \in \mathbf{R}^d$, $\|\theta\| = 1$, defined by $f_{\infty,\theta}(1|x) = 1$ if $\langle \theta, x \rangle > 0$, 0 if $\langle \theta, x \rangle < 0$ and $1/2$ if $\langle \theta, x \rangle = 0$. Denote by $\overline{\Theta}$ the parameter space obtained by adding to $\mathbf{R}^d$ elements of the form (∞, θ). We note that MLE exists in $\overline{\mathcal{F}}$ in the separated case, although it is not unique since it depends on the choice of a separating hyperplane defined by θ. Given a choice of MLE, we let

$$\widehat{\theta}_n^{(x,y)} = \operatorname*{arg\,min}_{\theta \in \overline{\Theta}} \Big\{ \sum_{i=1}^n \ell(f_\theta, (X_i, Y_i)) + \ell(f_\theta, (x, y)) \Big\} \tag{7.42}$$

for any $(x, y) \in \mathbf{R}^d \times \{-1, 1\}$. It is also convenient to let $Z_i = -Y_i X_i$; then, one has $\widehat{\theta}_n^{(x,y)} = \widehat{\theta}_n^{-yx}$, where for $z \in \mathbf{R}^d$ we define (with a slight abuse of notation for $\theta \in \overline{\Theta} \setminus \mathbf{R}^d$)

$$\widehat{\theta}_n^z = \operatorname*{arg\,min}_{\theta \in \overline{\Theta}} \Big\{ \sum_{i=1}^n \ell(\langle \theta, Z_i \rangle) + \ell(\langle \theta, z \rangle) \Big\} . \tag{7.43}$$

7.5.2 SMP for logistic regression

Let us now instantiate SMP as well as Theorem 7.2 to the logistic family.

Proposition 7.4. *For the family of logistic conditional distributions* (7.38)*, SMP writes*

$$\widetilde{f}_n(y|x) = \frac{f_{\widehat{\theta}_n^{(x,y)}}(y|x)}{f_{\widehat{\theta}_n^{(x,1)}}(1|x) + f_{\widehat{\theta}_n^{(x,-1)}}(-1|x)} = \frac{\sigma(\langle \widehat{\theta}_n^{(x,y)}, yx \rangle)}{\sigma(\langle \widehat{\theta}_n^{(x,1)}, x \rangle) + \sigma(\langle \widehat{\theta}_n^{(x,-1)}, -x \rangle)} \tag{7.44}$$

for every $x \in \mathbf{R}^d$ and $y \in \{-1, 1\}$. Unlike the MLE (7.42)*, SMP is always well-defined and unique. We always have that $\widetilde{f}_n(y|x) \in (0, 1)$ and it does not depend on the choice of a MLE in the linearly separated case. In addition, it satisfies the following excess risk bound*:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}_{Z_1^n, Z}\big[\sigma(\langle \widehat{\theta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\theta}_n^{Z}, Z \rangle)\big] , \tag{7.45}$$

where $Z_1, \dots, Z_n, Z$ are i.i.d. variables distributed as $-YX$.

The proof of Proposition 7.4 is given in Section 7.7.4 below. Unlike MLE, SMP is always well-defined and outputs predictions in $(0, 1)$. Indeed, the numerator in (7.44) belongs to $(0, 1]$, and whenever the points $Y_1 X_1, \dots, Y_n X_n, yx$ belong to a half-space (so that MLE does not exist in $\mathbf{R}^d$), we have $f_{\widehat{\theta}_n^{(x,y)}}(y|x) = 1$, so that the prediction of SMP is well-defined and does not depend on the choice of MLE in (7.42), see the proof of Proposition 7.4 for details.

Comparison with MLE. SMP corrects a well-known deficiency of MLE, which tends to produce overly confident and ill-calibrated predictions (Sur and Candès, 2019). To emphasize this effect, consider the case of a point x for which the virtual datasets $(X_1, Y_1), \dots, (X_n, Y_n), (x, y)$ are separated for both $y = -1$ and $y = 1$. Then, the prediction $\widehat{f}_n(1|x)$ of an MLE $\widehat{f}_n \in \overline{\mathcal{F}}$ can be either 1 or 0, both being possible depending on the specific choice of separating hyperplane. Hence, in this case the prediction of MLE is both highly confident and dependent on an arbitrary choice. By contrast, in this situation SMP gives equal probability $1/2$ to both classes, reflecting the uncertainty for such points x.

A non-Bayesian approach to calibration. As for the Gaussian linear model (Section 7.4), SMP returns more uncertain conditional distributions for input points x with high "leverage", namely strong influence on the prediction of MLE at this point. This provides a simple and natural approach to calibration of probabilistic predictions for logistic regression, which does not rely on Bayesian methods. Such an approach is appealing on computational grounds, since the prediction $\widetilde{f}(\cdot|x)$ of SMP is obtained by solving two logistic regressions (7.42), bypassing the need for approximate posterior sampling.

Comparison with stability approaches. Approaches based on stability of the loss (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Srebro et al., 2010; Koren and Levy, 2015) would lead to a control of the excess risk involving $\ell(\langle \widehat{\theta}_n^{-Z}, Z\rangle) - \ell(\langle \widehat{\theta}_n^{Z}, Z\rangle)$, while Proposition 7.4 involves $\sigma(\langle \widehat{\theta}_n^{-Z}, Z\rangle) - \sigma(\langle \widehat{\theta}_n^{Z}, Z\rangle)$, where we recall that $\ell(u) = \log(1+e^u)$ and $\sigma(u) = 1/(1+e^{-u})$. Whenever $u' \approx u \gg 1$, we have $\ell(u') - \ell(u) \approx \ell'(u)\cdot(u'-u) \approx u'-u$, while $\sigma(u') - \sigma(u) \approx \sigma'(u)\cdot(u'-u) \approx e^{-u}\cdot(u'-u)$. In this case, the SMP bound is exponentially smaller than the loss stability bound. This roughly explains why we are able to remove terms of order e^{BR} from our upper bound on the excess risk of SMP, provided in the next section.

7.5.3 Excess risk bounds for Ridge-regularized SMP

In order to obtain explicit and precise non-asymptotic guarantees, we consider a Ridge-regularized variant of SMP for logistic regression. Specifically, for $\lambda > 0$ we consider the penalty $\phi(\theta) = \lambda\|\theta\|^2/2$. The corresponding penalized SMP can be computed as follows: for every $z \in \mathbf{R}^d$, let

$$\widehat{\theta}^z_{\lambda,n} := \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d} \left\{ \frac{1}{n+1}\Big(\sum_{i=1}^n \ell(\langle \theta, Z_i\rangle) + \ell(\langle \theta, z\rangle)\Big) + \frac{\lambda}{2}\|\theta\|^2 \right\}. \tag{7.46}$$

Note that $\widehat{\theta}^z_{\lambda,n} \in \mathbf{R}^d$ exists and is unique, since the regularized objective in (7.46) is strongly convex, hence strictly convex and diverging as $\|\theta\| \to +\infty$. As before, we let $\widehat{\theta}^{(x,y)}_{\lambda,n} = \widehat{\theta}^{-yx}_{\lambda,n}$ for $(x,y) \in \mathbf{R}^d \times \{-1,1\}$. Now, following Theorem 7.2, the regularized SMP writes in this case

$$\widetilde{f}_{\lambda,n}(y|x) = \frac{\sigma(y\langle \widehat{\theta}^{(x,y)}_{\lambda,n}, x\rangle)\, e^{-\lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2/2}}{\sigma(\langle \widehat{\theta}^{(x,1)}_{\lambda,n}, x\rangle)\, e^{-\lambda\|\widehat{\theta}^{(x,1)}_{\lambda,n}\|^2/2} + \sigma(-\langle \widehat{\theta}^{(x,-1)}_{\lambda,n}, x\rangle)\, e^{-\lambda\|\widehat{\theta}^{(x,-1)}_{\lambda,n}\|^2/2}} \tag{7.47}$$

for any $(x,y) \in \mathbf{R}^d \times \{-1,1\}$, and comes as before at the cost of two ridge-regularized logistic regressions.

We will work under Assumption 7.5, namely $\|X\| \leqslant R$ almost surely, as in Section 7.4.3 for the Gaussian linear model. Our main guarantee for Ridge-regularized SMP is stated in a nonparametric setting, where dependence on the dimension d is kept implicit through the degrees of freedom (7.29).

Theorem 7.6. *Grant Assumption 7.5, and assume that $\lambda \geqslant 2R^2/(n+1)$. Then, the Ridge-regularized logistic SMP given by (7.47) satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] \leqslant R(f_\theta) + e\cdot\frac{\mathsf{df}_{4\lambda}(\Sigma)}{n} + \frac{\lambda}{2}\|\theta\|^2 \tag{7.48}$$

for every $\theta \in \mathbf{R}^d$, where we recall that $\mathsf{df}_\lambda(\Sigma) = \mathrm{Tr}[(\Sigma+\lambda I)^{-1}\Sigma]$.

The upper bound (7.48) is a *fast rate* excess risk guarantee; it is worth noting that it only requires bounded covariates (Assumption 7.5). In particular, it requires no assumption on the conditional distribution of Y given X. Furthermore, when the feature X comes from a bounded kernel (see the discussion in Section 7.4.3 above), the bound (7.48) is valid *under no assumption* on the distribution of (X,Y).

We note that Marteau-Ferey et al. (2019) established nonparametric fast rate guarantees akin to (7.48) for the Ridge-regularized estimator in the *well-specified* case. Compared

to (7.48), their bias term, while also equal to λB^2 under the sole assumption $\|\theta\| \leqslant B$, can be further improved under stronger assumptions on θ (namely, faster coefficient decay, or *source condition*, Caponnetto and De Vito, 2007). On the other hand, this result relies on the assumption of a well-specified model, and under our general assumptions such rates would exhibit exponential dependence in BR (Hazan et al., 2014).

Since $\mathsf{df}_{4\lambda}(\Sigma) \leqslant d$ for every λ, we deduce the following result in finite dimension.

Corollary 7.2. *Under Assumption 7.5, the Ridge-regularized logistic SMP* $\widetilde{f}_{\lambda,n}$ (7.47) *with* $\lambda = 2R^2/(n+1)$ *satisfies, for every* $B > 0$,

$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - \inf_{\|\theta\| \leqslant B} R(f_\theta) \leqslant \frac{e \cdot d + B^2 R^2}{n} \,. \tag{7.49}$$

Note that under the well-conditioned scaling of dimension d with constant signal strength, namely $BR = O(\sqrt{d})$ (see Remark 7.2 from Section 7.4.3), Corollary 7.2 yields an excess risk of $O(d/n)$.

Bypassing a lower bound. Under Assumption 7.5, Corollary 7.2 leads to an upper bound for Ridge SMP of $O((d + B^2R^2)/n)$ with respect to the ball $\|\theta\| \leqslant B$. By contrast, Hazan et al. (2014) showed a lower bound for any *proper* estimator (including the norm-constrained or Ridge-penalized MLE, or any stochastic optimization procedure) of order $\min(BR/\sqrt{n}, de^{BR}/n)$ in the worst case. We note that SMP is an improper estimator, as the log-odds ratio $\log(\widetilde{f}_{\lambda,n}(1|x)/\widetilde{f}_{\lambda,n}(-1|x))$ is nonlinear in x, and that it bypasses the lower bound for proper estimators.

A practical improper estimator. Fast rates of order $O(d \log(BRn)/n)$ are obtained by Kakade and Ng (2005); Foster et al. (2018) under Assumption 7.5, by applying online-to-offline conversion (averaging) to a Bayes mixture sequential procedure, with prior on θ uniform over the ball of radius B (Foster et al., 2018) or Gaussian (Kakade and Ng, 2005). This bound has an even better dependence on B (logarithmic instead of quadratic) than Corollary 7.2, although it also has a slightly worse dependence in n (additional $\log n$ factor); Theorem 7.6 additionally replaces d by $\mathsf{df}_{4\lambda}(\Sigma)$. The main advantage of SMP over Bayes is that it is computationally less demanding: it replaces a problem of posterior sampling by one of optimization, since it requires training two updated logistic regressions, starting for instance at the Ridge-penalized MLE. Therefore, we partly answer an open problem from Foster et al. (2018), about finding an efficient alternative with fast rate, at least in the batch statistical learning case. We note however that SMP is still more computationally demanding at prediction time than MLE, because of the required updates of the logistic risk minimization problem.

Overview of guarantees for logistic regression. Logistic regression with bounded features $\|X\| \leqslant R$ over the ball $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ is (when restricting to proper estimators) a convex and R-Lipschitz stochastic optimization problem over a bounded domain. This implies that a *slow rate* of $O(BR/\sqrt{n})$ can be achieved by properly-tuned averaged projected online gradient descent (Robbins and Monro, 1951; Zinkevich, 2003; Shalev-Shwartz, 2012; Bubeck, 2015; Hazan, 2016), Ridge-regularized ERM over $\mathbf{R}^d$ (Bousquet and Elisseeff, 2002; Sridharan et al., 2009), or (as a linear prediction problem) constrained ERM over $\mathcal{F}_B$ (Kakade et al., 2009; Bartlett and Mendelson, 2002; Meir and Zhang, 2003). Under the same assumptions,

the logistic loss is e^{-BR}-exp-concave over $\mathcal{F}_B$, implying that a rate of $O(de^{BR}/n)$ can be achieved (up to potential $\log n$ factors) through the (averaged) Exponential Weights (Hazan et al., 2007; Vovk, 1998) or Online Newton Step algorithms (Hazan et al., 2007; Mahdavi et al., 2015), as well as ERM over $\mathcal{F}_B$ (Koren and Levy, 2015; Gonen and Shalev-Shwartz, 2018; Mehta, 2017). The improved dependence on n in this bound is typically outweighed by the prohibitive exponential dependence on parameter norm. As mentioned before, a lower bound of Hazan et al. (2014) shows that, without further assumptions, no proper (within model $\mathcal{F}$) estimator can improve over the $O(\min(BR/\sqrt{n}, de^{BR}/n))$ guarantee. In order to bypass this lower bound, one has to resort to improper procedures (Foster et al., 2018). This is the approach taken by Foster et al. (2018); Kakade and Ng (2005) and ourselves, enabling improved guarantees without further assumptions, as discussed above.

Another line of work (Bach, 2010, 2014; Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019) studies the behavior of specific (within-model) estimators, such as Ridge-regularized MLE or stochastic approximation procedures, in a distribution-dependent fashion. A key technique in these refined analyses is the use of (generalized) *self-concordance* of logistic loss, introduced by Bach (2010), namely a control of the third derivative in terms of the second. Following progress in Bach (2010, 2014), Bach and Moulines (2013) introduces a stochastic approximation algorithm with excess risk $O(\rho^3 d(BR)^4/n)$, where ρ is a distribution-dependent curvature parameter. This bound eliminates dependence on the smallest eigenvalue of Hessian at the optimum (Bach, 2014), but does not lead to the correct scaling in the finite-dimensional case with $BR = O(\sqrt{d})$, or in the nonparametric setting due to dependence on d instead of $\mathsf{df}_\lambda(\Sigma)$ (see Remark 7.2). In finite dimension, a tight non-asymptotic guarantee for MLE is obtained by Ostrovskii and Bach (2018), with an excess risk of $O(d_{\text{eff}}/n)$ for $n \gtrsim \max(\rho d_{\text{eff}}, d\log d)$, where d_{eff} denotes the *effective dimension* characterizing the asymptotic risk of MLE (7.3). These results are extended by Marteau-Ferey et al. (2019) in the well-specified nonparametric setting, with sharp risk bounds for the Ridge-regularized MLE. In the worst case, the distribution-dependent constants ρ and d_{eff} scale with e^{BR} (Bach and Moulines, 2013), although they can be much smaller for more favorable distributions. Despite the difference in assumptions, from a technical point of view, our analysis of the bound on the SMP excess risk also uses self-concordance.

In addition to these non-asymptotic analyses, a recent line of work (Sur and Candès, 2019; Barbier et al., 2019; Candès and Sur, 2018) studies logistic regression under high-dimensional asymptotics where $d \asymp n$. This asymptotic approach differs from the non-asymptotic one in that it provides an exact characterization of the error, but under highly specific distributional assumptions (well-specified model and Gaussian or jointly independent features).

7.6 Conclusion

In this chapter, we derive excess risk bounds for predictive density estimation under logarithmic loss, which hold under misspecification. Minimizing these excess risk bounds naturally leads to a new improper (out-of-model) procedure, which we call *Sample Minmax Predictor* (SMP). On several problems, we show that the resulting bound, which is based on a refinement of the stability argument tailored for the logarithmic loss, scales as d/n, irrespective of the true distribution. This contrasts with estimators taking values within the model, whose performance typically degrade under misspecification, where it exhibits unbounded constants. This estimator provides an alternative to approaches based on online-to-offline conversion (Barron,

1987; Catoni, 2004; Cesa-Bianchi et al., 2004; Audibert, 2009) of sequential procedures, whose rates feature an additional logarithmic dependence on sample size, and may be infinite for unbounded models.

We apply SMP to the Gaussian linear model. In this case, SMP can be described explicitly, and achieves in the general misspecified case at most twice the minimax risk in the well-specified case, for every distribution of covariates. We then consider a Ridge-regularized variant, which achieves nonparametric fast rates, as well as a bound with a logarithmic dependence on the diameter of the comparison class in the finite-dimensional case.

We then consider logistic regression. Here, (Ridge-penalized) SMP is a simple explicit procedure, whose predictions can be computed at the cost of two logistic regressions. From a statistical perspective, it achieves fast excess risk rates even for worst-case distributions; such guarantees are known to be out of reach for any *proper* procedure (Hazan et al., 2014). In the batch i.i.d. case, this provides a more practical alternative to the improper estimator from Foster et al. (2018), which relies on Bayesian mixtures, thereby partly addressing an open question from this article. This work leaves a number of open problems and future directions:

- First, the excess risk bounds in this chapter only hold in expectation, and not with exponential probability. This limitation is shared by procedures relying on online-to-batch conversion (Catoni, 2004; Audibert, 2008, 2009; Foster et al., 2018). In particular, the high-probability bound stated by Foster et al. (2018) for a procedure based on a "confidence boosting" technique from Mehta (2017) appears to be incorrect: specifically, Equation (17) herein is obtained by applying Markov's inequality to the excess risk; however, this quantity can take negative values since the predictor is outside the class. Designing procedures that achieve high (exponential) probability excess risk bounds that do not degrade under model misspecification is an interesting direction for future work.

- Second, it could be interesting to adapt the proposed method to online logistic regression, with a regret bound for individual sequences. We believe this to be feasible by adapting the procedure and proof technique, and leave this task to future work.

- Another possibility is to apply SMP to other (conditional or otherwise) models beyond the Gaussian linear and logistic ones considered here, such as generalized linear models (McCullagh and Nelder, 1989), or (even in the logistic case) nonparametric classes beyond the RKHS balls considered here.

- Finally, Theorem 7.3 shows that in the Gaussian model, the Bayes predictive posterior under uniform prior *equalizes* excess risk over all distributions in the misspecified case. This reveals the critical role of averaging under misspecification, where it can mitigate slower posterior concentration rate. It would be interesting to extend this finding to other models, and investigate conditions on the model and prior under which uniform non-asymptotic bounds (such as Theorem 7.3 or our guarantees for SMP) hold for Bayesian methods.

On a more general note, statistical learning with logarithmic loss (that is, misspecified Kullback-Leibler density estimation) possesses specific properties, which can be exploited to obtain more precise results than generic approaches applicable to general loss functions (which often suffer from the unboundedness of logarithmic loss). This has been exploited successfully in the sequential case where cumulative criteria are considered (Merhav and Feder, 1998;

Cesa-Bianchi and Lugosi, 2006); while the present work provides similar guarantees for the statistical learning setting, we expect that further advances are possible on this subject.

7.7 Proofs

7.7.1 Proofs of general excess risk bounds (Section 7.2)

Proof of Theorem 7.1. Let Z_1^n, Z denote $n+1$ i.i.d. variables distributed as P. We have

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] &= \mathbb{E}_{Z_1^n,Z}[\ell(\widehat{g}_n, Z)] - \inf_{f\in\mathcal{F}} \mathbb{E}_{Z_1^n,Z}\left[\frac{1}{n+1}\left\{\sum_{i=1}^n \ell_\phi(f,Z_i) + \ell_\phi(f,Z)\right\}\right] \\ &= \mathbb{E}_{Z_1^n,Z}[\ell(\widehat{g}_n, Z)] - \mathbb{E}_{Z_1^n,Z}\left[\inf_{f\in\mathcal{F}} \frac{1}{n+1}\left\{\sum_{i=1}^n \ell_\phi(f,Z_i) + \ell_\phi(f,Z)\right\}\right] - \Delta_n\end{aligned}$$

where we denoted

$$\Delta_n = \inf_{f\in\mathcal{F}} \mathbb{E}\left[\frac{1}{n+1}\left\{\sum_{i=1}^n \ell_\phi(f,Z_i)+\ell_\phi(f,Z)\right\}\right] - \mathbb{E}\left[\inf_{f\in\mathcal{F}} \frac{1}{n+1}\left\{\sum_{i=1}^n \ell_\phi(f,Z_i)+\ell_\phi(f,Z)\right\}\right] \geqslant 0. \tag{7.50}$$

In particular, by definition of $\widehat{f}^Z_{\phi,n}$,

$$\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] + \Delta_n = \mathbb{E}_{Z_1^n,Z}[\ell(\widehat{g}_n, Z)] - \frac{1}{n+1}\mathbb{E}\left[\sum_{i=1}^n \ell_\phi(\widehat{f}^Z_{\phi,n}, Z_i) + \ell_\phi(\widehat{f}^Z_{\phi,n}, Z)\right]. \tag{7.51}$$

Since the distribution of the i.i.d. sample $(Z_1,\dots,Z_n,Z)$ is preserved by exchanging Z and Z_i, we have $\mathbb{E}[\ell_\phi(\widehat{f}^Z_{\phi,n}, Z_i)] = \mathbb{E}[\ell_\phi(\widehat{f}^Z_{\phi,n}, Z)]$ for $i = 1,\dots,n$ (recall that $\widehat{f}^Z_{\phi,n}$ is chosen symmetrically in $Z_1,\dots,Z_n,Z$). Hence, (7.51) becomes

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] + \Delta_n &= \mathbb{E}_{Z_1^n,Z}\big[\ell(\widehat{g}_n, Z) - \ell_\phi(\widehat{f}^Z_{\phi,n}, Z)\big] \\ &= \mathbb{E}_{Z_1^n,X}\mathbb{E}_{Y|X}\big[\ell(\widehat{g}_n(X), Y) - \ell_\phi(\widehat{f}^{(X,Y)}_{\phi,n}(X), Y)\big] \\ &\leqslant \mathbb{E}_{Z_1^n,X}\left[\sup_{y\in\mathcal{Y}}\big\{\ell(\widehat{g}_n(X), y) - \ell_\phi(\widehat{f}^{(X,y)}_{\phi,n}(X), y)\big\}\right],\end{aligned} \tag{7.52}$$

which implies the bound (7.10) since $\Delta_n \geqslant 0$. The remaining claims follow directly. □

Proof of Theorem 7.2. In the case of the logarithmic loss $\ell(p,(x,y)) = -\log p(y|x)$, we have for every density p on $\mathcal{Y}$ and $x \in \mathcal{X}$:

$$\sup_{y\in\mathcal{Y}}\big\{\ell(p,y) - \ell_\phi(\widehat{f}^{(x,y)}_{\phi,n}(x), y)\big\} = \sup_{y\in\mathcal{Y}} \log \frac{\widehat{f}^{(x,y)}_{\phi,n}(y|x) e^{-\phi(\widehat{f}^{(x,y)}_{\phi,n})}}{p(y)}. \tag{7.53}$$

Now, Theorem 7.2 follows from Theorem 7.1 together with Lemma 7.1 below, where we consider $g(y) = \widehat{f}^{(x,y)}_{\phi,n}(y|x) e^{-\phi(\widehat{f}^{(x,y)}_{\phi,n})}$. □

Lemma 7.1. *Let $g : \mathcal{Y} \to [0,+\infty]$ be a measurable function such that $\int_{\mathcal{Y}} g \mathrm{d}\mu \in \mathbf{R}_+^*$. Then,*

$$\inf_p \sup_{y\in\mathcal{Y}} \log \frac{g(y)}{p(y)} = \log\left(\int_{\mathcal{Y}} g(y)\mu(\mathrm{d}y)\right), \tag{7.54}$$

where the infimum in (7.54) spans over all probability densities $p : \mathcal{Y} \to \mathbf{R}^+$ with respect to μ, and the infimum is reached at

$$p^* = \frac{g}{\int_{\mathcal{Y}} g \mathrm{d}\mu} \,. \tag{7.55}$$

Proof. For every density p, denote $C(p) = \sup_{y\in\mathcal{Y}} \log g(y)/p(y)$. By definition, $p(y) \geqslant e^{-C(p)} g(y)$, so that since p is a density

$$1 = \int_{\mathcal{Y}} p(y)\mu(\mathrm{d}y) \geqslant e^{-C(p)} \int_{\mathcal{Y}} g(y)\mu(\mathrm{d}y) \,,$$

so that $C(p) \geqslant \log\big(\int_{\mathcal{Y}} g \mathrm{d}\mu\big)$. Since $C(p^*) = \log\big(\int_{\mathcal{Y}} g \mathrm{d}\mu\big)$, this concludes the proof. □

We will sometimes also use the following observation:

Lemma 7.2. *The expected excess risk of the SMP is equal to*:

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] = \mathbb{E}_{Z_1^n,X}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X) e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})} \mu(\mathrm{d}y)\Big)\Big] - \Delta_n \,, \tag{7.56}$$

where, letting $Z_1, \dots, Z_{n+1}$ be i.i.d. sample from P and f^* a risk minimizer (when it exists),

$$\begin{aligned} \Delta_n &= \frac{1}{n+1} \inf_{f\in\mathcal{F}} \mathbb{E}\Big[\sum_{i=1}^{n+1} \ell_\phi(f, Z_i) - \sum_{i=1}^{n+1} \ell_\phi(\widehat{f}_{\phi,n+1}, Z_i)\Big] \\ &= \frac{1}{n+1} \mathbb{E}\Big[\sum_{i=1}^{n+1} \ell_\phi(f^*, Z_i) - \sum_{i=1}^{n+1} \ell_\phi(\widehat{f}_{\phi,n+1}, Z_i)\Big] . \end{aligned} \tag{7.57}$$

Proof. This follows from the fact that inequality (7.52) is an equality when $\widehat{g}_n = \widetilde{f}_{\phi,n}$ (see Lemma 7.1). □

7.7.2 Proofs for density estimation (Section 7.3)

Proof of Proposition 7.1. Since the MLE $\widehat{f}_n$ writes $\widehat{f}_n(y) = N_n(y)/n$, we have for every $y \in \mathcal{Y}$:

$$\widehat{f}_n^y(y) = \frac{N_n(y)+1}{n+1} \propto N_n(y) + 1 \,, \tag{7.58}$$

so that, since $\sum_{y\in\mathcal{Y}} N_n(y) = n$,

$$\sum_{y\in\mathcal{Y}} \widehat{f}_n^y(y) = \frac{n+d}{n+1} \,. \tag{7.59}$$

It proves that the SMP $\widetilde{f}_n$ (7.14) is the Laplace estimator (7.16) and that the excess risk bound (7.15) becomes $\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] \leqslant \log \frac{n+d}{n+1} \leqslant \frac{d-1}{n}$ (since $\log(1+u) \leqslant u$ for $u \geqslant 0$). □

Proof of Proposition 7.2. First, let us prove that a risk minimizer $f_{\theta^*,\Sigma} \in \mathcal{F}$ exists if and only if $\mathbb{E}\|Y\| < +\infty$ and that $\theta^* = \mathbb{E}[Y]$ in this case. Let μ be the distribution $\mathcal{N}(0,\Sigma)$, and define the log loss with respect to μ. Then, for every $\theta, y \in \mathbf{R}^d$, $\ell(f_{\theta,\Sigma}, y) = -\langle \Sigma^{-1}\theta, y\rangle + \frac{1}{2}\|\theta\|^2_{\Sigma^{-1}}$. Assume that there exists $\theta^* \in \mathbf{R}^d$ such that $\mathbb{E}[\ell(f_{\theta^*+\theta,\Sigma}, Y) - \ell(f_{\theta^*,\Sigma}, Y)]$ is well-defined and in $[0, +\infty]$ for every $\theta \in \mathbf{R}^d$. This implies that $\mathbb{E}[(\ell(f_{\theta^*+\theta,\Sigma}, Y) - \ell(f_{\theta^*,\Sigma}, Y))_-] < +\infty$, and hence that $\mathbb{E}[(\langle \Sigma^{-1}\theta, Y\rangle)_-] < +\infty$. Taking $\theta = \pm\Sigma e_j$ for $1 \leqslant j \leqslant d$ (where $(e_j)_{1\leqslant j\leqslant d}$ is

the canonical basis of $\mathbf{R}^d$), this implies that $\mathbb{E}|Y_j| < +\infty$, and hence that $\mathbb{E}\|Y\| \leqslant \mathbb{E}\|Y\|_1 = \sum_{j=1}^d \mathbb{E}[|Y_j|] < +\infty$. Conversely, if $\mathbb{E}\|Y\| < +\infty$, so that $\mathbb{E}[Y] \in \mathbf{R}^d$ exists, then for every $\theta \in \mathbf{R}^d$, $R(f_{\theta,\Sigma}) = \mathbb{E}[\ell(f_{\theta,\Sigma}, Y)] = -\langle \Sigma^{-1}\theta, \mathbb{E}[Y]\rangle + \frac{1}{2}\theta^\top \Sigma^{-1}\theta$, which is minimized by $\theta^* = \mathbb{E}[Y]$.

We now proceed to determine the SMP and establish the excess risk bound (7.18). The MLE is $f_{\bar{Y}_n,\Sigma} = \mathcal{N}(\bar{Y}_n, \Sigma)$, so that for $y \in \mathbf{R}^d$, $\widehat{f}_n^y = f_{\widehat{\theta}_n^y,\Sigma}$ with $\widehat{\theta}_n^y = \frac{n\bar{Y}_n + y}{n+1}$. Since $y - \widehat{\theta}_n^y = \frac{n}{n+1}(y - \bar{Y}_n)$, we have, considering densities with respect to the measure $(2\pi)^{-d/2}\mathrm{d}y$:

$$\begin{aligned} f_{\widehat{\theta}_n^y}(y) &= (\det \Sigma)^{-1/2} \exp\Big(-\frac{1}{2}\|y - \widehat{\theta}_n^y\|_{\Sigma^{-1}}^2 \Big) \\ &= (\det \Sigma)^{-1/2} \exp\Big(-\frac{1}{2}\Big(\frac{n}{n+1}\Big)^2 \|y - \bar{Y}_n\|_{\Sigma^{-1}}^2 \Big) \\ &= (\det \Sigma)^{-1/2} \det((1+1/n)^2\Sigma)^{1/2} f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y) \\ &= \Big(1 + \frac{1}{n}\Big)^d f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y)\,, \end{aligned} \tag{7.60}$$

so that (after normalization) $\widetilde{f}_n = \mathcal{N}(\bar{Y}_n, (1+1/n)^2\Sigma)$ and

$$\int_{\mathbf{R}^d} f_{\widehat{\theta}_n^y}(y)(2\pi)^{-d/2}\mathrm{d}y = \int_{\mathbf{R}^d} \Big(1 + \frac{1}{n}\Big)^d f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y)(2\pi)^{-d/2}\mathrm{d}y = \Big(1 + \frac{1}{n}\Big)^d, \tag{7.61}$$

which yields the excess risk bound (7.18) using Theorem 7.2.

Now, assume that the model is well-specified, namely $Y \sim \mathcal{N}(\theta^*, \Sigma)$ for some $\theta^* \in \mathbf{R}^d$. Using Lemma 7.2, we have

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] = \mathbb{E}\Big[\log\Big(\int_{\mathbf{R}^d} f_{\widehat{\theta}_n^y}(y)(2\pi)^{-d/2}\mathrm{d}y\Big)\Big] - \Delta_n = d\log\Big(1 + \frac{1}{n}\Big) - \Delta_n\,,$$

where Δ_n is defined as in (7.50), *i.e.*

$$\begin{aligned} \Delta_n &= \frac{1}{n+1}\mathbb{E}\Big[\sum_{i=1}^{n+1} \ell(f_{\theta^*,\Sigma}, Y_i) - \inf_{\theta\in\mathbf{R}^d} \sum_{i=1}^{n+1} \ell(f_{\theta,\Sigma}, Y_i)\Big] \\ &= \frac{1}{2}\mathbb{E}\Big[\frac{1}{n+1}\sum_{i=1}^{n+1} \|Y_i - \theta^*\|_{\Sigma^{-1}}^2 - \frac{1}{n+1}\sum_{i=1}^{n+1} \|\bar{Y}_{n+1} - Y_i\|_{\Sigma^{-1}}^2\Big] \\ &= \frac{1}{2}\mathbb{E}\big[\|\bar{Y}_{n+1} - \theta^*\|_{\Sigma^{-1}}^2\big] \\ &= \frac{1}{2}\mathrm{Tr}\Big(\Sigma^{-1}\mathbb{E}\big[(\bar{Y}_{n+1} - \theta^*)(\bar{Y}_{n+1} - \theta^*)^\top\big]\Big) \\ &= \frac{1}{2}\mathrm{Tr}\Big(\Sigma^{-1} \times \frac{1}{n+1}\Sigma\Big) = \frac{d}{2(n+1)} \end{aligned}$$

where we used the fact that $\mathbb{E}[(Y-\theta^*)(Y-\theta^*)^\top] = \Sigma$. It follows that $\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] = d\log(1+1/n) - d/(2n) \leqslant d/(2n)$, which completes the proof of Proposition 7.2. □

Proof of Theorem 7.3. Define the densities and the log-loss with respect to the measure $(2\pi)^{-d/2}\mathrm{d}y$ on $\mathbf{R}^d$. For every $\sigma^2 > 0$, $\theta \in \mathbf{R}$ and $y \in \mathbf{R}^d$, we have

$$\ell(f_{\theta,\sigma^2\Sigma}, y) = -\log f_{\theta,\sigma^2\Sigma}(y) = \frac{d}{2}\log\sigma^2 + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2\sigma^2}\|y - \theta\|_{\Sigma^{-1}}^2$$

so that, denoting $\theta^* = \mathbb{E}[Y]$ and $\Sigma_Y := \mathbb{E}[(Y-\theta^*)(Y-\theta^*)^\top]$, we obtain

$$\begin{aligned} R(f_{\theta,\sigma^2\Sigma}) - \frac{1}{2}\log\det(\Sigma) &= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\mathbb{E}\big[\|Y-\theta\|^2_{\Sigma^{-1}}\big] \\ &= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\theta-\theta^*\|^2_{\Sigma^{-1}} + \frac{1}{2\sigma^2}\mathbb{E}\,\mathrm{Tr}\big(\Sigma^{-1}(Y-\theta^*)(Y-\theta^*)^\top\big) \\ &= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\theta-\theta^*\|^2_{\Sigma^{-1}} + \frac{1}{2\sigma^2}\mathrm{Tr}\big(\Sigma^{-1}\Sigma_Y\big) \end{aligned}$$

so that

$$\begin{aligned} \mathcal{E}(f_{\theta,\sigma^2\Sigma}) &= R(f_{\theta,\sigma^2\Sigma}) - R(f_{\theta^*,\Sigma}) \\ &= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\theta-\theta^*\|^2_{\Sigma^{-1}} + \frac{1}{2}\Big(\frac{1}{\sigma^2}-1\Big)\mathrm{Tr}\big(\Sigma^{-1}\Sigma_Y\big)\,. \end{aligned} \tag{7.62}$$

Now, since

$$\mathbb{E}\big[\|\bar{Y}_n-\theta^*\|^2_{\Sigma^{-1}}\big] = \mathrm{Tr}\Big(\Sigma^{-1}\mathbb{E}\big[(\bar{Y}_n-\theta^*)(\bar{Y}_n-\theta^*)^\top\big]\Big) = \frac{\mathrm{Tr}(\Sigma^{-1}\Sigma_Y)}{n}\,,$$

equation (7.62) implies that, for $\sigma^2 = 1+1/n$,

$$\mathbb{E}\big[\mathcal{E}(f_{\bar{Y}_n,\sigma^2\Sigma})\big] = \frac{d}{2}\log\sigma^2 + \frac{1}{2}\Big[\Big(1+\frac{1}{n}\Big)\frac{1}{\sigma^2}-1\Big]\mathrm{Tr}(\Sigma^{-1}\Sigma_Y) = \frac{d}{2}\log\Big(1+\frac{1}{n}\Big)\,. \tag{7.63}$$

In order to conclude that $\widehat{f}_n = \mathcal{N}(\bar{Y}_n, (1+1/n)\Sigma)$, which has constant risk, achieves minimax excess risk over the class of distributions of Y with finite variance, it suffices to note that $\widehat{f}_n$ achieves minimax excess risk for Y a Gaussian from $\{\mathcal{N}(\theta^*,\Sigma) : \theta^* \in \mathbf{R}^d\}$ (*i.e.*, in the well-specified case). Indeed, if $Y \sim \mathcal{N}(\theta^*,\Sigma)$, then $\mathcal{E}(f) = \mathrm{KL}(\mathcal{N}(\theta^*,\Sigma), f)$ for every density f, and $\widehat{g}_n$ achieves minimax KL-risk on the Gaussian location family (Ng, 1980; Murray, 1977). □

7.7.3 Proofs for the Gaussian linear model (Section 7.4)

Proof of Theorem 7.4. Let us first recall that $\mathcal{F} = \{f_\theta(y|x) = \mathcal{N}(\langle\theta,x\rangle, 1) : \theta \in \mathbf{R}^d\}$ and that $\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^n X_iX_i^\top$ and $\widehat{S}_n = n^{-1}\sum_{i=1}^n Y_iX_i$. The MLE is given by $\widehat{\theta}_n = \widehat{\Sigma}_n^{-1}\widehat{S}_n$ and, for every $x \in \mathbf{R}^d$ and $y \in \mathbf{R}$,

$$\widehat{\theta}_n^{(x,y)} = (n\widehat{\Sigma}_n + xx^\top)^{-1}(n\widehat{S}_n + yx).$$

Hence, we have

$$\begin{aligned} y - \langle\widehat{\theta}_n^{(x,y)}, x\rangle &= y - \big\langle(n\widehat{\Sigma}_n + xx^\top)^{-1}(n\widehat{S}_n + yx), x\big\rangle \\ &= \big(1 - \big\langle(n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle\big)y - \big\langle(n\widehat{\Sigma}_n + xx^\top)^{-1}n\widehat{S}_n, x\big\rangle \\ &= \sigma_n(x)^{-1}(y - \mu_n(x)), \end{aligned}$$

where we defined

$$\sigma_n(x) = \big(1 - \big\langle(n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle\big)^{-1} \quad \text{and} \quad \mu_n(x) = \frac{\big\langle(n\widehat{\Sigma}_n + xx^\top)^{-1}n\widehat{S}_n, x\big\rangle}{1 - \big\langle(n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle}.$$

Note that both quantities are well-defined under since $\widehat{\Sigma}_n$ is invertible almost surely by Assumption 7.2. Moreover, these quantities can be simplified thanks to the following lemma.

Lemma 7.3. *Assume that S is a symmetric positive d-dimensional matrix and that $v \in \mathbf{R}^d$. Then, one has*

$$\big(1 - \langle (S + vv^\top)^{-1} v, v\rangle\big)^{-1} = 1 + \langle S^{-1} v, v\rangle, \tag{7.64}$$

and, for any $u \in \mathbf{R}^d$,

$$\frac{\big\langle (S + vv^\top)^{-1} S u, v\big\rangle}{1 - \big\langle (S + vv^\top)^{-1} v, v\big\rangle} = \langle u, v\rangle\,. \tag{7.65}$$

The proof of Lemma 7.3 is given below. It also follows from the Sherman-Morrison formula. Using (7.64) with $S = n\widehat{\Sigma}_n$ and $v = x$ leads to

$$\sigma_n(x) = 1 + \big\langle (n\widehat{\Sigma}_n)^{-1} x, x\big\rangle$$

while the fact that $\widehat{S}_n = \widehat{\Sigma}_n \widehat{\theta}_n$ together with (7.65) for $S = n\widehat{\Sigma}_n$, $v = x$ and $u = \widehat{\theta}_n$ leads to

$$\mu_n(x) = \frac{\big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} n\widehat{S}_n, x\big\rangle}{1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} x, x\big\rangle} = \frac{\big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} n\widehat{\Sigma}_n \widehat{\theta}_n, x\big\rangle}{1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} x, x\big\rangle} = \langle \widehat{\theta}_n, x\rangle.$$

Consider the dominating measure $\mu(\mathrm{d}y) = (2\pi)^{-1/2}\mathrm{d}y$ on $\mathbf{R}$. The computations above entail that for every $y \in \mathbf{R}$, we have

$$f_{\widehat{\theta}_n^{(x,y)}}(y|x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(y - \langle \widehat{\theta}_n^{(x,y)}, x\rangle\big)^2\Big) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2\sigma_n^2(x)}\big(y - \mu_n(x)\big)^2\Big).$$

Note that

$$\int_{\mathbf{R}} f_{\widehat{\theta}_n^{(x,y)}}(y|x)\mu(\mathrm{d}y) = \sigma_n(x),$$

which shows after normalization (7.14) that the SMP is given by

$$\widetilde{f}_n(y|x) = \mathcal{N}(\mu_n(x), \sigma_n^2(x)) \tag{7.66}$$

and that its excess risk writes

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\big[\log \sigma_n(X)\big] = \mathbb{E}\Big[-\log\Big(1 - \big\langle (n\widehat{\Sigma}_n + XX^\top)^{-1} X, X\big\rangle\Big)\Big]\,. \tag{7.67}$$

This proves the first inequality in (7.26). Let us prove now the second inequality in (7.26). Let us recall that the covariance Σ and rescaled design $\widetilde{X}, \widetilde{X}_i$ and rescaled covariance $\widetilde{\Sigma}_n$ are given by (7.22) and (7.24). We have

$$\begin{aligned}\big\langle (n\widehat{\Sigma}_n + XX^\top)^{-1} X, X\big\rangle &= \big\langle \Sigma^{1/2}(n\widehat{\Sigma}_n + XX^\top)^{-1}\Sigma^{1/2}\Sigma^{-1/2}X, \Sigma^{-1/2}X\big\rangle \\ &= \big\langle (n\widetilde{\Sigma}_n + \widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}, \widetilde{X}\big\rangle,\end{aligned} \tag{7.68}$$

hence, combining (7.67), (7.68) and (7.64), we have

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\Big[-\log\Big(1 - \big\langle (n\widetilde{\Sigma}_n + \widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}, \widetilde{X}\big\rangle\Big)\Big] = \mathbb{E}\Big[\log\Big(1 + \big\langle (n\widetilde{\Sigma}_n)^{-1}\widetilde{X}, \widetilde{X}\big\rangle\Big)\Big],$$

which leads, using Jensen's inequality, together with $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = I_d$ and the fact that $\widetilde{\Sigma}_n$ and $\widetilde{X}$ are independent, to

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \log\Big(1 + \frac{1}{n}\mathbb{E}\big[\mathrm{Tr}\big(\widetilde{\Sigma}_n^{-1}\widetilde{X}\widetilde{X}^\top\big)\big]\Big) &= \log\Big(1 + \frac{1}{n}\mathrm{Tr}\big\{\mathbb{E}[\widetilde{\Sigma}_n^{-1}]\mathbb{E}[\widetilde{X}\widetilde{X}^\top]\big\}\Big) \\ &= \log\Big(1 + \frac{1}{n}\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]\Big).\end{aligned}$$

This concludes the proof of Theorem 7.4. □

Proof of Lemma 7.3. First, (7.65) clearly holds if $v = 0$. Now, for $u, v \in \mathbf{R}^d$, $v \neq 0$:

$$\begin{aligned}\big\langle (S + vv^\top)^{-1} S u, v \big\rangle &= \big\langle (S + vv^\top)^{-1}(S + vv^\top - vv^\top)u, v \big\rangle \\ &= \big\langle (I_d - (S + vv^\top)^{-1} vv^\top)u, v \big\rangle \\ &= \langle u, v \rangle \big(1 - \langle (S + vv^\top)^{-1} v, v \rangle\big)\,. \end{aligned} \tag{7.69}$$

Letting $u = S^{-1}v$ in (7.69), the left-hand side is $\langle (S + vv^\top)^{-1} v, v\rangle > 0$ (since $S + vv^\top \succcurlyeq S$ is positive, and $v \neq 0$) so that the right-hand side is positive and thus $1 - \big\langle (S+vv^\top)^{-1} v, v \big\rangle > 0$. Dividing both sides of (7.69) by this quantity establishes (7.65), which implies (7.64) by taking $u = S^{-1}v$. □

Proof of Theorem 7.5 and Proposition 7.3. Let us recall that we consider the family $\mathcal{F} = \{f_\theta(\cdot|x) = \mathcal{N}(\langle \theta, x\rangle, \sigma^2) : \theta \in \mathbf{R}^d\}$, together with the Ridge penalization $\phi(\theta) = \lambda\|\theta\|^2/2$ for some $\lambda > 0$. Let

$$\widehat{\theta}_{\lambda,n} := \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(f_\theta, (X_i, Y_i)) + \frac{\lambda}{2}\|\theta\|^2 \right\} = (\widehat{\Sigma}_n + \lambda I_d)^{-1} \widehat{S}_n,$$

denote the Ridge estimator, where $\widehat{\Sigma}_n$ and $\widehat{S}_n$ are the same as in the proof of Theorem 7.4. Defining

$$\widehat{\Sigma}^x_\lambda = n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d \quad \text{and} \quad \widehat{K}^x_\lambda = (\widehat{\Sigma}^x_\lambda)^{-1},$$

we have

$$\widehat{\theta}^{(x,y)}_{\lambda,n} = \big(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\big)^{-1}(n\widehat{S}_n + yx) = \widehat{K}^x_\lambda(n\widehat{S}_n + yx)$$

for any $y \in \mathbf{R}$ and $x \in \mathbf{R}^d$. Note that we have

$$y - \langle \widehat{\theta}^{(x,y)}_{\lambda,n}, x\rangle = y - \big\langle \widehat{K}^x_\lambda(n\widehat{S}_n + yx), x\big\rangle = \big(1 - \|x\|^2_{\widehat{K}^x_\lambda}\big)y - \langle n\widehat{S}_n, x\rangle_{\widehat{K}^x_\lambda} \tag{7.70}$$

and that

$$\begin{aligned}\lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2 &= \lambda\|\widehat{K}^x_\lambda(n\widehat{S}_n + yx)\|^2 = \lambda\|n\widehat{S}_n + yx\|^2_{(\widehat{K}^x_\lambda)^2} \\ &= y^2\lambda\|x\|^2_{(\widehat{K}^x_\lambda)^2} + 2y\lambda\langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^2} + \lambda\|n\widehat{S}_n\|^2_{(\widehat{K}^x_\lambda)^2}\,.\end{aligned}$$

The SMP is given in this setting by

$$\widetilde{f}_{\lambda,n}(y|x) = \frac{f_{\widehat{\theta}^{(x,y)}_{\lambda,n}}(y|x) e^{-\lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2/2}}{\int_{\mathbf{R}} f_{\widehat{\theta}^{(x,y')}_{\lambda,n}}(y'|x) e^{-\lambda\|\widehat{\theta}^{(x,y')}_{\lambda,n}\|^2/2}\mu(\mathrm{d}y')},$$

where $\mu(\mathrm{d}y) = (2\pi)^{-1/2}\mathrm{d}y$, see (7.14), and where

$$f_{\widehat{\theta}^{(x,y)}_{\lambda,n}}(y|x) e^{-\lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2/2} = \exp\left(-\frac{1}{2}\Big\{ \big(y - \langle \widehat{\theta}^{(x,y)}_{\lambda,n}, x\rangle\big)^2 + \lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2 \Big\}\right).$$

Now, the equality (7.70) gives, after a straightforward computation,

$$\big(y - \langle \widehat{\theta}^{(x,y)}_{\lambda,n}, x\rangle\big)^2 + \lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2 = \frac{1}{\sigma_\lambda(x)^2}\big(y - \mu_\lambda(x)\big)^2 + C,$$

where C is a quantity that does not depend on y and where we introduced, respectively,

$$\sigma_\lambda(x)^2 = \Big((1-\|x\|^2_{\widehat{K}^x_\lambda})^2 + \lambda\|x\|^2_{(\widehat{K}^x_\lambda)^2}\Big)^{-1}$$
$$\mu_\lambda(x) = \frac{(1-\|x\|^2_{\widehat{K}^x_\lambda})\langle n\widehat{S}_n, x\rangle_{\widehat{K}^x_\lambda} - \lambda\langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^2}}{(1-\|x\|^2_{\widehat{K}^x_\lambda})^2 + \lambda\|x\|^2_{(\widehat{K}^x_\lambda)^2}}\,.$$

This entails that the SMP is given by

$$\widetilde{f}_{\lambda,n}(\cdot|x) = \mathcal{N}\big(\mu_\lambda(x), \sigma_\lambda(x)^2\big)\,. \tag{7.71}$$

By definition of $\widehat{\theta}_{\lambda,n}$ we have

$$n\widehat{S}_n = \big(n\widehat{\Sigma}_n + \lambda(n+1)I_d\big)\widehat{\theta}_{\lambda',n}$$

where $\lambda' = (n+1)\lambda/n$, so that for $\alpha \in \{1,2\}$ we have

$$\begin{aligned}\langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^\alpha} &= \big\langle \big(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\big)^\alpha n\widehat{S}_n, x\big\rangle \\ &= \big\langle \big(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\big)^\alpha (n\widehat{\Sigma}_n + \lambda(n+1)I_d + xx^\top - xx^\top)\widehat{\theta}_{\lambda',n}, x\big\rangle \\ &= \langle \widehat{\theta}_{\lambda',n}, x\rangle_{(\widehat{K}^x_\lambda)^{\alpha-1}} - \langle \widehat{\theta}_{\lambda',n}, x\rangle \, \|x\|^2_{(\widehat{K}^x_\lambda)^\alpha},\end{aligned}$$

namely

$$\langle n\widehat{S}_n, x\rangle_{\widehat{K}^x_\lambda} = \big(1-\|x\|^2_{\widehat{K}^x_\lambda}\big)\langle \widehat{\theta}_{\lambda',n}, x\rangle \quad \text{and} \quad \langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^2} = \langle \widehat{\theta}_{\lambda',n}, x\rangle_{\widehat{K}^x_\lambda} - \langle \widehat{\theta}_{\lambda',n}, x\rangle \, \|x\|^2_{(\widehat{K}^x_\lambda)^2}.$$

This allows, after straightforward computations, to express $\mu_\lambda(x)$ as a function of $\widehat{\theta}_{\lambda',n}$ as follows:

$$\mu_\lambda(x) = \langle \widehat{\theta}_{\lambda',n}, x\rangle - \lambda\sigma_\lambda(x)^2\langle \widehat{\theta}_{\lambda',n}, x\rangle_{\widehat{K}^x_\lambda}.$$

We know from Theorem 7.2 that the penalized excess risk of SMP satisfies

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] &\leqslant \mathbb{E}_{Z_1^n,X}\Big[\log\Big(\int_{\mathbf{R}} f_{\widehat{\theta}^{(X,y)}_{\lambda,n}}(y|X)e^{-\lambda\|\widehat{\theta}^{(X,y)}_{\lambda,n}\|^2/2}\mu(\mathrm{d}y)\Big)\Big] \\ &\leqslant \mathbb{E}_{Z_1^n,X}\Big[\log\Big(\int_{\mathbf{R}} f_{\widehat{\theta}^{(X,y)}_{\lambda,n}}(y|X)\mu(\mathrm{d}y)\Big)\Big].\end{aligned}$$

We know from the computations above that

$$\big(y - \langle \widehat{\theta}^{(x,y)}_{\lambda,n}, x\rangle\big)^2 = \big(1-\|x\|^2_{\widehat{K}^x_\lambda}\big)^2\big(y - \langle \widehat{\theta}_{\lambda',n}, x\rangle\big)^2,$$

so that, after integrating with respect to y,

$$\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] \leqslant \mathbb{E}_{X_1^n,X}\left[\log\Big(\frac{1}{1-\|X\|^2_{\widehat{K}^X_\lambda}}\Big)\right] = \mathbb{E}_{X_1^n,X}\Big[-\log\big(1-\langle(\widehat{\Sigma}^X_\lambda)^{-1}X, X\rangle\big)\Big]. \tag{7.72}$$

Note that, by the identity (7.64) from Lemma 7.3, and since $\|X\| \leqslant R$ almost surely (Assumption 7.5) we have

$$\langle(\widehat{\Sigma}^X_\lambda)^{-1}X, X\rangle = \frac{\langle(n\widehat{\Sigma}_n + \lambda(n+1)I_d)^{-1}X, X\rangle}{1+\langle(n\widehat{\Sigma}_n + \lambda(n+1)I_d)^{-1}X, X\rangle} \leqslant \frac{R^2/\big(\lambda(n+1)\big)}{1+R^2/\big(\lambda(n+1)\big)}\,. \tag{7.73}$$

In addition, the function $g(u) = -\log(1-u)/u$ defined on $(0,1)$ is nondecreasing, since its derivative writes:

$$g'(u) = \frac{1}{u^2}\left[\frac{u}{1-u} - \log\left(1+\frac{u}{1-u}\right)\right] \geqslant 0\,,$$

where we used the inequality $\log(1+v) \leqslant v$ for $v \geqslant 0$. Combining this fact with (7.73) shows that

$$-\log\big(1-\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big) \leqslant g\bigg(\frac{R^2/\big(\lambda(n+1)\big)}{1+R^2/\big(\lambda(n+1)\big)}\bigg)\cdot\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\,. \tag{7.74}$$

Next, by exchangeability of $(X_1,\dots,X_n,X)$, we have

$$\begin{aligned}\mathbb{E}\big[\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big] &= \frac{1}{n+1}\,\mathbb{E}\bigg[\sum_{i=1}^n\langle(\widehat{\Sigma}_\lambda^X)^{-1}X_i, X_i\rangle + \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\bigg]\\ &= \frac{1}{n+1}\,\mathbb{E}\bigg[\mathrm{Tr}\bigg\{\bigg(\sum_{i=1}^n X_iX_i^\top + XX^\top + \lambda(n+1)I_d\bigg)^{-1}\bigg(\sum_{i=1}^n X_iX_i^\top + XX^\top\bigg)\bigg\}\bigg]. \end{aligned}\tag{7.75}$$

In addition, the function $A \mapsto \mathrm{Tr}((A+I_d)^{-1}A)$ is concave on positive matrices. Indeed, it writes $d - \mathrm{Tr}[(A+I_d)^{-1}]$, and $A \mapsto \mathrm{Tr}(A^{-1})$ is convex on positive matrices since $x \mapsto x^{-1}$ is convex on $\mathbf{R}_+^*$, by a general result on the convexity of trace functionals, see e.g. Bhatia (2009); Boyd and Vandenberghe (2004). Hence, applying Jensen's inequality to (7.75) and using the fact that

$$\mathbb{E}\bigg[\sum_{i=1}^n X_iX_i^\top + XX^\top\bigg] = (n+1)\Sigma\,,$$

we obtain:

$$\mathbb{E}\big[\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big] \leqslant \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}\,. \tag{7.76}$$

Finally, combining the bounds (7.72), (7.74) and (7.76) yields:

$$\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] \leqslant g\bigg(\frac{R^2/\big(\lambda(n+1)\big)}{1+R^2/\big(\lambda(n+1)\big)}\bigg)\cdot\frac{\mathsf{df}_\lambda(\Sigma)}{n+1}\,. \tag{7.77}$$

Nonparametric rates (Theorem 7.5). Assume that $\lambda(n+1) \geqslant 2R^2$. The quantity inside $g(\cdot)$ in (7.77) is then bounded by $(1/2)/(1+1/2) = 1/3$, and since $g(1/3) = 3\log(3/2) \leqslant 1.25$, (7.77) becomes, by definition of $\mathcal{E}_\lambda$:

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta\in\mathbf{R}^d}\bigg\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\bigg\} \leqslant 1.25\cdot\frac{\mathsf{df}_\lambda(\Sigma)}{n+1}\,. \tag{7.78}$$

which is precisely the announced bound (7.33).

Finite-dimensional case: improved dependence on the norm (Proposition 7.3). Now, let $\lambda = d/\big(B^2(n+1)\big)$ for some $B > 0$ (which will be a bound on the norm of the comparison parameter θ). Then, $R^2/\big(\lambda(n+1)\big) = B^2R^2/d$. Now, note that for every $v > 0$

$$g\bigg(\frac{v}{1+v}\bigg) = \frac{-\log\big(1-v/(1+v)\big)}{v/(1+v)} = \frac{(1+v)\log(1+v)}{v}\,.$$

In addition, if $v \leqslant 1$, then $(1+v)\log(1+v)/v \leqslant 1+v \leqslant 2$. On the other hand, if $v \geqslant 1$, then $(1+v)/v \leqslant 2$; it follows that for every $v > 0$:

$$g\left(\frac{v}{1+v}\right) \leqslant 2\log(e+v) \leqslant 2\log(4+4\sqrt{v}+v) = 4\log(2+\sqrt{v})\,. \tag{7.79}$$

Now, the excess risk bound (7.77) implies that, for every $\theta \in \mathbf{R}^d$ such that $\|\theta\| \leqslant B$,

$$\begin{aligned}
\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - R(f_\theta) &\leqslant g\left(\frac{B^2R^2/d}{1+B^2R^2/d}\right)\cdot\frac{\mathsf{df}_\lambda(\Sigma)}{n+1} + \frac{\lambda}{2}\|\theta\|^2 \\
&\leqslant 4\log\left(2+\frac{BR}{\sqrt{d}}\right)\times\frac{d}{n+1} + \frac{d}{B^2(n+1)}\times\frac{B^2}{2} \qquad (7.80)\\
&= \frac{d}{n+1}\left\{4\log\left(2+\frac{BR}{\sqrt{d}}\right)+\frac{1}{2}\right\} \\
&\leqslant \frac{5d\log\left(2+BR/\sqrt{d}\right)}{n+1} \qquad (7.81)
\end{aligned}$$

where inequality (7.80) uses the bound (7.79) with $v = B^2R^2/d$, the bound $\mathsf{df}_\lambda(\Sigma) \leqslant d$ (7.30) and the fact that $\|\theta\| \leqslant B$, while inequality (7.81) uses the fact that $1/2 \leqslant \log 2$. □

7.7.4 Proofs for logistic regression (Section 7.5)

Proof of Proposition 7.4. Let us first discuss the properties of predictions produced by the SMP, and compare it to the MLE. First, if the points $Z_1, \ldots, Z_n$ do not lie within a half-space, the MLE is uniquely determined and belongs to $\mathbf{R}^d$; in addition, for any $x \in \mathbf{R}^d$ and $y \in \{-1, 1\}$, $Z_1, \ldots, Z_n, -yx$ are not separated either, so $\widehat{\theta}_n^{(x,y)} \in \mathbf{R}^d$ is also well-defined and unique, and so is the prediction $\widetilde{f}_n(1|x) \in (0,1)$.

Let $\Lambda_n = \{\sum_{1\leqslant i\leqslant n}\lambda_i Z_i : \lambda_i \in \mathbf{R}^+, 1 \leqslant i \leqslant n\}$ denote the convex cone generated by $Z_1, \ldots, Z_n$. Assume that $\Lambda_n \cap (-\Lambda_n) = \{0\}$ and that all Z_i are distinct from 0. Then, convex separation implies that there exists $\theta \in \mathbf{R}^d$ such that $\langle\theta, z\rangle < 0$ for all $z \in \Lambda_n \setminus \{0\}$, so that the Z_i lie within a strict half-space: $\langle\theta, Z_i\rangle < 0$ for all i. Hence, any MLE $f_{\widehat{\theta}_n}$ in $\overline{\mathcal{F}}$ belongs to $\overline{\mathcal{F}} \setminus \mathcal{F}$, and corresponds to a separating hyperplane $(+\infty, \widehat{\theta}_n)$ for some $\widehat{\theta}_n \in S^{d-1}$ (such that $\langle\widehat{\theta}_n, z\rangle < 0$ for all $z \in \Lambda_n \setminus \{0\}$). Its predictions $f_{\widehat{\theta}_n}(1|x)$ are as follows:

- If $x = 0$, then $f_{\widehat{\theta}_n}(1|x) = 1/2$.
- If $x \in \Lambda_n \setminus \{0\}$, then $\langle\widehat{\theta}_n, x\rangle < 0$ and thus $f_{\widehat{\theta}_n}(1|x) = 0$. Likewise, if $x \in (-\Lambda_n) \setminus \{0\}$, then $f_{\widehat{\theta}_n}(1|x) = 1$;
- If $x \in \mathbf{R}^d \setminus [\Lambda_n \cup (-\Lambda_n)]$, then both x and $-x$ are linearly separated from Λ_n. Hence, one can choose $\widehat{\theta}_n$ with $\langle\widehat{\theta}_n, z\rangle < 0$ for $z \in \Lambda_n \setminus \{0\}$ such that either $\langle\widehat{\theta}_n, x\rangle > 0$ or $\langle\widehat{\theta}_n, x\rangle < 0$ (or even $\langle\widehat{\theta}_n, x\rangle = 0$). In other words, one can choose an MLE $\widehat{\theta}_n$ such that $f_{\widehat{\theta}_n}(1|x)$ is either 1, 0 or 1/2: the prediction of the MLE is ill-determined in this region, since it depends on the specific choice of the MLE.

By contrast, let us consider the prediction of the SMP $\widetilde{f}_n$. Let $z = -yx \in \mathbf{R}^d \setminus \{0\}$. As before, if $z \in \mathbf{R}^d \setminus (-\Lambda_n)$, then there exists θ with $\langle\theta, z\rangle < 0$ and $\langle\theta, Z_i\rangle = -\langle\theta, -Z_i\rangle < 0$. Hence, $f_{\widehat{\theta}_n^{(x,y)}}(y|x) = 1$. On the other hand, if $z \in (-\Lambda_n) \setminus \{0\}$, then the dataset $Z_1, \ldots, Z_n, z$ is not separated, so that $f_{\widehat{\theta}_n^{(x,y)}}(y|x) \in (0,1)$. Hence, for $x \in \mathbf{R}^d$:

- If $x = 0$, then $\widetilde{f}_n(1|x) = 1/2$.
- If $x \in \Lambda_n$, then $-x \in (-\Lambda_n)$ so that $f_{\widehat{\theta}_n^{(x,1)}}(1|x) \in (0,1)$, while $x \in \mathbf{R}^d \setminus (-\Lambda_n)$ so that $f_{\widehat{\theta}_n^{(x,-1)}}(-1|x) = 1$; hence, $\widetilde{f}_n(1|x) \in (0, 1/2)$. Likewise, if $x \in (-\Lambda_n)$, then $\widetilde{f}_n(1|x) \in (1/2, 1)$.
- If $x \in \mathbf{R}^d \setminus [\Lambda_n \cup (-\Lambda_n)]$, then $f_{\widehat{\theta}_n^{(x,1)}}(1|x) = f_{\widehat{\theta}_n^{(x,-1)}}(-1|x) = 1$, so that $\widetilde{f}_n(1|x) = 1/2$.

Finally, the excess risk bound (7.45) is established in the proof of Theorem 7.5 below, letting $\lambda = 0$. □

Proof of Theorem 7.6. Let (X, Y) be a test sample, and $Z = -YX$. Since $\{Z, -Z\} = \{X, -X\}$, the excess risk bound (7.15) of the SMP $\widetilde{f}_{\lambda,n}$ (7.47) writes:

$$\begin{aligned}
&\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta\in\mathbf{R}^d}\Big\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\Big\} \\
&\leqslant \mathbb{E}\left[\log\left(\sigma(\langle\widehat{\theta}_{\lambda,n}^{(X,1)}, X\rangle)\, e^{-\lambda\|\widehat{\theta}_{\lambda,n}^{(X,1)}\|^2/2} + \sigma(-\langle\widehat{\theta}_{\lambda,n}^{(X,-1)}, X\rangle)\, e^{-\lambda\|\widehat{\theta}_{\lambda,n}^{(X,-1)}\|^2/2}\right)\right] \\
&= \mathbb{E}\left[\log\left(\sigma(\langle\widehat{\theta}_{\lambda,n}^{-Z}, Z\rangle)\, e^{-\lambda\|\widehat{\theta}_{\lambda,n}^{-Z}\|^2/2} + \sigma(-\langle\widehat{\theta}_{\lambda,n}^{Z}, Z\rangle)\, e^{-\lambda\|\widehat{\theta}_{\lambda,n}^{Z}\|^2/2}\right)\right] \\
&\leqslant \mathbb{E}\left[\log\left(1 + \sigma(\langle\widehat{\theta}_{\lambda,n}^{-Z}, Z\rangle) - \sigma(\langle\widehat{\theta}_{\lambda,n}^{Z}, Z\rangle)\right)\right] \qquad (7.82) \\
&\leqslant \mathbb{E}\left[\sigma(\langle\widehat{\theta}_{\lambda,n}^{-Z}, Z\rangle) - \sigma(\langle\widehat{\theta}_{\lambda,n}^{Z}, Z\rangle)\right] \qquad (7.83)
\end{aligned}$$

where inequality (7.82) is obtained by lower-bounding $e^{-\lambda\|\cdot\|^2/2} \leqslant 1$ and using the identity $\sigma(-u) = 1 - \sigma(u)$. Now, defining for $\theta \in \mathbf{R}^d$

$$\widehat{R}_{\lambda,n}^{Z}(\theta) := \frac{1}{n+1}\left\{\sum_{i=1}^{n}\ell(\langle\theta, Z_i\rangle) + \ell(\langle\theta, Z\rangle)\right\} + \frac{\lambda}{2}\|\theta\|^2\,,$$

we have, respectively,

$$\widehat{\theta}_{\lambda,n}^{Z} = \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d}\widehat{R}_{\lambda,n}^{Z}(\theta) \qquad (7.84)$$

$$\widehat{\theta}_{\lambda,n}^{-Z} = \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d}\left\{\widehat{R}_{\lambda,n}^{Z}(\theta) - \frac{1}{n+1}\langle\theta, Z\rangle\right\}, \qquad (7.85)$$

where (7.85) comes from the fact that $\ell(-u) = \ell(u) - u$ for $u \in \mathbf{R}$.

Now, the function $\widehat{R}_n^Z$ is λ-strongly convex, as the sum of a convex function (recall that ℓ is convex since $\ell'' = \sigma(1-\sigma) \geqslant 0$) and a $\lambda\|\theta\|^2/2$ term. It follows from Lemma 7.4 that

$$R\cdot\big\|\widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^{Z}\big\| \leqslant R\cdot\frac{\|Z/(n+1)\|}{\lambda} \leqslant \frac{R^2}{\lambda(n+1)} \leqslant \frac{1}{2}\,, \qquad (7.86)$$

where we used the assumption that $\lambda \geqslant 2R^2/(n+1)$. In addition, still by Lemma 7.4,

$$0 \leqslant \langle\widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^{Z}, Z\rangle \leqslant 1/2\,. \qquad (7.87)$$

Now, since $(\log\sigma')' = \sigma''/\sigma' = 1 - 2\sigma \leqslant 1$, we have for every $u \in \mathbf{R}$ and $v \in [0, 1/2]$, $\log\sigma'(u+v) - \log\sigma'(u) \leqslant v$, namely $\sigma'(u+v) \leqslant e^v\sigma'(u) \leqslant e\cdot\sigma'(u)$. Hence, $\sigma(u+v) \leqslant e\cdot\sigma'(u)\cdot v$

for every $u \in \mathbf{R}$ and $v \in [0, 1/2]$. By (7.87), applying this inequality to $u = \langle \widehat{\theta}^Z_{\lambda,n}, Z\rangle$ and $v = \langle \widehat{\theta}^{-Z}_{\lambda,n} - \widehat{\theta}^Z_{\lambda,n}, Z\rangle$ yields:

$$\sigma\big(\langle \widehat{\theta}^{-Z}_{\lambda,n}, Z\rangle\big) - \sigma\big(\langle \widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) \leqslant e^{1/2} \cdot \sigma'\big(\langle \widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) \cdot \langle \widehat{\theta}^{-Z}_{\lambda,n} - \widehat{\theta}^Z_{\lambda,n}, Z\rangle\,. \tag{7.88}$$

Let us now consider the function $\widehat{R}^Z_{\lambda,n}$; its third derivative can be controlled in terms of its Hessian, as shown by Bach (2010). Fix $\theta, \theta \in \mathbf{R}^d$, and define the function $g(t) = \widehat{R}^Z_{\lambda,n}(\theta + t\theta)$ for $t \in \mathbf{R}$. We have respectively, denoting $\theta_t = \theta + t\theta$,

$$g''(t) = \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\theta_t)\theta, \theta\rangle = \frac{1}{n+1}\left\{\sum_{i=1}^n \sigma'(\langle \theta_t, Z_i\rangle)\langle \theta, Z_i\rangle^2 + \sigma'(\langle \theta_t, Z\rangle)\langle \theta, Z\rangle^2\right\} + \lambda\|\theta\|^2$$
$$g'''(t) = \nabla^3 \widehat{R}^Z_{\lambda,n}(\theta_t)[\theta,\theta,\theta] = \frac{1}{n+1}\left\{\sum_{i=1}^n \sigma''(\langle \theta_t, Z_i\rangle)\langle \theta, Z_i\rangle^3 + \sigma''(\langle \theta_t, Z\rangle)\langle \theta, Z\rangle^3\right\}$$

Now, since $|\sigma''| = |\sigma(1-\sigma)(1-2\sigma)| \leqslant \sigma(1-\sigma) = \sigma'$ (as $0 \leqslant \sigma \leqslant 1$), and since by the Cauchy-Schwarz inequality $|\langle \theta, Z_i\rangle| \leqslant R\|\theta\|$ $(1 \leqslant i \leqslant n)$ and $|\langle \theta, Z\rangle| \leqslant R\|\theta\|$, we have

$$\begin{aligned}|g'''(t)| &= \frac{1}{n+1}\left\{\sum_{i=1}^n \big|\sigma''(\langle \theta_t, Z_i\rangle)\langle \theta, Z_i\rangle^3\big| + \big|\sigma''(\langle \theta_t, Z\rangle)\langle \theta, Z\rangle^3\big|\right\} \\ &\leqslant R\|\theta\| \cdot \frac{1}{n+1}\left\{\sum_{i=1}^n \sigma'(\langle \theta_t, Z_i\rangle)\langle \theta, Z_i\rangle^2 + \sigma'(\langle \theta_t, Z\rangle)\langle \theta, Z\rangle^2\right\} \leqslant R\|\theta\| \cdot g''(t)\,. \end{aligned} \tag{7.89}$$

The property (7.89) is the pseudo-self-concordance condition introduced by Bach (2010); in particular, by Proposition 1 therein, we have for every $\theta, \theta \in \mathbf{R}^d$:

$$\nabla^2 \widehat{R}^Z_{\lambda,n}(\theta + \theta) \succcurlyeq e^{-R\|\theta\|} \cdot \nabla^2 \widehat{R}^Z_{\lambda,n}(\theta)\,. \tag{7.90}$$

It follows from (7.90) (letting $\theta = \widehat{\theta}^Z_{\lambda,n}$ and $\theta = \theta' - \widehat{\theta}^Z_{\lambda,n}$) that $\widehat{R}^Z_{\lambda,n}$ is $e^{-(1/2+\varepsilon)}\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})$-strongly convex on the open convex ball $\Omega_\varepsilon = \{\theta' \in \mathbf{R}^d : R\|\theta' - \widehat{\theta}^Z_{\lambda,n}\| < 1/2 + \varepsilon\}$ for every $\varepsilon > 0$. In addition, the inequality (7.86) shows that the function $\widehat{R}^Z_{\lambda,n}(\theta) - \langle \theta, Z\rangle/(n+1)$ reaches its minimum $\widehat{\theta}^{-Z}_{\lambda,n}$ on Ω_ε, so that by Lemma 7.4,

$$\langle \widehat{\theta}^{-Z}_{\lambda,n} - \widehat{\theta}^Z_{\lambda,n}, Z/(n+1)\rangle \leqslant e^{1/2+\varepsilon}\left\|\frac{Z}{n+1}\right\|^2_{\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1}}\,.$$

Taking $\varepsilon \to 0$ in the above bound and multiplying by $n+1$, we obtain:

$$\langle \widehat{\theta}^{-Z}_{\lambda,n} - \widehat{\theta}^Z_{\lambda,n}, Z\rangle \leqslant \frac{e^{1/2}}{n+1} \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1} Z, Z\rangle\,, \tag{7.91}$$

so that by combining inequalities (7.88) and (7.91),

$$\sigma\big(\langle \widehat{\theta}^{-Z}_{\lambda,n}, Z\rangle\big) - \sigma\big(\langle \widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) \leqslant \frac{e}{n+1} \cdot \sigma'\big(\langle \widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1} Z, Z\rangle\,. \tag{7.92}$$

It thus remains to control the expectation of the right-hand side of (7.92). By exchangeability of $(Z_1, \dots, Z_n, Z)$ (and since $\widehat{R}^Z_{\lambda,n}, \widehat{\theta}^Z_{\lambda,n}$ are unchanged after permutation of Z_i and Z), we have:

$$\begin{aligned}
&\mathbb{E}\big[\sigma'\big(\langle\widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) \cdot \langle\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1} Z, Z\rangle\big] \\
&= \frac{1}{n+1}\mathbb{E}\bigg[\sum_{i=1}^n \sigma'\big(\langle\widehat{\theta}^Z_{\lambda,n}, Z_i\rangle\big) \cdot \langle\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1} Z_i, Z_i\rangle + \sigma'\big(\langle\widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) \cdot \langle\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1} Z, Z\rangle\bigg] \\
&= \mathbb{E}\bigg[\mathrm{Tr}\bigg\{\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\theta}^Z_{\lambda,n})^{-1} \cdot \frac{1}{n+1}\bigg(\sum_{i=1}^n \sigma'\big(\langle\widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) Z_i Z_i^\top + \sigma'\big(\langle\widehat{\theta}^Z_{\lambda,n}, Z\rangle\big) Z Z^\top\bigg)\bigg\}\bigg] \\
&= \mathbb{E}\left[\mathrm{Tr}\left\{\big[\nabla^2 \widehat{R}^Z_n(\widehat{\theta}^Z_{\lambda,n}) + \lambda I_d\big]^{-1} \nabla^2 \widehat{R}^Z_n(\widehat{\theta}^Z_{\lambda,n})\right\}\right] ; \qquad (7.93)
\end{aligned}$$

in (7.93), we defined

$$\widehat{R}^Z_n(\theta) = \widehat{R}^Z_{\lambda,n}(\theta) - \frac{\lambda}{2}\|\theta\|^2 = \frac{1}{n+1}\left\{\sum_{i=1}^n \ell(\langle\theta, Z_i\rangle) + \ell(\langle\theta, Z\rangle)\right\},$$

whose Hessian writes

$$\nabla^2 \widehat{R}^Z_n(\theta) = \frac{1}{n+1}\left\{\sum_{i=1}^n \sigma'(\langle\theta, Z_i\rangle) Z_i Z_i^\top + \sigma'(\langle\theta, Z\rangle) Z Z^\top\right\}.$$

Finally, by concavity of the map $A \mapsto \mathrm{Tr}[(A + \lambda I_d)^{-1} A]$ on positive matrices (shown in the proof of Theorem 7.5), denoting $\widetilde{H}_{\lambda,n} := \mathbb{E}[\nabla^2 \widehat{R}^Z_n(\widehat{\theta}^Z_{\lambda,n})] = \mathbb{E}[\nabla^2 \widehat{R}_{n+1}(\widehat{\theta}_{\lambda,n+1})]$ we have

$$\mathbb{E}\left[\mathrm{Tr}\left\{\big[\nabla^2 \widehat{R}^Z_n(\widehat{\theta}^Z_{\lambda,n}) + \lambda I_d\big]^{-1} \nabla^2 \widehat{R}^Z_n(\widehat{\theta}^Z_{\lambda,n})\right\}\right] \leqslant \mathrm{Tr}\big\{[\widetilde{H}_{\lambda,n} + \lambda I_d]^{-1} \widetilde{H}_{\lambda,n}\big\} = \mathsf{df}_\lambda(\widetilde{H}_{\lambda,n}) . \quad (7.94)$$

Combining inequalities (7.83), (7.92), (7.93) and (7.94), we conclude that

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta \in \mathbf{R}^d}\left\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\right\} \leqslant e \cdot \frac{\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n})}{n+1} . \qquad (7.95)$$

Finally, the bound (7.48) is obtained by noting that, by exchangeability and since $\sigma' = \sigma(1-\sigma) \leqslant 1/4$ and $Z_1 Z_1^\top = X_1 X_1^\top$,

$$\widetilde{H}_{\lambda,n+1} = \mathbb{E}\big[\sigma'(\langle\widehat{\theta}_{\lambda,n+1}, Z_1\rangle) Z_1 Z_1^\top\big] \leqslant \mathbb{E}\big[X_1 X_1^\top\big]/4 = \Sigma/4 ,$$

so that $\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n}) \leqslant \mathsf{df}_\lambda(\Sigma/4) = \mathsf{df}_{4\lambda}(\Sigma)$. □

Lemma 7.4 (Stability)**.** *Let Ω be a nonempty open convex subset of $\mathbf{R}^d$, and $F : \Omega \to \mathbf{R}$ a differentiable function. Assume that F is* Σ-strongly convex *on Ω (where Σ is a $d \times d$ symmetric positive matrix), in the sense that, for every $x, x' \in \Omega$,*

$$F(x') \geqslant F(x) + \langle\nabla F(x), x' - x\rangle + \frac{1}{2}\|x' - x\|^2_\Sigma . \qquad (7.96)$$

Assume that F reaches its minimum at $x^* \in \Omega$. Let $g \in \mathbf{R}^d$, and assume that the function $x \mapsto F(x) - \langle g, x\rangle$ reaches its minimum at some $\widetilde{x} \in \Omega$. Then,

$$\|\widetilde{x} - x^*\|_\Sigma \leqslant \|g\|_{\Sigma^{-1}} , \qquad \langle g, \widetilde{x} - x^*\rangle \leqslant \|g\|^2_{\Sigma^{-1}} . \qquad (7.97)$$

Proof. First, since $\widetilde{x} \in \Omega$ minimizes the function $x \mapsto F(x) - \langle g, x\rangle$, we have $0 = \nabla F(\widetilde{x}) - g$. This implies

$$\langle \nabla F(\widetilde{x}), \widetilde{x} - x^* \rangle = \langle g, x \rangle \,. \tag{7.98}$$

Now, by substituting x' and x in inequality (7.96) and adding the resulting inequality to (7.96), we obtain for every $x, x' \in \Omega$,

$$\langle \nabla F(x') - \nabla F(x), x' - x \rangle \geqslant \|x' - x\|_\Sigma^2 \,.$$

Setting $x' = \widetilde{x}$ and $x = x^*$, and using that $\nabla F(x^*) = 0$ (since $x^* \in \Omega$ minimizes F), we obtain $\langle \nabla F(\widetilde{x}), \widetilde{x} - x^* \rangle \geqslant \|\widetilde{x} - x^*\|_\Sigma^2$. On the other hand, the Cauchy-Schwarz inequality implies that

$$\langle g, \widetilde{x} - x^* \rangle \leqslant \|g\|_{\Sigma^{-1}} \cdot \|\widetilde{x} - x^*\|_\Sigma \,. \tag{7.99}$$

Plugging the previous inequalities in (7.98) yields $\|x' - x\|_\Sigma^2 \leqslant \|g\|_{\Sigma^{-1}} \cdot \|\widetilde{x} - x^*\|_\Sigma$, hence $\|x' - x\|_\Sigma \leqslant \|g\|_{\Sigma^{-1}}$; the inequality $\langle g, \widetilde{x} - x^* \rangle \leqslant \|g\|_{\Sigma^{-1}}^2$ then follows by (7.99). □

Chapter 8

Complements

In this chapter, we include results that complement those in the previous two chapters. In Section 8.1, we provide the minimax excess risk for Gaussian linear density estimation in the *well-specified* setting, which relates to the result obtained in Chapter 7 for the SMP in the general *misspecified* setting; we also compare this risk to that of the best proper estimator (namely, the MLE) in high dimension. In Section 8.2, we complement the minimax lower bound for least squares in Chapter 6 by a lower bound on the Bayes risk under isotropic Gaussian prior for arbitrary signal-to-noise ratio; this amounts to a general lower bound for Stieltjes transforms of empirical spectral distributions of random vectors with identity covariance.

Contents

8.1 Gaussian linear density estimation in high dimension

In this complement, we determine the minimax excess risk for predictive (conditional) density estimation with respect to the linear Gaussian model in the well-specified case, which was referred to in Section 7.4.1 of Chapter 7 as well as Section 1.4.5 of the introduction.

Specifically, the setting is that of conditional density estimation, see Section 1.4 as well as Chapter 7. Here, the space of covariates is $\mathcal{X} = \mathbf{R}^d$, the response lies in $\mathcal{Y} = \mathbf{R}$. The considered (conditional) model is the *Gaussian linear model*, given by the conditional densities of the form

$$\mathcal{F} = \left\{ f_\beta(\cdot|x) := \mathcal{N}(\langle \beta, x \rangle, \sigma^2) : \beta \in \mathbf{R}^d \right\}, \tag{8.1}$$

where we set the base measure on $\mathbf{R}^d$ to be $\mu(\mathrm{d}y) = (2\pi)^{-d/2}\mathrm{d}y$ and identify densities with respect to μ with the corresponding densities. Here, σ^2 is fixed, and without loss of generality we assume that $\sigma^2 = 1$. Finally, we consider in this section the *well-specified case*, where the true conditional distribution of Y given X belongs to the class $\mathcal{F}$. The results here (and their proof) are similar in spirit to those of Chapter 6 on regression with square loss.

Setting. We assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. samples from a distribution P, such that the conditional distribution of Y given X belongs to the class $\mathcal{F}$, *i.e.* such that $Y =$

$\langle \beta^*, X\rangle + \varepsilon$ where $\varepsilon|X \sim \mathcal{N}(0,1)$. Hence, the corresponding set of distributions P of (X,Y) is characterized by the distribution P_X of covariates X, and is denoted $\mathcal{P} := \mathcal{P}_{\mathrm{Gauss}}(P_X, 1)$ (with the notation of Chapter 6). Recall from Sections 1.1.1 and 1.4 of the introduction that the *risk* of a conditional density g is

$$R(g) := \mathbb{E}[\ell(g,(X,Y))] = \mathbb{E}[-\log g(Y|X)]\,,$$

where ℓ denotes the logarithmic loss. Also, the *minimax excess risk* is by definition

$$\mathcal{E}_n^*(P_X) := \inf_{\widehat{g}_n} \sup_{P\in\mathcal{P}} \mathbb{E}[\mathcal{E}(\widehat{g}_n)] = \inf_{\widehat{g}_n} \sup_{P\in\mathcal{P}} \Big\{\mathbb{E}[R(\widehat{g}_n)] - \inf_{\beta\in\mathcal{F}} R(f_\beta)\Big\}\,, \tag{8.2}$$

where $\widehat{g}_n$ spans all estimators of Y given X. In what follows, we assume that $\mathbb{E}[\|X\|^2] < +\infty$ and that the covariance matrix $\Sigma = \mathbb{E}[XX^\top]$ is invertible.

Main result. Theorem 8.1 below provides the minimax risk, as a function of the distribution P_X of covariates.

Theorem 8.1. *If the distribution P_X is degenerate (in the sense of Definition 6.1, Chapter 6) or if $n < d$, then the minimax risk (8.2) is infinite. If P_X is non-degenerate and $n \geqslant d$, then the minimax excess risk (8.2) in the well-specified case is given by*

$$\frac{1}{2}\mathbb{E}\Big[\log\big(1 + \langle (n\widehat{\Sigma}_n)^{-1}X, X\rangle\big)\Big] = \frac{1}{2}\mathbb{E}\big[-\log(1-\widehat{\ell}_{n+1})\big]\,, \tag{8.3}$$

where $\widehat{\ell}_{n+1} = \langle (n\widehat{\Sigma}_n + X_{n+1}X_{n+1}^\top)^{-1}X_{n+1}, X_{n+1}\rangle$ denotes the leverage score *of the point X_{n+1} in the sample $X_1,\dots,X_{n+1}$. This minimax risk is achieved by the Bayes predictive posterior under uniform prior on $\mathbf{R}^d$, namely*

$$\widehat{g}_n(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle, (1 + \langle (\widehat{\Sigma}_n)^{-1}x, x\rangle))\,,$$

where $\widehat{\beta}_n^{\mathrm{LS}}$ is the OLS estimator.

First, it is worth noting that, as in the case of least-squares regression (Theorem 6.2, Chapter 6), the minimax excess risk in density estimation is characterized by the distribution of statistical leverage scores: the more uneven they are, the higher the minimax risk.

Second, the minimax risk in the well-specified case (Theorem 8.1) is precisely *half* the worst-case risk of the SMP estimator in the general misspecified case (Theorem 7.4 in Chapter 7). This implies in particular that the minimax risk in the misspecified case is at most twice that in the well-specified case.

High dimension and suboptimality of proper estimators. By convexity of the function $u \mapsto -\log(1-u)$, and since $\mathbb{E}[\widehat{\ell}_{n+1}] = d/(n+1)$ under the conditions of Theorem 8.1 (see Section 6.2.2), for every distribution P_X, the minimax risk (8.3) is at least

$$\mathcal{E}_n^*(P_X) \geqslant -\frac{1}{2}\log\big(1 - \mathbb{E}[\widehat{\ell}_{n+1}]\big) = -\frac{1}{2}\log\Big(1 - \frac{d}{n+1}\Big)\,.$$

On the other hand, by concavity of the log function, the minimax risk (8.3) is smaller than

$$\frac{1}{2}\log\Big(1 + \mathbb{E}\big[\langle (\widehat{\Sigma}_n)^{-1}X, X\rangle\big]\Big) = \frac{1}{2}\log\Big(1 + \mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big]\Big)\,;$$

when the features are Gaussian, namely $X \sim \mathcal{N}(0, \Sigma)$, we have $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})] = d/(n-d-1)$ for $n > d+1$ (Breiman and Freedman, 1983), so that

$$\mathcal{E}_n^*(P_X) \leqslant \frac{1}{2}\log\Big(1 + \frac{d}{n-d-1}\Big) = -\frac{1}{2}\log\Big(1 - \frac{d}{n-1}\Big).$$

In particular, in the high-dimensional asymptotic regime where $n, d \to \infty$ while the "hardness" of the problem is fixed, namely $d/n \to \gamma \in (0,1)$ (if $\gamma > 1$, the minimax risk is infinite by Theorem 8.1), both the above distribution-independent lower bound and the upper bound for Gaussian covariates converge to the same limit, namely

$$\frac{1}{2}\log\Big(1 + \frac{\gamma}{1-\gamma}\Big) = -\frac{1}{2}\log(1-\gamma). \tag{8.4}$$

As in the least-squares problem (Chapter 6), Gaussian covariates are almost the "easiest" covariates in terms of minimax risk in high dimension, owing to the fact that the distribution of leverage scores converges to a Dirac mass at γ.

In addition, when restricting to proper (within $\mathcal{F}$) conditional distributions, the problem is equivalent to least-squares regression, with square loss $\ell(\beta, (x,y)) = \frac{1}{2}(y - \langle \beta, x\rangle)^2$ (see e.g. Section 7.3.2). In particular, by the results in Section 6.2, the minimax *proper* estimator is the MLE $\widehat{f}_n(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle, 1)$, with risk

$$\frac{\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]}{2n} \geqslant \frac{d}{2(n-d+1)}.$$

In particular, if $d/n \to \gamma$, this quantity is asymptotically at least $\gamma/(2(1-\gamma))$, with equality in the case of Gaussian covariates. This limiting risk is strictly larger than the one for general (improper) estimators (8.4). Hence, even when the true distribution belongs to the model, using improper estimators can be advantageous in high dimension. This contrasts with the asymptotic optimality of the MLE in the classical regime where d is fixed while $n \to \infty$, and complements the results of Chapter 7 which highlight the degradation of the MLE under model misspecification.

Proof of Theorem 8.1. The proof of Theorem 8.1 follows from similar arguments as that of Theorem 6.1, hence we only highlight the part that differs. The main difference is the computation of the risk of Bayes estimators under Gaussian prior.

Lemma 8.1 (Risk of Bayes predictive posteriors)**.** *Let $\lambda > 0$. The Bayes predictive posterior under Gaussian prior $\Pi_\lambda = \mathcal{N}(0, (\lambda n)^{-1} I_d)$ on β^* is*

$$\widehat{g}_{\lambda,n}(\cdot|x) = \mathcal{N}\big(\langle \widehat{\beta}_{\lambda,n}, x\rangle, 1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} x, x\rangle\big), \tag{8.5}$$

where $\widehat{\beta}_{\lambda,n}$ denotes the Ridge estimator (6.33); when P_X is non-degenerate and $n \geqslant d$, the above is well-defined for $\lambda = 0$ and equals $\widehat{g}_n$. Then, if $\lambda > 0$ or if the previous conditions apply, we have, assuming that $Y = \langle \beta^*, X\rangle + \varepsilon$ with $\varepsilon | X \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[\mathcal{E}(\widehat{g}_{\lambda,n})] = \frac{1}{2}\mathbb{E}\big[\log\big(1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle\big)\big] + \frac{\lambda^2}{2}\cdot \mathbb{E}\bigg[\frac{\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, \beta^*\rangle^2}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\bigg]$$
$$- \frac{\lambda}{2}\cdot \mathbb{E}\bigg[\frac{n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-2} X, X\rangle}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\bigg]. \tag{8.6}$$

Proof of Lemma 8.1. A standard computation shows that the posterior distribution $\widehat{\Pi}_\lambda$ on $\beta^* \in \mathbf{R}^d$ is $\mathcal{N}(\widehat{\beta}_{\lambda,n}, n^{-1}(\widehat{\Sigma}_n + \lambda I_d)^{-1})$. The predictive posterior given $x \in \mathbf{R}^d$ is then the distribution of $Y_x \sim \mathcal{N}(\langle \beta, x\rangle, 1)$, where $\beta \sim \widehat{\Pi}_\lambda$. Now, $\beta = \widehat{\beta}_{\lambda,n} + n^{-1/2}(\widehat{\Sigma}_n + \lambda I_d)^{-1/2} Z$, where $Z \sim \mathcal{N}(0, I_d)$, while $Y_x = \langle \beta, x\rangle + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$ independent of Z. Hence, $Y_x = \langle \widehat{\beta}_{\lambda,n}, x\rangle + n^{-1/2}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1/2} Z, x\rangle + \varepsilon$ is Gaussian (conditionally on $\widehat{\beta}_{\lambda,n}, \widehat{\Sigma}_n$) with mean $\langle \widehat{\beta}_{\lambda,n}, x\rangle$ and variance $\mathrm{Var}(\varepsilon) + \mathrm{Var}(n^{-1/2}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1/2} x, Z\rangle) = 1 + n^{-1}\|(\widehat{\Sigma}_n + \lambda I_d)^{-1/2} x\|^2$, *i.e.* $Y_x \sim \widehat{g}_{\lambda,n}$.

First, consider a conditional density of the form $g(\cdot|x) = \mathcal{N}(\mu(x), \sigma^2(x))$, so that $g(y|x) = \sigma(x)^{-1}\exp(-[y-\mu(x)]^2/[2\sigma^2(x)])$. We have, for every $(x,y) \in \mathbf{R}^d \times \mathbf{R}$,

$$\ell(g,(x,y)) = \frac{1}{2}\log\sigma^2(x) + \frac{1}{2\sigma^2(x)}(y-\mu(x))^2\,,$$

so that

$$R(g) = \frac{1}{2}\mathbb{E}\big[\log\sigma^2(X)\big] + \frac{1}{2}\mathbb{E}\left[\frac{(Y-\mu(X))^2}{\sigma^2(X)}\right]$$

In particular, this risk is minimized by $g(\cdot|x) = \mathcal{N}(\langle \beta^*, x\rangle, 1)$, for which it equals $1/2$. Hence, in the case of $\widehat{g}_{\lambda,n}$, we get

$$2\,\mathbb{E}[\mathcal{E}(\widehat{g}_{\lambda,n})] = \mathbb{E}[\log(1+n^{-1}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle)] + \mathbb{E}\left[\frac{(Y-\langle \widehat{\beta}_{\lambda,n}, X\rangle)^2}{1+n^{-1}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle}\right] - 1. \quad (8.7)$$

Now, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{(Y-\langle \widehat{\beta}_{\lambda,n}, X\rangle)^2}{1+n^{-1}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle}\right] &= \mathbb{E}\left[\frac{(\langle \widehat{\beta}_{\lambda,n}-\beta^*, X\rangle - \varepsilon)^2}{1+n^{-1}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle}\right] \\
&= \mathbb{E}\left[\frac{\langle \widehat{\beta}_{\lambda,n}-\beta^*, X\rangle^2 + 1}{1+n^{-1}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle}\right] \quad (8.8) \\
&= \mathbb{E}\left[\frac{\mathbb{E}[\langle \widehat{\beta}_{\lambda,n}-\beta^*, X\rangle^2 | X_{1:n}, X] + 1}{1+n^{-1}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle}\right] \quad (8.9)
\end{aligned}$$

where $X_{1:n} = (X_1, \ldots, X_n)$ and (8.8) comes from the fact that, conditionally on $(X_{1:n}, Y_{1:n}, X)$, ε is centered with unit variance. Now since

$$\widehat{\beta}_{\lambda,n} = (\widehat{\Sigma}_n + \lambda I_d)^{-1}\cdot\frac{1}{n}\sum_{i=1}^n Y_i X_i = (\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\beta^* + (\widehat{\Sigma}_n + \lambda I_d)^{-1}\cdot\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\,,$$

we have

$$\langle \widehat{\beta}_n - \beta^*, X\rangle = \frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X_i, X\rangle - \lambda\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}\beta^*, X\rangle\,,$$

so that, using that $\mathbb{E}[\varepsilon_i|X_{1:n}, X] = 0$ and $\mathbb{E}[\varepsilon_i^2|X_{1:n}, X] = 1$,

$$\begin{aligned}
\mathbb{E}[\langle \widehat{\beta}_{\lambda,n}-\beta^*, X\rangle^2 | X_{1:n}, X] &= \frac{1}{n^2}\sum_{i=1}^n \langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X_i, X\rangle^2 + \lambda^2\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}\beta^*, X\rangle^2 \\
&= \frac{1}{n}\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n+\lambda I_d)^{-1}X, X\rangle + \lambda^2\langle (\widehat{\Sigma}_n+\lambda I_d)^{-1}X, \beta^*\rangle^2.
\end{aligned}$$

Plugging this into (8.9), we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{(Y-\langle\widehat{\beta}_{\lambda,n},X\rangle)^2}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right]-1\\
&=\mathbb{E}\left[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle+\lambda^2\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2+1}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right]-1\\
&=\mathbb{E}\left[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle+\lambda^2\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2-n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right]\\
&=\lambda^2\cdot\mathbb{E}\left[\frac{\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right]-\lambda\cdot\mathbb{E}\left[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-2}X,X\rangle}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right],
\end{aligned}$$

which together with (8.7) establishes Lemma 8.1. □

In particular, when P_X is non-degenerate and $n \geqslant d$, we obtain by setting $\lambda = 0$ in Lemma 8.1:

$$\mathbb{E}[\mathcal{E}(\widehat{g}_n)] = \frac{1}{2}\mathbb{E}\Big[\log\big(1+\langle(n\widehat{\Sigma}_n)^{-1}X,X\rangle\big)\Big] = \frac{1}{2}\mathbb{E}\big[-\log(1-\widehat{\ell}_{n+1})\big]\,,$$

where the second inequality comes from the Sherman-Morrison identity (Horn and Johnson, 1990), see Lemma 6.1. This establishes an upper bound on the minimax risk.

A matching lower bound on the minimax risk (including in the case where P_X is degenerate or $n < d$) is then obtained similarly to Theorem 6.1, from the following:

Corollary 8.1 (Bayes risk under Gaussian prior)**.** *Let $\lambda > 0$. Then, the Bayes optimal risk under Gaussian prior $\Pi_\lambda = \mathcal{N}(0,(\lambda n)^{-1}I_d)$ equals*

$$\frac{1}{2}\mathbb{E}\big[\log\big(1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle\big)\big]\,.$$

Proof. Let $\mathcal{L}(\beta^*,\widehat{f}_n) = \mathbb{E}_{\beta^*}[\mathcal{E}(\widehat{f}_n)]$ denote the Kullback-Leibler expected excess risk of the estimator $\widehat{f}_n$ under the distribution P_{β^*}, namely when $Y|X \sim \mathcal{N}(\langle\beta^*,X\rangle,1)$. The Bayes optimal estimator under prior Π_λ and under Kullback-Leibler loss is simply the predictive posterior (Berger, 1985; Lehmann and Casella, 1998), which is $\widehat{g}_{\lambda,n}$. Hence, we have

$$\begin{aligned}
&\inf_{\widehat{f}_n}\mathbb{E}_{\beta^*\sim\Pi_\lambda}[\mathbb{E}_{\beta^*}[\mathcal{E}(\widehat{f}_n)]] = \mathbb{E}_{\beta^*\sim\Pi_\lambda}[\mathcal{L}(\beta^*,\widehat{g}_{\lambda,n})]\\
&=\frac{1}{2}\,\mathbb{E}\big[\log\big(1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle\big)\big]+\\
&+\frac{\lambda^2}{2}\cdot\mathbb{E}_{\beta^*\sim\Pi_\lambda}\left[\mathbb{E}_{X_{1:n},X}\left[\frac{\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right]\right]-\frac{\lambda}{2}\cdot\mathbb{E}\left[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-2}X,X\rangle}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\right]
\end{aligned}$$

Now, by Fubini's theorem and since

$$\begin{aligned}
\mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2\big] &= \mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[X^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}\beta^*(\beta^*)^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}X\big]\\
&= X^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}\mathbb{E}_{\beta^*\sim\Pi_\lambda}[\beta^*(\beta^*)^\top](\widehat{\Sigma}_n+\lambda I_d)^{-1}X\\
&= (\lambda n)^{-1}X^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}(\widehat{\Sigma}_n+\lambda I_d)^{-1}X\\
&= (\lambda n)^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-2}X,X\rangle\,,
\end{aligned}$$

the second and third terms of the above sum compensate. This proves Corollary 8.1. □

8.2 A Marchenko-Pastur lower bound on Stieltjes transforms of ESDs of covariance matrices

In this section, we let X be a random vector in $\mathbf{R}^d$, with unit covariance: $\mathbb{E}[XX^\top] = I_d$. Given n i.i.d. variables $X_1, \dots, X_n$ distributed as X, define the *sample covariance matrix* as

$$\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top . \tag{8.10}$$

$\widehat{\Sigma}_n$ is a symmetric, positive semi-definite $d \times d$ matrix. Let $\lambda_1(\widehat{\Sigma}_n) \geqslant \dots \geqslant \lambda_d(\widehat{\Sigma}_n)$ denote the (ordered) eigenvalues of $\widehat{\Sigma}_n$, and denote $\widehat{\lambda}_{j,n} = \lambda_j(\widehat{\Sigma}_n)$ for $1 \leqslant j \leqslant d$. The *empirical spectral distribution* (ESD) of $\widehat{\Sigma}_n$ is by definition the distribution $\widehat{\mu}_n = (1/d) \sum_{j=1}^d \delta_{\widehat{\lambda}_{j,n}}$, with cumulative distribution function

$$\widehat{F}_n(x) = \frac{1}{d} \sum_{j=1}^d \mathbf{1}(\widehat{\lambda}_{j,n} \leqslant x)$$

for $x \in \mathbf{R}$. The celebrated Marchenko-Pastur theorem (Marchenko and Pastur, 1967) states that, if $X \sim \mathcal{N}(0, I_d)$, as $d, n \to \infty$ while $d/n \to \gamma \in (0, 1)$, the ESD $\widehat{\mu}_n$ converges almost surely in distribution to the *Marcenko-Pastur distribution* μ_γ^{MP}, with density

$$x \longmapsto \frac{\sqrt{(b_\gamma - x)(x - a_\gamma)}}{2\pi\gamma x} \cdot \mathbf{1}(a_\gamma \leqslant x \leqslant b_\gamma)$$

with respect to the Lebesgue measure, where $a_\gamma = (1 - \sqrt{\gamma})^2$ and $b_\gamma = (1 + \sqrt{\gamma})^2$. This behavior has a form of *universality*, in the sense that it remains true whenever the coordinates of X are independent, centered and with unit variance (Wachter, 1978; Yin, 1986). On the other hand, the independence assumption that underlies this "universal" behavior is quite strong, especially in high dimension where it implies a very specific "incoherent" geometry for the X_i's (including near-constant norm and pairwise orthogonality, see Section 1.3.2).

In this section, we show a form of *extremality* of the Marchenko-Pastur distribution among ESDs of empirical covariance matrices of general (unit covariance) random vectors in $\mathbf{R}^d$. Define the *Stieltjes transform* $S_\mu : \mathbf{R}_+^* \to \mathbf{R}$ of a probability distribution μ supported on $\mathbf{R}^+$ by

$$S_\mu(\lambda) := \int_{\mathbf{R}} (x + \lambda)^{-1} \mu(\mathrm{d}x) .$$

The Stieltjes transform (extended to $\lambda \in \mathbf{C} \setminus \mathbf{R}^-$) plays an important role in the spectral analysis of random matrices, and in particular in the proof of the Marchenko-Pastur law (Bai and Silverstein, 2010). Also, define the *expected ESD* $\bar{\mu}_n = \mathbb{E}[\widehat{\mu}_n]$ (such that $\bar{\mu}_n(A) = (1/d) \sum_{j=1}^d \mathbb{P}(\widehat{\lambda}_{j,n} \in A)$ for every measurable subset A of $\mathbf{R}$) and its cumulative distribution function $\bar{F}_n(x) := \mathbb{E}[\widehat{F}_n(x)] = (1/d) \sum_{j=1}^d \mathbb{P}(\widehat{\lambda}_{j,n} \leqslant x)$. Our main result is the following:

Theorem 8.2 (Marchenko-Pastur lower bound)**.** *Let X be a random vector in $\mathbf{R}^d$ such that $\mathbb{E}[XX^\top] = I_d$. Then, the expected Stieltjes transform of the ESD $\widehat{\mu}_n$ is lower bounded in terms of that of the Marchenko-Pastur distribution $\mu_{\gamma'}^{\mathsf{MP}}$ with $\gamma' = d/(n+1)$. Specifically, for every*

$\lambda > 0$, denoting $\lambda' = [n/(n+1)]\lambda$,

$$S_{\bar{\mu}_n}(\lambda) = \frac{1}{d}\,\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\}\big] \geqslant \frac{n}{n+1}\,\frac{-(1-\gamma'+\lambda') + \sqrt{(1-\gamma'+\lambda')^2 + 4\gamma'\lambda'}}{2\lambda'\gamma'}$$
$$= \frac{n}{n+1} S_{\mu_{\gamma'}^{\mathsf{MP}}}(\lambda')\,. \tag{8.11}$$

In particular, if $n, d \to \infty$ with $d/n \to \gamma \in (0,1)$, $\liminf_{n\to\infty} \inf_{P_X} S_{\bar{\mu}_n}(\lambda) \geqslant S_{\mu_\gamma^{\mathsf{MP}}}(\lambda)$ for every $\lambda > 0$.

Theorem 8.2 states that the Marchenko-Pastur law, which is a limiting distribution of ESDs of vectors with *independent coordinates*, also provides a non-asymptotic lower bound (in terms of associated Stieltjes transforms) for ESDs of *general* random vectors in $\mathbf{R}^d$.

Before giving the proof of Theorem 8.2 (which is elementary and relies on a combination of the Sherman-Morrison formula with a fixed-point argument), let us indicate some consequences for least-squares regression and Gaussian linear density estimation.

Let us fix a distribution P_X of covariates X such that $\Sigma := \mathbb{E}[XX^\top]$ is invertible. For $\sigma^2 > 0$, consider the statistical model $\mathcal{P} = \mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) = \{P_{(X,Y)} : Y|X \sim \mathcal{N}(\langle \beta^*, X\rangle, \sigma^2), \beta^* \in \mathbf{R}^d\}$. For $\lambda > 0$, define the prior distribution $\Pi_\lambda = \mathcal{N}(0, \sigma^2/(\lambda n)\Sigma^{-1})$ on β^*. Π_λ has constant density on the sets $\{\beta^* \in \mathbf{R}^d : \|\beta^*\|_\Sigma = t\}$ of constant *signal strength* $\|\beta^*\|_\Sigma = \mathbb{E}[\langle \beta^*, X\rangle^2]^{1/2}$. Let us also define the *signal-to-noise ratio* (SNR) $\eta^2 = \eta^2(\lambda) := \mathbb{E}_{\beta^*\sim\Pi_\lambda}[\|\beta^*\|_\Sigma^2]/\sigma^2 = d/(\lambda n)$.

Corollary 8.2 (Lower bound on Bayes risk in regression in terms of SNR)**.** *Let $\lambda > 0$, and $\eta := \eta(\lambda)$ be the corresponding SNR. Then for every distribution P_X such that $\mathbb{E}[XX^\top] = \Sigma$, the Bayes optimal risk $B_{d,n}(P_X, \eta, \sigma^2)$ under prior Π_λ for prediction under square loss $\ell(\beta, (x,y)) = (y - \langle \beta, x\rangle)^2$ is lower bounded as*

$$B_{d,n}(P_X, \eta, \sigma^2) \geqslant \sigma^2 \cdot \frac{-\big(n+1-d+d/\eta^2\big) + \sqrt{(n+1-d+d/\eta^2)^2 + 4d^2/\eta^2}}{2d/\eta^2}\,. \tag{8.12}$$

In particular, under the limit scaling $n, d \to \infty$ with $d/n \to \gamma \in (0,1)$, the Bayes risk is asymptotically lower bounded by

$$\liminf_{n\to\infty,\ d/n\to\gamma} \inf_{P_X} B_{d,n}(P_X, \eta, \sigma^2) \geqslant \sigma^2 \cdot \frac{-(1-\gamma+\gamma/\eta^2) + \sqrt{(1-\gamma+\gamma/\eta^2)^2 + 4\gamma^2/\eta^2}}{2\gamma/\eta^2}\,.$$

This lower bound is tight: indeed, when $X \sim \mathcal{N}(0, \Sigma)$, the Bayes risk converges to this limit; for fixed $\lambda > 0$ this follows from the Marchenko-Pastur law and dominated convergence, see Bai et al. (2003); Dicker (2016) for rates of convergence. This extends the observation that the minimax risk is approximately minimized in the case of Gaussian covariates (see Section 6.2.2 of Chapter 6) to the Bayes risk with arbitrary signal strength.

Proof of Theorem 8.2. First, write

$$S_{\bar{\mu}_n}(\lambda) = \mathbb{E}\bigg[\int_{\mathbf{R}} (x+\lambda)^{-1}\widehat{\mu}_n(\mathrm{d}x)\bigg] = \mathbb{E}\bigg[\frac{1}{d}\sum_{j=1}^d (\widehat{\lambda}_{j,n} + \lambda)^{-1}\bigg] = \frac{1}{d}\,\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\}\big]\,.$$

Then, denoting $\rho := (d/n)S_{\bar{\mu}_n}(\lambda)$, we have

$$\begin{aligned}
\rho &= \frac{1}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\}\big] \\
&= \mathbb{E}\big[\langle (n\widehat{\Sigma}_n + \lambda n I_d)^{-1} X, X\rangle\big] \\
&= \mathbb{E}\left[\frac{\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle}{1 - \langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle}\right] \qquad (8.13) \\
&\geqslant \frac{\mathbb{E}\big[\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle\big]}{1 - \mathbb{E}\big[\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle\big]} \qquad (8.14)
\end{aligned}$$

where (8.13) uses the Sherman-Morrison identity, while (8.14) comes from the convexity of $x \mapsto x/(1-x)$. Now, by exchangeability of $(X_1, \dots, X_n, X)$, letting $\lambda' = [n/(n+1)]\lambda$,

$$\begin{aligned}
\mathbb{E}\big[\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle\big] &= \frac{1}{n+1}\mathbb{E}\left[\sum_{i=1}^{n+1} \big\langle \big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I_d\big)^{-1} X_i, X_i\big\rangle\right] \\
&= \mathbb{E}\big[\mathrm{Tr}\big\{\big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I_d\big)^{-1}\widehat{\Sigma}_{n+1}\big\}\big] \\
&= \frac{1}{n+1}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_{n+1} + \lambda' I_d)^{-1}\widehat{\Sigma}_{n+1}\big\}\big] \\
&= \frac{d}{n+1} - \lambda' \cdot \frac{1}{n+1}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_{n+1} + \lambda' I_d)^{-1}\big\}\big] \\
&\geqslant \frac{d}{n+1} - \lambda'\rho\,, \qquad (8.15)
\end{aligned}$$

where (8.15) comes from the fact that, since $(n+1)\widehat{\Sigma}_{n+1} \geqslant n\widehat{\Sigma}_n$,

$$\begin{aligned}
\frac{1}{n+1}\mathbb{E}\big[\mathrm{Tr}\big\{\big(\widehat{\Sigma}_{n+1} + \lambda' I_d\big)^{-1}\big\}\big] &= \mathbb{E}\big[\mathrm{Tr}\big\{\big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I_d\big)^{-1}\big\}\big] \\
&\leqslant \mathbb{E}\big[\mathrm{Tr}\big\{\big(n\widehat{\Sigma}_n + \lambda n I_d\big)^{-1}\big\}\big] = \rho\,. \qquad (8.16)
\end{aligned}$$

Let $\gamma' := d/(n+1)$. It follows from (8.14) and (8.15) that ρ satisfies:

$$\rho \geqslant \frac{\gamma' - \lambda'\rho}{1 - \gamma' + \lambda'\rho}\,,$$

which after rearranging (using $1 - \gamma' + \lambda'\rho > 0$) amounts to

$$\lambda'\rho^2 + (1 - \gamma' + \lambda')\rho - \gamma' \geqslant 0\,. \qquad (8.17)$$

Since the polynomial of order 2 in ρ in equation (8.17) equals $-\gamma' < 0$ at 0, it has a positive and a negative root. Since $\rho > 0$, (8.17) is equivalent to saying that ρ is larger than the positive root of the polynomial in (8.17), which writes:

$$\rho \geqslant \frac{-(1 - \gamma' + \lambda') + \sqrt{(1 - \gamma' + \lambda')^2 + 4\gamma'\lambda'}}{2\lambda'}\,. \qquad (8.18)$$

Theorem 8.2 ensues since $S_{\bar{\mu}_n}(\lambda) = (n/d)\rho = [n/(n+1)]\rho/\gamma'$, while the Stieltjes transform of the Marchenko-Pastur distribution may be found in Bai and Silverstein (2010, Lemma 3.11). □

Proof of Corollary 8.2. Up to the changes of variables $\widetilde{X} = \Sigma^{-1/2}X$ and $\widetilde{Y} = Y/\sigma$, we reduce to the case where $\Sigma = I_d$ and $\sigma^2 = 1$. As shown in the proof of Theorem 6.1 (Chapter 6), the Bayes risk under prior $\Pi_\lambda = \mathcal{N}(0, (\lambda n)^{-1} I_d)$ is equal to $n^{-1}\mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\}]$. Corollary 8.2 then follows from Theorem 8.2 after substituting for λ', γ'. □

Conclusion and future work

In this thesis, we studied problems and methods for learning and prediction.

The first part was devoted to Random forest methods, specifically to a variant called *Mondrian forests* (MF) introduced by Lakshminarayanan et al. (2014). Our main contribution was a statistical analysis of this nonparametric procedure, relying on exact computations of relevant local and global properties of the underlying recursive partitions. We deduced minimax rates of convergence for both trees and forests; these rates extend results from Arlot and Genuer (2014) for purely uniformly random forests in dimension one, and highlight an advantage of forests over single trees by a bias reduction due to smoothing of discontinuities (Chapter 2). We also amended the original Mondrian forest procedure, which was introduced for computational reasons (in order to obtain an efficient online algorithm), in order to obtain an exact procedure with risk guarantees, and experimented with this method in a conditional density estimation context (Chapter 3).

The second part dealt with sequential prediction with expert advice. Our main contribution was an analysis of the behavior of the standard exponential weights algorithm, tuned for the worst-case adversarial setting, in the stochastic setting. We showed that the variant with decreasing learning rate achieves optimal adaptation to the sub-optimality gap of the stochastic instance in the same way as more sophisticated algorithms, but unlike the latter fails to adapt to more general Bernstein conditions on losses, and can therefore perform worse in the presence of near-optimal experts (Chapter 4). We also studied a variant of the problem with growing expert classes, and designed efficient algorithms with optimal regret in this setting (Chapter 5).

In the third part, we investigated problems of regression and density estimation, with an emphasis on linear methods. Our first main contribution was a study of random-design least-squares prediction, where we considered the minimax excess risk with respect to linear classes, as a function of the distribution of covariates. We showed that the ordinary least-squares (OLS) estimator is exactly minimax optimal in the absence of an approximation error, and asymptotically as $d = o(n)$ in the general case. In addition, we expressed the minimax excess risk in terms of the distribution of leverage scores, and deduced tight lower bounds for this problem, highlighting the fact that Gaussian design is nearly most favorable for prediction in high dimension. We also obtained upper bounds in expectation for the OLS estimator for non-Gaussian design, under weak distributional assumptions. These latter results relied on a study of the lower tail and negative moments of sample covariance matrices, for which we obtained matching upper and lower bounds under a minimal "small-ball" regularity condition (Chapter 6). Our second main contribution is the introduction of a procedure for statistical learning under logarithmic loss, which satisfies a general excess risk bound valid under model misspecification. This procedure, called *Sample Minmax Predictor* (SMP), is improper and improves over guarantees achievable by proper (within-model) predictors (which degrade

under model misspecification) as well as ones obtained through online-to-offline comparison (whose rates contain additional $\log n$ terms, and which cannot achieve uniform bounds over some unbounded classes), partially answering an open problem from Grünwald and Kotłowski (2011). We investigated this procedure in detail for conditional density estimation, with comparison classes formed by the Gaussian linear and logistic models. For logistic regression, the SMP is a simple procedure whose predictions can be computed at the cost of two logistic regressions, and it achieves a fast risk rate under weak assumptions, partly addressing an open problem from Foster et al. (2018) on efficient algorithms with such guarantees (Chapter 7). We complemented these results by a minimax analysis of Gaussian linear density estimation in the well-specified case, showing an advantage of improper estimators even in this case, provided that the dimension is moderately large. In addition, we established a non-asymptotic lower bound for the Stieltjes transform of empirical spectral distributions of sample covariance matrices of general isotropic random vectors in terms of that of the Marchenko-Pastur distribution, which extends the minimax lower bound of Chapter 6 for least-squares regression to Bayes risks depending on the signal-to-noise ratio (Chapter 8).

This work leaves a number of open questions for future research:

1. The results of Chapter 2 for Mondrian Forests, as well as those of Arlot and Genuer (2014) for Purely Random Forests, show that for the proposed (stylized) variants of Random Forests and in the considered regime, the advantage of forests over single trees lies in a *bias* reduction. As highlighted in Section 1.5.2 of the introduction, this result runs counter to the initial motivation for introducing forests, which was to reduce the *variance* of the procedure. Indeed, the results on bias reduction suggest to use *shallower* trees inside a forest than for single trees, which contrasts with the use of deep, completely developed individual trees in Random forests (Breiman, 2001a). To the best of our knowledge, no currently available result justifies the use of fully developed randomized trees with the parameters for bagging used in practice. One way to investigate this could be to consider partly randomized but more adaptive partitions, whose splits are partly data-dependent, and show some variance reduction in this case. This would formalize the intuition put forward by Breiman (1996) for bagging decision trees, that the data-dependent choice of splits makes this procedure unstable, so that bagging may reduce variance.

2. Another way to study this problem, which may be more amenable to precise analysis, would be to investigate the effect of bagging and features subsampling on simpler methods. A natural choice is linear predictors with enough variables that they can fit the dataset (in the same way as fully developed trees), which may be simple enough to be analytically tractable, yet rich enough that they can convey insights about the behavior of complex predictors in the interpolating regime, as shown by recent work (Advani and Saxe, 2017; Liang and Rakhlin, 2018; Bartlett et al., 2019; Hastie et al., 2019; Belkin et al., 2019; Muthukumar et al., 2019).

3. The bound on the lower tail of covariance matrices of Chapter 6 holds in the regime where $n \geqslant 6d$ (that is, for "tall" rectangular design matrices); while we did not attempt to optimize the factor 6, our argument does not extend to square of nearly square matrices with $d \sim n$. Extending the bound to this regime (in order to control moments of the condition number for such matrices) may require refining the proof by using the techniques from Rudelson and Vershynin (2008, 2009); Tao and Vu (2009b,a).

4. The non-asymptotic Marchenko-Pastur lower bound of Section 8.2 on Stieltjes transforms of expected ESDs of general empirical covariance matrices can be seen as a form of extremality of the Marchenko-Pastur distribution among such ESDs. Indeed, it states that
$$\int_{\mathbf{R}} f(x)\bar{\mu}_n(\mathrm{d}x) \geqslant \frac{n}{n+1}\int_{\mathbf{R}} f(x)\mu^{\mathsf{MP}}_{d/(n+1)}(\mathrm{d}x) \tag{8.19}$$
for $f = f_\lambda : x \mapsto (x+\lambda)^{-1}$, for every $\lambda > 0$ (and therefore for any positive linear combination of such functions, which is necessarily convex and in fact totally positive, in the sense that $(-1)^p f^{(p)} \geqslant 0$ for every $n \geqslant 0$). In a restricted and somewhat imprecise sense, this suggests that the Marchenko-Pastur distribution is essentially the least spread-out expected ESD given the aspect ratio d/n in high dimension. It would be interesting to see if this statement could be made more precise, by extending inequality (8.19) (or a similar one) to a wider subclass of convex functions f.

5. The distribution-free excess risk guarantees for the SMP in Chapter 7 hold in expectation, similarly to those obtained through online-to-batch conversion. It would be interesting to complement this by a procedure that achieves distribution-free high (exponential) probability excess risk bounds, for instance in the case of logistic regression.

6. Theorem 7.3 in Chapter 7 shows that the Bayes predictive posterior on the Gaussian location model under uniform prior equalizes the expected excess risk across all (misspecified) distributions. By local asymptotic normality, we expect this behavior to extend asymptotically to smooth parametric models with smooth priors. It would be interesting to see if non-asymptotic distribution-free expected excess risk bounds (similar to those of the SMP) can be obtained for Bayes predictive posteriors (possibly with a learning rate smaller than 1) without averaging, for more general exponential families with suitable (presumably log-concave) priors.

7. Finally, another research direction is to extend or refine the results on the logistic SMP. One possible direction is to obtain analogous regret guarantees for the online problem with individual sequences; the sequential analogue of the SMP may be the Sequentially Normalized Maximum Likelihood (SNML) algorithm (Roos and Rissanen, 2008; Kotłowski and Grünwald, 2011). Another direction would be to obtain excess risk bounds (in the batch setting) with logarithmic rather than quadratic dependence on the norm $\|\beta\|$, similarly to the bound for (computationally expensive) Bayes mixtures (Kakade and Ng, 2005; Foster et al., 2018) or the Ridge SMP in for the Gaussian linear model (Proposition 7.3 Chapter 7). In our setting, a technical difficulty arises in the case of logistic regression due to the use of self-concordance to control the error of the local quadratic approximation, which prevents using regularization parameters as small as in the Gaussian case. A last important question is whether fast rates under weak distributional assumptions such as those of SMP or Bayes mixtures with averaging can be obtained in a computationally more efficient way, and more generally to better understand the possible tradeoffs between computational efficiency and statistical robustness to worst-case distributions.

Bibliography

Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561.

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2011). Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3-4):195–200.

Advani, M. S. and Saxe, A. M. (2017). High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, 62(3):547–554.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588.

Anderson, G. W., Guionnet, A., and Zeitouni, O. (2010). *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley New York.

Arlot, S. (2008). V-fold cross-validation improved: V-fold penalization. *arXiv preprint arXiv:0802.0566*.

Arlot, S. and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *The Annals of Statistics*, 34(6):2921–2938.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Audibert, J.-Y. (2004). *PAC-Bayesian statistical learning theory*. PhD thesis, Université Paris VI.

Audibert, J.-Y. (2008). Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20*, pages 41–48.

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646.

Audibert, J.-Y. and Catoni, O. (2010). Linear regression through PAC-Bayesian truncation. *arXiv preprint arXiv:1010.0072*.

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794.

Auer, P., Cesa-Bianchi, N., and Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75.

Auer, P. and Chiang, C.-K. (2016). An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory (COLT)*, volume 49, pages 116–120.

Azoury, K. S. and Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627.

Bach, F. and Moulines, É. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781.

Bai, Z., Miao, B., and Yao, J.-F. (2003). Convergence rates of spectral distributions of large sample covariance matrices. *SIAM Journal on Matrix Analysis and Applications*, 25(1):105–127.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer-Verlag New York, 2 edition.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Barron, A. R. (1987). Are bayes rules consistent in information? In *Open Problems in Communication and Computation*, pages 85–91. Springer.

Barron, A. R., Rissanen, J. J., and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Grünwald, P. D., Harremoës, P., Hedayati, F., and Kotłowski, W. (2013). Horizon-independent optimal prediction with log-loss in exponential families. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 639–661.

Bartlett, P. L., Koolen, W. M., Malek, A., Takimoto, E., and Warmuth, M. K. (2015). Minimax fixed-design linear regression. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 226–239.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019). Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334.

Belkin, M., Hsu, D., and Xu, J. (2019). Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*.

Bennett, G., Dor, L. E., Goodman, V., Johnson, W. B., and Newman, C. M. (1977). On uncomplemented subspaces of l^p, $1 < p < 2$. *Israel Journal of Mathematics*, 26(2):178–187.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Stochastic Modelling and Applied Probability*. Springer-Verlag Berlin Heidelberg.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag New York.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Bhatia, R. (2009). *Positive Definite Matrices*, volume 16 of *Princeton Series in Applied Mathematics*. Princeton University Press.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(1):1063–1095.

Biau, G., Cérou, F., and Guyader, A. (2010). On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11(Feb):687–712.

Biau, G. and Devroye, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Blackwell, D. (1956). An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8.

Blanchard, G. (1999). The "progressive mixture" estimator for regression trees. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 35(6):793–820.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526.

Bousquet, O. and Warmuth, M. K. (2002). Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 577, Statistics departement, University of California Berkeley.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Breiman, L. (2004). Consistency for a simple model of random forests. Technical Report 670, Statistics departement, University of California Berkeley.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. CRC, Monterey, CA.

Brown, L. D., George, E. I., and Xu, X. (2008). Admissible predictive density estimation. *The Annals of Statistics*, pages 1156–1170.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23, pages 42.1–42.23.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4):927–961.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697.

Candès, E. J. and Sur, P. (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Catoni, O. (1997). The mixture approach to universal model selection. Technical report, École Normale Supérieure.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg.

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM*, 44(3):427–485.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. (2012). Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25*, pages 980–988.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, New York, USA.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2007). Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity analysis in linear regression*, volume 327 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York.

Chaudhuri, K., Freund, Y., and Hsu, D. J. (2009). A parameter-free hedging algorithm. In *Advances in Neural Information Processing Systems 22*, pages 297–305.

Chernov, A. and Vovk, V. (2009). Prediction with expert evaluators' advice. In *Proceedings of the 20th conference on Algorithmic Learning Theory (ALT)*, pages 8–22.

Chernov, A. and Vovk, V. (2010). Prediction with advice of unknown number of experts. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 117–125.

Chernov, A. and Zhdanov, F. (2010). Prediction with expert advice under discounted loss. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 255–269.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Clarke, B. S. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60.

Clémençon, S., Depecker, M., and Vayatis, N. (2013). Ranking forests. *Journal of Machine Learning Research*, 14(Jan):39–73.

Cover, T. M. (1968). Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pages 413–415.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, USA, 2nd edition.

Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.

Cucker, F. and Smale, S. (2002a). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428.

Cucker, F. and Smale, S. (2002b). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Cui, Y., Zhu, R., Zhou, M., and Kosorok, M. (2017). Some asymptotic results of survival tree and forest models. *arXiv preprint arXiv:1707.09631*.

Cutler, A. and Zhao, G. (2001). PERT–Perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497.

Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 537–546. ACM.

de Rooij, S., van Erven, T., Grünwald, P., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316.

De Vito, E., Caponnetto, A., and Rosasco, L. (2005). Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85.

Denil, M., Matheson, D., and de Freitas, N. (2013). Consistency of online random forests. In *Proceedings of the 30th Annual International Conference on Machine Learning (ICML)*, pages 1256–1264.

Denil, M., Matheson, D., and de Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *Proceedings of the 31st Annual International Conference on Machine Learning (ICML)*, pages 665–673.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A bayesian CART algorithm. *Biometrika*, 85(2):363–377.

Devroye, L. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 61(4):467–481.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer-Verlag.

Devroye, L. and Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018.

Devroye, L. and Wagner, T. (1979a). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207.

Devroye, L. and Wagner, T. (1979b). Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157.

Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

Domingos, P. and Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 71–80.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

Dudley, R. M. (1984). A course on empirical processes. In *École d'Été de Probabilités de Saint-Flour XII - 1982*, pages 1–142. Springer.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Number 63 in Cambridge Studies in advanced mathematics. Cambridge University Press.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128.

Durrett, R. (2010). *Probability: theory and examples*. Cambridge University Press, 4th edition edition.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175.

El Karoui, N. and Kösters, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*.

Esseen, C. G. (1966). On the Kolmogorov-Rogozin inequality for the concentration function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(3):210–216.

Faure, M., Gaillard, P., Gaujal, B., and Perchet, V. (2015). Online learning and game theory. a quick overview with recent results and applications. *ESAIM: Proceedings and Surveys*, 51:246–271.

Federer, H. (1996). *Geometric measure theory*. Springer.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.

Fernique, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. In *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*, pages 1–96. Springer.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018). Logistic regression: the importance of being improper. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, pages 167–208.

Foster, D. P. (1991). Prediction in the worst case. *The Annals of Statistics*, 19:1084–1090.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Freund, Y., Schapire, R. E., Singer, Y., and Warmuth, M. K. (1997). Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC)*, pages 334–343.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer series in statistics, New York.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Gaillard, P., Stoltz, G., and van Erven, T. (2014). A second-order bound with excess losses. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 176–196.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Sumtibus F. Perthes et I. H. Besser.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562.

George, E. I., Liang, F., and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics*, 34(1):78–91.

Gerchinovitz, S., Ménard, P., and Stoltz, G. (2017). Fano's inequality for random variables. *arXiv preprint arXiv:1702.05985*.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 353–360.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Ghosh, J. K. (1994). *Higher order asymptotics*. Institute of Mathematical Statistics.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–989.

Gofer, E., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2013). Regret minimization for branching experts. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 618–638.

Gonen, A. and Shalev-Shwartz, S. (2018). Average stability is invariant to data preconditioning. implications to exp-concave empirical risk minimization. *Journal of Machine Learning Research*, 18(222):1–13.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.

Grünwald, P. D. and Kotłowski, W. (2011). Open problem: Bounds on individual risk for log-loss predictors. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, volume 19, pages 813–816. PMLR.

Grünwald, P. D. and Mehta, N. A. (2019). A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory (ALT)*, pages 433–465.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Györfi, L., Lugosi, G., and Morvai, G. (1999). A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650.

Gyorgy, A., Linder, T., and Lugosi, G. (2012). Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 58(11):6709–6725.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 175–194.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, 76(4):675–684.

Hartigan, J. A. (1998). The maximum likelihood prior. *The Annals of Statistics*, 26(6):2083–2103.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Haussler, D., Kivinen, J., and Warmuth, M. K. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

Hazan, E., Koren, T., and Levy, K. Y. (2014). Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 197–209.

Hazan, E. and Seshadhri, C. (2009). Efficient learning algorithms for changing environments. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 393–400.

Hedayati, F. and Bartlett, P. L. (2017). Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction. *IEEE Transactions on Information Theory*, 63(10):6767–6773.

Helmbold, D. P. and Schapire, R. E. (1997). Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, 27(1):51–68.

Herbster, M. and Warmuth, M. K. (1998). Tracking the best expert. *Machine Learning*, 32(2):151–178.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press, Cambridge.

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2010). Party: A laboratory for recursive partytioning.

Hsu, D., Kakade, S. M., and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. (2017). Following the leader and fast rates in online linear prediction: curved constraint sets and other regularities. *Journal of Machine Learning Research*, 18(145):1–31.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821.

Huber, P. J. (1981). *Robust statistics*. John Wiley and Sons.

Hyafil, L. and Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Information processing letters*, 5(1):15–17.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical estimation: asymptotic theory*. Springer Science & Business Media.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Juditsky, A., Rigollet, P., and Tsybakov, A. B. (2008). Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206.

Jun, K.-S., Orabona, F., Wright, S., and Willett, R. (2017). Improved strongly adaptive online learning using coin betting. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 943–951.

Kakade, S. M. and Ng, A. Y. (2005). Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems 17*, pages 641–648.

Kakade, S. M., Sridharan, K., and Tewari, A. (2009). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800.

Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453.

Keener, R. W. (2010). *Theoretical statistics: Topics for a core course*. Springer Texts in Statistics. Springer.

Klusowski, J. M. (2018). Complete analysis of a random forest model. *arXiv preprint arXiv:1805.02587*.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *École d'Été de Probabilités de Saint-Flour*. Springer-Verlag Berlin Heidelberg.

Koltchinskii, V. and Mendelson, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, 83(2):299–313.

Koolen, W. M. (2018). Bounded excess risk for AdaHedge. Available at http://blog.wouterkoolen.info/ConstantRegret4AdaHedge/post.html. [Online; accessed 3-04-2019].

Koolen, W. M., Adamskiy, D., and Warmuth, M. K. (2012). Putting Bayes to sleep. In *Advances in Neural Information Processing Systems 25*, pages 135–143.

Koolen, W. M. and de Rooij, S. (2013). Universal codes from switching strategies. *IEEE Transactions on Information Theory*, 59(11):7168–7185.

Koolen, W. M., Grünwald, P., and van Erven, T. (2016). Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems 29*, pages 4457–4465.

Koolen, W. M. and van Erven, T. (2015). Second-order quantile methods for experts and combinatorial games. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1155–75.

Koolen, W. M., van Erven, T., and Grünwald, P. D. (2014). Learning the learning rate for prediction with expert advice. In *Advances in Neural Information Processing Systems 27*, pages 2294–2302.

Koren, T. and Levy, K. (2015). Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems 28*, pages 1477–1485.

Kotłowski, W. and Grünwald, P. D. (2011). Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 457–476.

Kushner, H. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Stochastic Modelling and Applied Probability*. Springer-Verlag New York.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems 27*, pages 3140–3148.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2016). Mondrian forests for large-scale regression when uncertainty matters. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Langford, J. and Shawe-Taylor, J. (2003). PAC-Bayes & margins. In *Advances in Neural Information Processing Systems 15*, pages 439–446.

Lattimore, T. and Szepesvári, C. (2019). Bandit algorithms. Available at https://tor-lattimore.com/downloads/book/book.pdf.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag New York.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in statistics: some basic concepts*. Springer Series in Statistics. Springer-Verlag New York.

Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. In Boucheron, S. and Vayatis, N., editors, *Topics in Learning Theory – Société Mathématique de France*.

Lecué, G. and Mendelson, S. (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*, volume 89. American Mathematical Society, Providence, RI.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: Isoperimetry and processes*. Springer Science & Business Media.

Legendre, A.-M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer.

Liang, F. and Barron, A. R. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11):2708–2726.

Liang, T. and Rakhlin, A. (2018). Just interpolate: Kernel "ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*.

Liang, T., Rakhlin, A., and Sridharan, K. (2015). Learning with square loss: localization through offset Rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 1260–1285.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110.

Littlestone, N. (1989). From on-line to batch learning. In *Proceedings of the 2nd annual workshop on Computational Learning Theory (COLT)*, pages 269–284. Morgan Kaufmann Publishers Inc.

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Information and computation*, 108(2):212–261.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: a survey. *Foundations of Computational Mathematics*, 19:1145–1190.

Luo, H. and Schapire, R. E. (2014). A drifting-games analysis for online learning and applications to boosting. In *Advances in Neural Information Processing Systems 27*, pages 1368–1376.

Luo, H. and Schapire, R. E. (2015). Achieving all with no parameters: AdaNormalHedge. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1286–1304.

Mahdavi, M., Zhang, L., and Jin, R. (2015). Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 1305–1320.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. (2019). Beyond least-squares: fast rates for regularized empirical risk minimization through self-concordance. *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, pages 2294–2340.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg.

Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366.

McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT)*, pages 203–215.

McAllester, D. A. (1999a). PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pages 164–170.

McAllester, D. A. (1999b). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, 2 edition.

McQuade, S. and Monteleoni, C. (2012). Global climate model tracking using geospatial neighborhoods. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Mehta, N. (2017). Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1085–1093.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Meir, R. and Zhang, T. (2003). Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860.

Mendelson, S. (2014). Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, volume 35, pages 25–39.

Mendelson, S. (2015). Learning without concentration. *Journal of the ACM*, 62(3):21.

Mendelson, S. and Paouris, G. (2014). On the singular values of random matrices. *Journal of the European Mathematical Society*, 16(4):823–834.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41.

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469.

Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Monteleoni, C., Schmidt, G. A., Saroha, S., and Asplund, E. (2011). Tracking climate models. *Statistical Analysis and Data Mining*, 4(4):372–392.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459.

Mourtada, J. (2019). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*.

Mourtada, J. and Gaïffas, S. (2019a). An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *arXiv preprint arXiv:1912.10784*.

Mourtada, J. and Gaïffas, S. (2019b). On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 83(1–28).

Mourtada, J., Gaïffas, S., and Scornet, E. (2017). Universal consistency and minimax rates for online Mondrian forests. In *Advances in Neural Information Processing Systems 30*, pages 3759–3768.

Mourtada, J., Gaïffas, S., and Scornet, E. (2018). Minimax optimal rates for Mondrian trees and forests. *arXiv preprint arXiv:1803.05784*.

Mourtada, J., Gaïffas, S., and Scornet, E. (2019). AMF: Aggregated Mondrian forests for online learning. *arXiv preprint arXiv:1906.10529*.

Mourtada, J. and Maillard, O.-A. (2017). Efficient tracking of a growing number of experts. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory (ALT)*, volume 76, pages 517–539.

Murray, G. D. (1977). A note on the estimation of probability density functions. *Biometrika*, 64(1):150–152.

Muthukumar, V., Vodrahalli, K., and Sahai, A. (2019). Harmless interpolation of noisy data in regression. *arXiv preprint arXiv:1903.09139*.

Nemirovski, A. (2000). Topics in non-parametric statistics. *Lectures on Probability Theory and Statistics: Ecole d'Ete de Probabilites de Saint-Flour XXVIII-1998*, 28:85–277.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nemirovski, A. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Springer US, 1 edition.

Nesterov, Y. and Nemirovskii, A. (1994). *Interior-point polynomial algorithms in convex programming*, volume 13. Society of Industrial and Applied Mathematics.

Ng, V. M. (1980). On the estimation of parametric density functions. *Biometrika*, 67(2):505–506.

Nguyen, H. H. and Vu, V. H. (2013). Small ball probability, inverse theorems, and applications. In *Erdös Centennial*, pages 409–463. Springer.

Oliveira, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194.

Orbanz, P. and Roy, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461.

Ostrovskii, D. and Bach, F. (2018). Finite-sample analysis of M-estimators using self-concordance. *arXiv preprint arXiv:1810.06838*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184.

Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, pages 1313–1320.

Rakhlin, A., Mukherjee, S., and Poggio, T. (2005). Stability results in learning theory. *Analysis and Applications*, 3(4):397–417.

Rakhlin, A. and Sridharan, K. (2015). Sequential probability assignment with binary alphabets and large classes of experts. *arXiv preprint arXiv:1501.07340*.

Raskutti, G. and Mahoney, M. W. (2016). A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(1):7508–7538.

Rissanen, J. J. (1985). *Minimum description length principle*. Wiley Online Library.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE transactions on information theory*, 42(1):40–47.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Robert, C. (2007). *The Bayesian choice*. Springer Texts in Statistics. Springer-Verlag New York.

Rockova, V. and van der Pas, S. (2017). Posterior concentration for bayesian regression trees and their ensembles. *arXiv preprint arXiv:1708.08734*.

Rogozin, B. A. (1987). The estimate of the maximum of the convolution of bounded densities. *Teoriya Veroyatnostei i ee Primeneniya*, 32(1):53–61.

Roos, T. and Rissanen, J. J. (2008). On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering*.

Rosasco, L. and Villa, S. (2015). Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems 28*, pages 1630–1638.

Roy, D. M. (2011). *Computability, inference and modeling in probabilistic programming*. PhD thesis, Massachusetts Institute of Technology.

Roy, D. M. and Teh, Y. W. (2009). The Mondrian process. In *Advances in Neural Information Processing Systems 21*, pages 1377–1384.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72.

Rudelson, M. and Vershynin, R. (2008). The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633.

Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739.

Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians*, volume III, pages 1576–1602.

Rudelson, M. and Vershynin, R. (2014). Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 2015(19):9594–9617.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Ryabko, B. Y. (1988). Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24(2):87–96.

Saffari, A., Leistner, C., Santner, J., Godec, M., and Bischof, H. (2009). On-line random forests. In *3rd IEEE ICCV Workshop on On-line Computer Vision*, pages 1393–1400.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763.

Sani, A., Neu, G., and Lazaric, A. (2014). Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems 27*, pages 810–818.

Saumard, A. (2018). On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203.

Schapire, R. E. and Freund, Y. (2012). *Boosting: Foundations and Algorithms*. MIT Press.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62:1485–1500.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269.

Seldin, Y. and Lugosi, G. (2017). An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory (COLT)*, pages 1743–1759.

Seldin, Y. and Slivkins, A. (2014). One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1287–1295.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670.

Shalizi, C. R., Jacobs, A. Z., Klinkner, K. L., and Clauset, A. (2011). Adapting to non-stationarity with growing expert ensembles. *arXiv preprint arXiv:1103.0949*.

Shamir, G. and Merhav, N. (1999). Low-complexity sequential lossless coding for piecewise-stationary memoryless sources. *IEEE Transactions on Information Theory*, 45(5):1498–1519.

Shamir, O. (2015). The sample complexity of learning linear predictors with the squared loss. *Journal of Machine Learning Research*, 16(108):3475–3486.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172.

Smola, A. J. and Schölkopf, B. (2002). *Learning with Kernels*, volume 4. MIT Press.

Spokoiny, V. (2012). Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

Srebro, N., Sridharan, K., and Tewari, A. (2010). Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, pages 2199–2207.

Sridharan, K., Shalev-Shwartz, S., and Srebro, N. (2009). Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552.

Srivastava, N. and Vershynin, R. (2013). Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111.

Stein, C. (1960). Multiple regression. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Information Science and Statistics. Springer-Verlag New York.

Steinwart, I., Hush, D., and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 79–93.

Stoltz, G. (2010). Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique*, 151(2):66–106.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sweeting, T. J., Datta, G. S., and Ghosh, M. (2006). Nonsubjective priors via predictive relative entropy regret. *The Annals of Statistics*, pages 441–468.

Taddy, M. A., Gramacy, R. B., and Polson, N. G. (2011). Dynamic trees for learning and design. *Journal of the American Statistical Association*, 106(493):109–123.

Takimoto, E. and Warmuth, M. K. (2000). The last-step minimax algorithm. In *International conference on Algorithmic Learning Theory (ALT)*, pages 279–290.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563.

Talagrand, M. (2014). *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media.

Tao, T. (2012). *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society.

Tao, T. and Vu, V. (2009a). From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices. *Bulletin of the American Mathematical Society*, 46(3):377–396.

Tao, T. and Vu, V. H. (2009b). Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. *Annals of Mathematics*, 169(2):595–632.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038.

Tjalkens, T. J., Shtarkov, Y. M., and Willems, F. M. J. (1993). Sequential weighting algorithms for multi-alphabet sources. In *6th Joint Swedish-Russian International Workshop on Information Theory*, pages 230–234.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer Berlin Heidelberg.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

Tulino, A. M. and Verdú, S. (2004). Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182.

Utgoff, P. E. (1989). Incremental induction of decision trees. *Machine learning*, 4(2):161–186.

van de Geer, S. (1999). *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge.

van de Geer, S. and Muro, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electronic Journal of Statistics*, 8(2):3031–3061.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., and Williamson, R. C. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(1):1793–1861.

van Erven, T., Koolen, W. M., de Rooij, S., and Grünwald, P. (2011). Adaptive Hedge. In *Advances in Neural Information Processing Systems 24*, pages 1656–1664.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Theory of Pattern Recognition: Statistical Learning Problems (in Russian)*. Nauka, Moscow.

Verma, N., Kpotufe, S., and Dasgupta, S. (2009). Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 565–574. AUAI Press.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press, Cambridge.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

Vovk, V. (1990). Aggregating strategies. In *Proceedings of Computational Learning Theory (COLT)*, pages 371–383.

Vovk, V. (1998). A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173.

Vovk, V. (1999). Derandomizing stochastic prediction strategies. *Machine Learning*, 35(3):247–282.

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review*, 69(2):213–248.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 6(1):1–18.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59. SIAM.

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327.

Willems, F. M. J. (1996). Coding for a binary independent piecewise-identically-distributed source. *IEEE Transactions on Information Theory*, 42(6):2210–2217.

Willems, F. M. J. (1998). The context-tree weighting method: Extensions. *IEEE Transactions on Information Theory*, 44(2):792–798.

Willems, F. M. J., Shtarkov, Y. M., and Tjalkens, T. J. (1995). The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664.

Wintenberger, O. (2017). Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362.

Xie, Q. and Barron, A. R. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445.

Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284.

Yang, Y. (2000). Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87.

Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116.

Yang, Y. and Barron, A. R. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

Yaskov, P. (2014). Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electronic Communications in Probability*, 19.

Yaskov, P. (2015). Sharp lower bounds on the least singular value of a random matrix without the fourth moment condition. *Electronic Communications in Probability*, 20.

Yin, Y. Q. (1986). Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis*, 20(1):50–68.

Ying, Y. and Pontil, M. (2008). Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596.

Zhang, T. (2006a). From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210.

Zhang, T. (2006b). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321.

Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. *arXiv preprint arXiv:1901.08779*.

Zimmert, J. and Seldin, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Contributions à l'apprentissage statistique: estimation de densité, agrégation d'experts et forêts aléatoires

Mots clés : Estimation de densité, prédiction à l'aide d'experts, forêts aléatoires, régression linéaire, régression logistique, apprentissage supervisé.

Résumé : L'apprentissage statistique fournit un cadre aux problèmes de prédiction, où l'on cherche à prédire des quantités inconnues à partir d'exemples. La première partie de cette thèse porte sur les méthodes de *Forêts aléatoires*, une famille d'algorithmes couramment utilisés en pratique, mais dont l'étude théorique s'avère délicate. Notre principale contribution est l'analyse précise d'une variante stylisée, les *forêts de Mondrian*, pour lesquelles nous établissons des vitesses de convergence non paramétriques minimax ainsi qu'un avantage des forêts sur les arbres. Nous étudions également une variante "en ligne" des forêts de Mondrian. La seconde partie est dédiée à l'agrégation d'experts, où il s'agit de combiner plusieurs sources de prédictions (experts) afin de prédire aussi bien que la meilleure d'entre elles. Nous analysons l'algorithme classique d'agrégation à poids exponentiels dans le cas stochastique, où il exhibe une certaine adaptativité à la difficulté du problème. Nous étudions également une variante du problème avec une classe croissante d'experts. La troisième partie porte sur des problèmes de régression et d'estimation de densité. Notre première contribution principale est une analyse minimax détaillée de la prédiction linéaire avec design aléatoire, en fonction de la loi des variables prédictives; nos bornes supérieures reposent sur un contrôle de la queue inférieure de matrices de covariance empiriques. Notre seconde contribution principale est l'introduction d'une procédure générale pour l'estimation de densité avec perte logarithmique, qui admet des bornes optimales d'excès de risque ne se dégradant pas dans le cas mal spécifié. Dans le cas de la régression logistique, cette procédure admet une forme simple et atteint des vitesses de convergence rapides inaccessibles aux estimateurs de type plug-in.

Title : Contributions to statistical learning: density estimation, expert aggregation and random forests

Keywords : Density estimation, prediction with expert advice, random forests, linear regression, logistic regression, supervised learning.

Abstract : The first part of this thesis is devoted to *Random forests*, a family of methods which are widely used in practice, but whose theoretical analysis has proved challenging. Our main contribution is the precise analysis of a simplified variant called *Mondrian forests*, for which we establish minimax nonparametric rates of convergence and an advantage of forests over trees. We also study an online variant of Mondrian forests. The second part is about prediction with expert advice, where one aims to sequentially combine different sources of predictions (experts) so as to perform almost as well as the best one in retrospect. We analyze the standard exponential weights algorithm on favorable stochastic instances, showing in particular that it exhibits some adaptivity to the hardness of the problem. We also study a variant of the problem with a growing expert class. The third part deals with regression and density estimation problems. Our first main contribution is a detailed minimax analysis of linear least squares prediction, as a function of the distribution of covariates; our upper bounds rely on a control of the lower tail of empirical covariance matrices. Our second main contribution is a general procedure for density estimation under entropy risk, which achieves optimal excess risk rates that do not degrade under model misspecification. When applied to logistic regression, this procedure has a simple form and achieves fast rates of convergence, bypassing some intrinsic limitations of plug-in estimators.

Institut Polytechnique de Paris
91120 Palaiseau, France