Rééchantillonnage et Sélection de modèles

Sylvain Arlot

N^o D'ORDRE : 8893

UNIVERSITÉ PARIS–SUD — FACULTÉ DES SCIENCES D'ORSAY

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ PARIS XI

Spécialité : **Mathématiques**

par

Sylvain ARLOT

RÉÉCHANTILLONNAGE ET SÉLECTION DE MODÈLES

Soutenue publiquement le **13 décembre 2007** devant la commission d'examen :

M. Patrice	**BERTAIL**	CREST et Université Paris–X	Examinateur
M. Philippe	**BERTHET**	Université Rennes–I	Examinateur
M. Gilles	**BLANCHARD**	Fraunhofer FIRST, Berlin	Examinateur
M. Stéphane	**BOUCHERON**	Université Paris–VII	Président
M. Olivier	**CATONI**	CNRS et Université Paris–VI	Examinateur
M. Pascal	**MASSART**	Université Paris–XI, Orsay	Directeur

Rapporteurs :	M. Peter L.	**BARTLETT**	University of California, Berkeley
	M. Yuhong	**YANG**	University of Minnesota

Remerciements – Acknowledgements

Mes premiers remerciements vont à Pascal, qui m'a fait découvrir le domaine de l'apprentissage statistique, en DEA puis en thèse. Tu m'as posé une question passionnante, à laquelle je suis loin d'avoir répondu intégralement, mais qui devrait m'accompagner encore de longues années. Nos discussions ont toujours été chaleureuses, franches, et riches d'enseignements. Si riches, même, qu'il me faudrait encore au moins la durée d'une ou deux thèses pour les exploiter pleinement.

I was honoured that Peter Bartlett and Yuhong Yang accepted to review this thesis. Their comments and suggestions really improved the manuscript, and raised several future directions of research. I hope they will be fruitful, and I am impatient to come to Berkeley and investigate some of them. Patrice Bertail, Philippe Berthet, Gilles Blanchard, Stéphane Boucheron et Olivier Catoni ont très gentiment accepté de participer à mon jury. Je les en remercie.

Étienne et Gilles, cela fera bientôt deux ans que nous travaillons ensemble sur ce qui est désormais le chapitre 10 de cette thèse. Nos très nombreux échanges — électroniques pour la plupart — m'ont énormément apporté, et cette thèse vous doit bien plus qu'il n'y paraît.

Je souhaite également remercier tout particulièrement Laurent Zwald, mon «grand frère» de thèse. Tes conseils avisés au cours des deux premières années de thèse m'ont été extrêmement précieux. De même que nos conversations autour d'un thé, qui ont bien souvent allongé la pause de midi.

Groupes de travail, séminaires, cours, conférences et autres voyages ont été l'occasion de nombreuses dialogues, pas uniquement mathématiques, mais toujours très enrichissants. Dans un désordre alphabétique, je remercie notamment Sylvain Baillet (qui m'a apporté le bonheur de découvrir qu'un travail théorique pouvait trouver une application *a posteriori* ; j'espère qu'elle sera fructeuse), Yannick Baraud, Lucien Birgé, Stéphane Boucheron, Alain Celisse (je suis étonné que nous n'ayions pas déjà collaboré, mais cela ne va plus tarder), Mathieu Cornec, Jérémie Jakubowicz, Aline Kurtzmann, Béatrice Laurent, Guillaume Lecué, Mathieu Lerasle, Bertrand Michel, Patricia Reynaud-Bouret, Vincent Rivoirard, Adrien Saumard, Gilles Stoltz (spécialiste ès administration), Christine Tuleau, Régis Vert, Nicolas Verzelen, Cédric Villani, et tant d'autres qui me pardonneront de ne pas les citer.

Un grand merci également à tous les relecteurs d'un ou plusieurs chapitres de cette thèse : Alain, Bertrand, Étienne, Jonas, Nicolas et Vincent.

J'ai eu la chance d'avoir des enseignants d'une qualité exceptionnelle tout au long de mon parcours. Mon goût pour l'aléatoire doit ainsi beaucoup à Jean-François Le Gall et Wendelin Werner. Je les remercie également de m'avoir permis des digressions dans mon cursus, vers la biologie et les systèmes dynamiques. Le temps m'a manqué pour les laisser apparaître dans cette thèse, mais j'espère bien qu'un jour ou l'autre, je pourrai les rapprocher des statistiques. Travailler en DEA sous la direction de Jean-Christophe Yoccoz fut un immense plaisir, j'espère que nous aurons l'occasion de poursuivre nos travaux un jour.

Les trois années très agréables que j'ai passées à Orsay doivent beaucoup à tous les (anciens) doctorants, notamment les inconditionnels de la pause thé. Je remercie ici tout d'abord Jonas (qui

a plus d'une fois égayé l'ambiance du bureau en faisant trembler ses murs), Samuel, Christophe et Jean-Patrick, avec qui partager le bureau 25 fut un plaisir. Je n'oublie pas Antoine, Frédéric, Ismaël B (organisateur talentueux et infatigable), Laurent T, Marie S, Marie T, Nicolas B, Niel, Sophie.

Un grand merci à Valérie Lavigne dont l'efficacité et la gentillesse ont rendu agréables et simples les nombreuses formalités administratives de la fin de thèse.

Enfin, mes plus vifs remerciements s'adressent à ceux que je n'ai pas cité dans ces deux pages et qui me le pardonneront. Que les autres grillent en enfer...

Toute ma gratitude va à ma famille et mes amis qui m'ont supporté pendant ces trois années, parmi lesquels je saluerais Vincent, pour le bonheur qu'il met dans ses crêpes et pour l'île d'Ouessant, Cédric, pour les fromages de chèvre à l'huile, Jérémie, pour les parties de ping-pong endiablées, Aline, Céline, Cyril, Federico, Flore, Martial, Merlin, Mickael, Noureddine, Perrine, Philippe, Sylvain, Vanya ; sans oublier Croquette, pour son incroyable faculté à reconnaître une boîte de thon. Et Anne, pour tout ce qui compte vraiment.

Au lecteur non-mathématicien

C'est à toi, lecteur curieux, qui a tourné la page des remerciements pour savoir ce qu'est une «thèse de mathématiques», et plus précisément cette thèse de «statistique», que ces deux pages sont dédiées.

Cette thèse s'inscrit dans le domaine de l'*apprentissage statistique*. Elle a pour objet principal d'étudier une méthode statistique, le *rééchantillonnage*, ainsi qu'un problème statistique, la *sélection de modèles*. Plus précisément, il s'agit d'étudier comment l'on peut utiliser cette méthode pour résoudre au mieux ce problème. Le restant de cette page a pour but d'expliquer ces trois termes de manière simple.

La *statistique* — d'après Littré — est la «science qui a pour but de faire connaître l'étendue, la population, les ressources agricoles et industrielles d'un État». En mathématique, le terme «statistique» désigne un domaine dont l'objet est d'étudier les procédés d'inférence : étant données des observations, que peut-on dire sur le processus qui les a générées ?

Par exemple, supposons que l'on dispose de mesures d'un indicateur de pollution[1] en un point de Paris à différents instants. Pour chaque mesure, on dispose d'indicateurs quantitatifs de différents phénomènes susceptibles d'expliquer[2] le taux de pollution (par exemple, la température, la pression, la vitesse du vent, la direction du vent, l'intensité du traffic automobile, l'ensoleillement, *etc.* ; mais aussi les valeurs de ces mêmes indicateurs et du taux de pollution au cours des jours qui ont précédé la mesure). L'ensemble de ces données constitue un *échantillon*.

Un problème statistique classique est alors celui de la *prédiction* : connaissant toutes ces covariables (température, *etc.*), quel est le taux de pollution attendu ? Ceci revêt un intérêt particulier si l'on à accès plus facilement aux covariables qu'au taux de pollution. Et si l'on dispose de prévisions pour toutes les covariables, on pourra réellement *prévoir* le taux de pollution.

Pour résoudre un tel problème, une idée naturelle est de fixer un *modèle* (*i.e.* un ensemble de règles de prédiction, de «prédicteurs» ; chacun associe une valeur du taux de pollution à un ensemble de covariables), puis d'«ajuster» le modèle aux données, *i.e.* déterminer le prédicteur du modèle qui se trompe le moins. Cependant, il existe de nombreuses manières de modéliser le lien entre le taux de pollution et ses covariables, aucune[3] n'étant universellement «la meilleure». Ainsi, un prédicteur issu d'un modèle très simple (*e.g.* n'utilisant qu'une ou deux covariables) ne dépend que peu du bruit[4], mais commet inévitablement des erreurs de par sa structure simpliste (on parle de *biais*). À l'inverse, un modèle très complexe (le cas extrême étant l'ensemble de toutes les règles de prédiction imaginables), n'est en général que peu biaisé, mais le prédicteur qui en résulte est beaucoup trop dépendant du bruit (on dit d'un tel modèle qu'il a une forte *variance*).

[1] Je précise que je n'ai jamais travaillé sur l'exemple du taux de pollution dans cette thèse.

[2] Il ne s'agit pas ici de relation de cause à effet, mais simplement d'une «corrélation». Ce n'est que dans le cadre d'une expérimentation scientifique qu'une étude statistique peut mettre en évidence la nature et la direction d'un lien entre deux phénomènes.

[3] en se limitant à des modèles «suffisamment génériques».

[4] le bruit modélisant les erreurs de mesures et des incertitudes liées à tous les paramètres non-observés.

Déterminer le niveau de complexité adapté à l'échantillon est la problématique de la *sélection de modèles*. Si l'on dispose de peu de données très bruité, le meilleur modèle est parmi les plus simples ; à l'inverse, avec beaucoup de données peu bruitées, il vaut mieux choisir un modèle complexe. Dans les situations intermédiaires, il s'agit de réaliser un *compromis entre le biais et la variance*. Ceci nécessite de quantifier précisément l'écart qu'il y a entre la qualité d'*ajustement aux données* (qui mesure l'erreur commise sur l'échantillon du prédicteur obtenu en utilisant ce même échantillon) et la qualité de *généralisation* (comment se comportera le prédicteur avec de nouvelles données). Cette différence est alors d'autant plus grande qu'un modèle est complexe. On appelle *pénalité* toute estimation de cette quantité. Le principal objectif de cette thèse est la *calibration précise de pénalités* pour la prédiction.

Une méthode naturelle pour évaluer la qualité d'un modèle pour la prédiction est de découper l'échantillon en deux parties : on n'utilise que la première (échantillon d'entraînement) pour ajuster le modèle aux données. Avec la seconde partie (échantillon de validation), on peut alors évaluer l'erreur commise sur de nouvelles données. Il s'agit bien là d'une erreur de *prédiction* et non d'une simple qualité d'ajustement. Une telle méthode — appelée validation — peut être utilisée pour la sélection de modèles. En général, il est préférable de répéter au moins une dizaine de fois cette opération de découpage en deux sous-échantillons. On évalue alors plus précisément l'erreur de prédiction. L'un des chapitres de cette thèse étudie les propriétés théoriques de cette méthode, très couramment utilisée.

On peut également penser utiliser cette technique de sous-échantillonnage pour construire une pénalité. L'idée est la suivante : la pénalité doit refléter l'écart qu'il y a entre l'échantillon et la distribution de nouvelles données (par exemple, un nouvel échantillon indépendant du précédent). D'une certaine manière, c'est déjà ce que fait la validation, en supprimant une partie de l'échantillon pour l'ajustement (l'échantillon de validation), pour mieux l'utiliser ensuite comme «nouvel échantillon». Le thème principal de cette thèse est l'idée de *rééchantillonnage*, qui généralise celle de sous-échantillonnage : au lieu de supprimer une partie de l'échantillon, on s'autorise notamment à donner un poids double ou triple à certaines observations et à en supprimer d'autres. L'heuristique sous-jacente est alors que la distance entre l'échantillon et le *rééchantillon*[5] est proportionnelle à la distance entre l'échantillon et un nouveau jeu de données indépendant.

[5] *i.e.*, ce «faux nouvel» échantillon.

Avertissement

La présente thèse réunit des travaux sur le rééchantillonnage et/ou la sélection de modèles, du point de vue non-asymptotique. Mise à part leur motivation commune, ces travaux ont des statuts quelque peu hétérogènes, allant du travail en cours à l'article soumis ou publié. À l'exception du Chap. 10, fruit d'une collaboration avec Gilles Blanchard et Étienne Roquain, tous les autres chapitres sont issus d'un travail personnel (et original, sauf indication contraire explicite).

Chaque chapitre peut être lu indépendamment des autres (ce qui est à l'origine de quelques répétitions), à l'exception des chapitres 8 (qui est un appendice aux Chap. 5 à 7) et 11 (qui est la conclusion de ce travail de thèse). Nous avons cependant souligné les liens qui unissent ces différents chapitres, en particulier en utilisant des notations consistantes (qui sont rassemblées pour la plupart avant le Chap. 2, puis redéfinies dans chaque chapitre où elles sont utilisées). Cette thèse est organisée comme suit :

- Le Chap. 1 est une présentation générale des connaissances actuelles
- Le Chap. 2 introduit et motive plus précisément les Chap. 3 à 9, qui visent tous à définir des procédures de sélection de modèles optimales.
- Les Chap. 3 et 4 traitent de sélection de modèles par pénalisation, pas nécessairement par rééchantillonnage.
- Dans les Chap. 5 à 9, nous considérons plusieurs procédures de choix de modèles par rééchantillonnage : la validation croisée V-fold au Chap. 5, la pénalisation V-fold et par rééchantillonnage aux Chap. 5 à 8, les complexités par rééchantillonnage globales au Chap. 9. Alors que les Chap. 5 et 6 sont sur le point d'être soumis, les trois derniers contiennent soit des résultats complémentaires, soit des travaux encore à compléter.
- Le Chap. 10 traite également de rééchantillonnage, mais avec des objectifs différents, à savoir les régions de confiances et les tests multiples. C'est une version enrichie d'un article publié [**ABR07**].
- Le Chap. 11 met en valeur les conclusions principales de ce travail, sous différents points de vue. Il rassemble également quelques problèmes ouverts qui mériteraient d'être l'objet de recherches théoriques futures.

À l'exception du Chap. 1, l'ensemble de cette thèse est rédigé en anglais. C'est pourquoi nous avons fait précéder chaque chapitre d'un cours résumé en français.

Notons également que les cadre de la régression sur de modèles d'histogrammes est considéré à plusieurs reprises, plusieurs preuves étant similaires et fondées sur les mêmes outils techniques. Bien que nous avons essayé de rendre les chapitres aussi indépendants les uns des autres que possible, il reste certains liens entre chapitres (limités aux outils probabilistes et à des résultats techniques). Ils sont organisés de la façon suivante. Mises à part quelques preuves techniques reportées au Chap. 8, le Chap. 5 peut être lu isolément. Les Chap. 3 et 6 contiennent les preuves complètes de leurs résultats principaux, mais les preuves de plusieurs résultats intermédiaires renvoient aux Chap. 5 et 8.

Foreword

The present dissertation collects works on resampling and/or model selection, from the non-asymptotical viewpoint. Apart from their common motivation, the status of these works are somehow heterogeneous (published, to be submitted, or still in progress). Chap. 10 is a joint work with Gilles Blanchard and Étienne Roquain, while all the other chapters are personal.

Each chapter can be read separately (this induces some redundancy), except Chap. 8 (which is mainly an appendix to Chap. 5 to 7) and the conclusion (Chap. 11). However, we have put a great attention on highlighting the links between chapters, in particular by using consistent notations (which are summed up before Chap. 2, and redefined in each chapter). This thesis is organized as follows:

- Chap. 1 is a general presentation (in french), of both the state-of-the art and the contributions of this thesis.
- Chap. 2 motivates more precisely Chap. 3 to 9, which deal with optimal model selection.
- Chap. 3 and 4 are about model selection by penalization, not necessarily through resampling.
- In Chap. 5 to 9, we consider several resampling model selection procedures: V-fold cross-validation in Chap. 5, V-fold and resampling penalties in Chap. 5 to 8, global resampling complexities in Chap. 9. While Chap. 5 and 6 are about to be submitted, the last three ones contain either additional material or works in progress.
- Chap. 10 is still about resampling, but with different aims, *i.e.* confidence regions and multiple testing. It is the long version of a published paper [**ABR07**].
- Chap. 11 highlights the main conclusions of this thesis, from several viewpoints. It also collects some open problems, which may deserve further theoretical researches.

Notice also that we consider several times the regression histogram framework, making several proofs similar and based upon the same technical tools. Although we have tried to make chapters as independent as possible, we could not avoid some cross-references (limited to probabilistic tools and technical results). They are organized as follows. Chap. 5 remains entirely self-contained (up to some technical proofs reported to Chap. 8). Then, both Chap. 3 and 6 contain the entire proofs of their main theorems, but several ingredients of these proofs can be found in Chap. 5 and 8.

Table des matières

CHAPITRE 1

Introduction

> La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. Ici, nous avons réuni théorie et pratique : rien ne fonctionne... et personne ne sait pourquoi!
>
> ALBERT EINSTEIN

Développer la théorie pour éclairer l'utilisation pratique de procédures statistiques est l'un des principaux objectifs de cette thèse. Il y a en effet un réel *fossé entre utilisateurs et théoriciens de la statistique*, qui semblent parfois ne joindre leurs efforts que pour le creuser davantage. On peut modérer le pessimisme de ce constat en raison du développement de la théorie statistique de l'apprentissage depuis les années 1960. Celle-ci aborde en effet le problème de l'inférence sans modéliser excessivement les données (approche *non-paramétrique*), et sans supposer la taille de l'échantillon très grande (approche *non-asymptotique*). Cette dernière hypothèse est en particulier déraisonnable lorsque la dimension des données est supérieure à la taille de l'échantillon[1], ce qui arrive fréquemment en pratique (par exemple les données de puces à ADN). Il y a donc une certaine convergence entre théorie et pratique sur ce qu'est une hypothèse «raisonnable», de même que sur les objectifs à atteindre (prédiction, adaptation, identification, *etc.*). Toutefois, les points de vue divergent sur ce qu'est une «bonne» façon d'atteindre l'un de ces objectifs.

En pratique, une bonne méthode est simple à expliquer, peu coûteuse en temps de calcul, valide dans un cadre assez général, robuste[2], stable[3] et retourne un résultat facilement utilisable[4]. Les méthodes de rééchantillonnage telles que la validation croisée «V-fold» et le bootstrap sont de bons candidats, trop mal connus encore en théorie non-asymptotique.

En général, pour un théoricien, une bonne méthode est avant tout simple à étudier, et se place dans un cadre aussi abstrait que possible. Par exemple, la validation («hold-out») repose sur un découpage en deux partie indépendantes de l'échantillon, ce qui simplifie grandement les preuves (et, en général, les rend tout simplement possibles). Mais, qu'on l'utilise telle quelle ou avec une méthode d'agrégation, la validation se fonde sur des estimateurs construits avec une partie des données seulement. À taille d'échantillon fixée, ses performances risquent donc d'être moins bonnes que celles d'une procédure fondée sur des estimateurs utilisant directement toutes les données.

Cette thèse se situe dans le cadre de la théorie statistique de l'apprentissage, en proposant une étude non-paramétrique et non-asymptotique de diverses méthodes de rééchantillonnage. L'un de nos principaux axes d'étude est la sélection de modèles, dont nous introduisons plus particulièrement les enjeux au Chap. 2. Nous proposons tout d'abord une méthode pour calibrer des pénalités à partir des données uniquement, en utilisant l'heuristique de pente (Chap. 3). Nous

[1] On parle alors de «fléau de la dimension».
[2] n'est pas trop perturbée par la présence d'une petite proportion de points aberrants.
[3] son résultat varie peu si une nouvelle donnée est ajoutée ou retirée. Voir à ce sujet Bousquet et Elisseeff **[BE02]**.
[4] une valeur prédite doit pouvoir être calculée rapidement ; en identification, on veut pouvoir interpréter simplement un résultat.

prouvons en particulier qu'elle reste valide dans un cadre *hétéroscédastique*, où de nombreuses pénalités classiques (linéaires en la dimension) ne fonctionnent plus (Chap. 4). Il apparaît donc nécessaire de définir des pénalités de forme plus générale, à partir des données.

Nous définissons alors une famille de pénalités par rééchantillonnage (Chap. 5 à 8), dont nous prouvons des qualités d'adaptation (notamment à la régularité en régression). Elles sont également robustes à l'hétéroscédasticité du bruit, là où les pénalité linéaires échouent. En comparaison avec la validation croisée V-fold, elles sont plus flexibles, ce qui se traduit par de meilleures performances, en particulier du point de vue non-asymptotique. Des éléments théoriques nous encouragent à penser que les bonnes propriétés mises en évidence dans un cadre de régression sont valables plus généralement, en particulier en classification (Chap. 7). Leur calibration devra alors être assurée par l'heuristique de pente, comme le souligne une étude théorique de la calibration de certaines pénalités globales par rééchantillonnage en classification (Chap. 9).

Enfin, nous montrons comment il est possible d'utiliser le rééchantillonnage pour construire des régions de confiance et des procédures de test multiple sur des données de grande dimension (Chap. 10). Il apparaît en particulier que le rééchantillonnage «apprend» implicitement les corrélations entre les coordonnées, qui n'ont donc pas besoin d'être modélisées ou supposées simples. Ceci nous permet de résoudre des problèmes qui se posent par exemple en imagerie cérébrale (voir Sect. 1.4.2).

Loin de mettre un point final à toutes ces questions, cette thèse se veut avant tout une ouverture vers de nombreuses voies de recherche. Nous en esquissons quelques-unes au Chap. 11.

1.1. Rééchantillonnage

Rééchantillonner, c'est utiliser un jeu de données (l'*échantillon*) pour construire un ou plusieurs nouveaux échantillons (les *rééchantillons*). Derrière une description aussi simple se cachent de nombreuses méthodes statistiques :

- le *sous-échantillonnage* consiste dans la sélection (aléatoire ou non) d'une partie des données initiales. C'est le cas du *jackknife*.
- le *bootstrap* (non-paramétrique) consiste dans la génération aléatoire un n-échantillon i.i.d., suivant la loi uniforme sur l'échantillon initialement observé. Autrement dit, on effectue n tirages avec remise dans les observations.

Dans ces deux cas, on peut — entre autres — estimer le biais ou la variance d'un estimateur en comparant le(s) rééchantillon(s) à l'échantillon initial, ou les rééchantillons entre eux.

Une autre méthode de rééchantillonnage est la *validation* (ou *hold-out*). Elle consiste à couper (aléatoirement ou non) l'échantillon en deux parties[5]. C'est donc une forme de sous-échantillonnage où l'on utilise aussi les données «supprimées». Le premier sous-échantillon (*échantillon d'entraînement* ou *d'apprentissage*) est utilisé par exemple pour construire un estimateur. Le second (*échantillon de validation*) permet alors d'évaluer la qualité de cet estimateur. Lorsqu'un tel découpage est répété, on parle de *validation croisée*. Suivant les manières d'effectuer ces découpages, on obtient par exemple le leave-one-out[6] ou la validation croisée par blocs («V-fold»).

Il existe bien d'autres formes de rééchantillonnages, et leur utilisation ne se limite pas à l'évaluation du biais et de la variance d'un estimateur. Nous ne verrons, dans cette introduction puis dans le reste de cette thèse, qu'un mince aperçu du vaste monde du rééchantillonnage en statistique.

[5] dans les applications, lorsque l'échantillon de validation sert à sélectionner un estimateur parmi plusieurs choix, on découpe souvent l'échantillon en trois parties : les deux premières sont l'échantillon d'entraînement et l'échantillon de validation ; la troisième, dite *échantillon test*, sert à évaluer (sans biais) l'erreur de l'estimateur final.

[6] l'échantillon de validation a alors un seul élément, et l'on décrit exhaustivement les n découpages possibles.

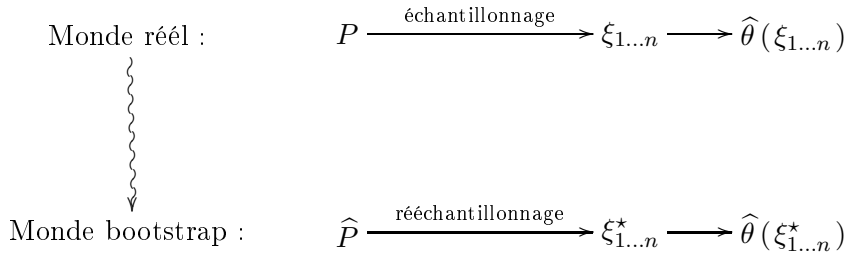


FIG. 1.1. L'heuristique de rééchantillonnage, selon Efron [**Efr79**]. Schéma inspiré de la Fig. 1 de [**Efr03**].

1.1.1. L'heuristique d'Efron. L'heuristique qui sous-tend la plupart des utilisations du rééchantillonnage est le principe de «plug-in», décrit par Efron au sujet du bootstrap [**Efr79, Efr03**]. Une autre description intuitive de cette heuristique se trouve dans l'introduction du livre de Hall [**Hal92**].

Considérons la situation suivante : n observations $\xi_1, \dots, \xi_n \in \Xi$ ont été générées indépendamment avec une même loi P inconnue. On souhaite estimer un paramètre d'intérêt $\theta = t(P)$ (par exemple la moyenne, la variance, un quantile ; mais aussi des paramètres plus complexes telles qu'une fonction de régression ou un prédicteur). Pour cela, on dispose d'un estimateur $\widehat{\theta}(\xi_{1\dots n})$. La question qui se pose alors est : comment évaluer la qualité de cet estimateur ? C'est-à-dire, évaluer son biais, sa variance, construire un intervalle de confiance, ou — s'il s'agit d'un prédicteur — déterminer son erreur de prédiction.

L'heuristique proposée par Efron est de construire un «monde bootstrap», miroir du «monde réel» mais dans lequel aucune quantité n'est inconnue (voir Fig. 1.1). La loi inconnue P est remplacée par une estimation $\widehat{P}$. Le processus d'échantillonnage est remplacé par celui de rééchantillonnage. Dans le cas du bootstrap, les processus d'échantillonnage et de rééchantillonnage sont identiques (*i.e.* $\xi^\star_1, \dots, \xi^\star_n$ sont i.i.d. de loi $\widehat{P}$). L'heuristique d'Efron se ramène donc à un principe de «plug-in» : on remplace P par son estimateur $\widehat{P}$, les autres quantités en découlant selon des processus inchangés.

La qualité d'un estimateur $\widehat{\theta}(\xi_{1\dots n})$ — de même que de nombreuses autres quantités — s'écrit sous la forme $R(P, \xi_{1\dots n})$. Cette quantité est inaccessible car P est inconnue, mais son équivalent dans le monde bootstrap $R(\widehat{P}, \xi^\star_{1\dots n})$ ne dépend que des observations. On peut donc le calculer, au besoin en faisant une approximation de type Monte-Carlo (voir Hall [**Hal92**], Appendice II et Efron et Tibshirani [**ET93**], Chap. 23). L'heuristique de rééchantillonnage se formalise donc ainsi ($\mathcal{L}(X)$ désignant la loi de la variable aléatoire X) :

$$\mathcal{L}\left(R\left(P, \xi_{1\dots n}\right)\right) \approx \mathcal{L}\left(R\left(\widehat{P}, \xi^\star_{1\dots n}\right) \,\middle|\, \xi_{1\dots n}\right) \quad .$$

L'un des principaux attraits du rééchantillonnage réside dans sa simplicité et sa généralité. En effet, la plupart des problèmes statistiques (fréquentistes) ont pour origine le fait que la loi qui a généré les observations est inconnue. En construisant un «monde bootstrap» où toutes les quantités sont accessibles (aux problèmes de temps de calcul près), on peut penser avoir trouvé la solution à tous ces problèmes. Il faut cependant garder à l'esprit que le rééchantillonnage a ses limites. Contrairement à ce que sous-entend le terme de bootstrap,[7] le «monde bootstrap» ne

[7]En anglais, l'expression «to pull oneself up by one's bootstraps» signifie «réussir uniquement par ses propres efforts ou capacités». Son étymologie serait à trouver dans les aventures du baron de Münchhausen (relatées en 1785 par Rudolph Erich Raspe, puis Gottfried August Bürger en 1786), lequel est censé être sorti d'un marécage

contient pas plus d'informations qu'il n'y en a dans les données. De plus, on connaît un certain nombre de situations dans lesquelles le bootstrap doit être modifié ou ne fonctionne pas du tout. Nous aborderons cette question plus loin dans cette introduction.

Rééchantillonnage paramétrique ou non-paramétrique. Dans l'heuristique de la Fig. 1.1, on remplace la distribution P inconnue par un estimateur $\widehat{P}$. Il y a deux catégories «classiques» d'estimateurs $\widehat{P}$:

- dans un cadre *non-paramétrique*, on ne dispose d'aucune information *a priori* sur P. On utilise alors la mesure empirique $\widehat{P} = P_n = n^{-1}\sum_{i=1}^n \delta_{\xi_i}$. Il est également possible d'estimer P à l'aide d'une version régularisée de P_n, on parle alors de «rééchantillonnage lisse»[8].
- dans un cadre *paramétrique*, on suppose que P appartient à une famille $(Q_\alpha)_{\alpha\in A}$ de distributions, avec $A \subset \mathbb{R}^d$. Étant donné un estimateur $\widehat{\alpha}$ du paramètre α_0 tel que $P = Q_{\alpha_0}$ (par exemple l'estimateur du maximum de vraisemblance), on pose alors $\widehat{P} = Q_{\widehat{\alpha}}$.

Dans cette thèse, nous nous plaçons exclusivement dans un cadre non-paramétrique, et nous ne considérerons que le cas où $\widehat{P} = P_n$.

Différents types de rééchantillonnages. Il existe une multitude de manières de rééchantillonner. Outre les méthodes déjà évoquées (jackknife, sous-échantillonnage, validation croisée, leave-one-out, bootstrap), on peut ajouter :

- le «delete d-jackknife» (équivalent au «leave-p-out») est un cas particulier de sous-échantillonnage : on supprime $d = p < n$ données choisies uniformément.
- la validation croisée V-fold consiste à partitionner les observations en V groupes. On en choisit alors un aléatoirement uniformément[9] : c'est l'échantillon de validation. La réunion des $V-1$ groupes restants constitue l'échantillon d'entraînement.
- le bootstrap «m out of n» est similaire au bootstrap, avec des rééchantillons de taille $m < n$.

Nous nous sommes pour l'instant limités à des cas où les rééchantillons peuvent s'écrire $\xi_1^\star, \ldots, \xi_m^\star$ pour un certain $m \in \mathbb{N}\backslash\{0\}$ (éventuellement aléatoire). En général, on ne considère que des quantités (estimateurs, statistiques de tests, *etc.*) qui ne dépendent que de la mesure empirique $P_n = n^{-1}\sum_{i=1}^n \delta_{\xi_i}$. On peut alors utiliser l'heuristique de la Fig. 1.1 en remplaçant le rééchantillon par une *mesure de probabilité aléatoire* $P_n^\star$, dont la loi conditionnellement à $\xi_{1\ldots n}$ est connue. Dans les cas précédemment cités, cela revient à poser $P_n^\star = m^{-1}\sum_{i=1}^m \delta_{\xi_i^\star}$.

En redéfinissant le rééchantillonnage ainsi, on augmente considérablement le nombre de façons de rééchantillonner. Mason et Newton [**MN92**] et Præstgaard et Wellner [**PW93**] ont notamment introduit la notion de «bootstrap à poids échangeables», où

$$P_n^\star := P_n^W := \frac{1}{n}\sum_{i=1}^n W_i \delta_{\xi_i}$$

où il était embourbé uniquement en se tirant par les bottes, se propulsant ainsi dans les airs. Les «bootstraps» sont les anneaux, en cuir ou en tissu, cousus sur le rebord des bottes et dans lesquels on passe les doigts pour s'aider à les enfiler. Le terme de «bootstrap» est désormais utilisé dans de nombreux domaines, notamment en informatique pour désigner le processus de démarrage d'un ordinateur.

[8] Dans le cas du bootstrap c'est le *smooth bootstrap* (Efron [**Efr82**], Silverman et Young [**SY87**], Hall, DiCiccio et Romano [**HDR89**]). Au sujet des avantages de celui-ci par rapport au bootstrap pour des problemes liés à l'estimation de quantiles ou de densités, *cf.* Falk et Reiss [**FR89**], Hall et Martin [**HM88, HM89**].

[9] Dans la validation croisée V-fold "classique" (en sélection de modèles), on parcourt exhaustivement les V choix possibles.

avec $W = (W_1, \dots, W_n) \in \mathbb{R}^n$ un vecteur de poids aléatoire, échangeable[10] et indépendant de $\xi_{1\dots n}$. Notons que l'on ne suppose pas nécessairement que les W_i sont entiers, ni même positifs. Par contre, il est en général préférable d'avoir $\mathbb{E}[W_i] = 1$. Parmi les poids classiques et/ou utilisés dans cette thèse, on trouve[11] :

- *Efron (m)*, $m \in \mathbb{N}\setminus\{0\}$: $mn^{-1}W$ suit une loi multinomiale de paramètres $(m; n^{-1}, \dots, n^{-1})$. Lorsque $m = n$, c'est le *bootstrap*. Lorsque $m < n$, c'est le *«m out of n» bootstrap*.
- *Random hold-out*[12] *(q)*, $q \in \{1, \dots, n-1\}$: $W_i = nq^{-1}\mathbb{1}_{i\in I}$ avec I un sous-ensemble aléatoire de $\{1, \dots, n\}$, choisi uniformément parmi les parties de cardinal q. Mise à part la contrainte sur la loi de I, il s'agit de poids de sous-échantillonnage (subsampling), également appelé «bootstrap without replacement». Lorsque $q = n-1$, on retrouve le leave-one-out.
- *Bootstrap à poids i.i.d.*[13] : $W_i = V_i/\overline{V}$ avec $V_1, \dots, V_n$ i.i.d. strictement positives et $\overline{V} = n^{-1}\sum_i V_i$.
  Des choix classiques pour la distribution de V_i sont la loi exponentielle de paramètre 1 (c'est le bootstrap bayésien de Rubin [**Rub81**]) et la loi Gamma de paramètre 4 (Weng [**Wen89**]).
- *Wild bootstrap* : $W_1, \dots, W_n$ sont indépendants, de même loi et vérifient $\mathbb{E}[W_i] = 1$, $\mathbb{E}\left[(W_i-1)^2\right] = 1$ et $\mathbb{E}\left[(W_i-1)^3\right] = 1$ (Wu [**Wu86**], Liu [**Liu88**], Härdle et Mammen [**HM93**]).
- *Rademacher (p)*, $p \in (0,1)$: $pW_1, \dots, pW_n$ sont i.i.d., de loi Bernoulli de paramètre p. Le nom Rademacher provient du cas $p = 1/2$ où les poids sont (à translation près) des variables Rademacher i.i.d. À la normalisation de P_n^W près, il s'agit de poids de sous-échantillonnage.
- *Poisson (μ)*, $\mu > 0$: $\mu W_1, \dots, \mu W_n$ sont i.i.d. de loi Poisson de paramètre μ.

On peut également considérer P_n^W avec des poids non-échangeables, en s'inspirant de la validation ou de la validation croisée V-fold :

- *Hold-out (q)* : $I \subset \{1, \dots, n\}$ de cardinal q étant fixé, on pose $W_i = nq^{-1}\mathbb{1}_{i\in I}$.
- *Validation croisée «V-fold»* : $(B_j)_{1\le j\le V}$ étant une partition fixe de $\{1, \dots, n\}$, on définit $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ avec J indépendant de $\xi_{1\dots n}$ et de loi uniforme dans $\{1, \dots, V\}$.

L'intérêt de disposer de nombreuses manières de rééchantillonner est qu'on peut choisir celle qui est la plus adaptée au problème que l'on se pose (Barbe et Bertail [**BB95**], Chap. 2).

1.1.2. Un peu d'histoire. Bien qu'il ait connu un essor sans précédent à partir des travaux fondateurs d'Efron [**Efr79, Efr82**], le rééchantillonnage a des origines plus anciennes. Tout d'abord, aux racines du bootstrap se trouve le jackknife, utilisé comme un moyen d'estimer le biais (Quenouille [**Que49**]) ou la variance (Tukey [**Tuk58**]) d'un estimateur. On peut également citer les tests par permutation, le sous-échantillonnage (voir notamment Hartigan [**Har69, Har75**]) et la validation croisée (Allen [**All74**], [**Sto74**]) comme méthodes de rééchantillonnage antérieures au bootstrap. D'autres références et d'autres travaux précurseurs ont également été recensés par Hall [**Hal03**].

De nombreux travaux ont suivi les articles initiaux d'Efron, sur le bootstrap lui-même, puis sur d'autres formes de rééchantillonnage. Ainsi, Mason et Newton [**MN92**] et Præstgaard et Wellner [**PW93**] ont introduit le bootstrap à poids échangeables. Plus encore, le bootstrap a permis de

[10] $W \in \mathbb{R}^n$ est échangeable si et seulement si pour toute permutation σ de $\{1, \dots, n\}$, W a la même distribution que $\left(W_{\sigma(1)}, \dots, W_{\sigma(n)}\right)$. Voir par exemple Aldous [**Ald85**].

[11] Dans cette énumération, nous indiquons en premier l'appelation utilisée dans cette thèse, qu'elle soit classique ou non.

[12] que l'on peut traduire par «validation aléatoire», à rapprocher de la validation croisée.

[13] Cette dénomination est classique mais trompeuse, car les poids W_i ne sont précisément pas i.i.d. car de somme égale à un. On parle parfois aussi de "i.i.d. generated weights". Faire attention à ne pas la confondre avec le wild bootstrap ci-après. Notons toutefois que le wild bootstrap est souvent utilisé en imposant *a posteriori* la normalisation $\sum_i W_i = 1$, auquel cas il relève également du «Bootstrap à poids i.i.d.».

porter un regard neuf sur l'idée ancienne de sous-échantillonnage, comme le montre le livre de Politis, Romano et Wolf [**PRW99**].

Aujourd'hui, on peut voir l'impact des méthodes de rééchantillonnage par une simple recherche sur Google Scholar. On trouve 350 000 résultats pour «bootstrap», 130 000 pour «cross validation» et 92 000 pour «resampling».

1.1.3. Champs d'application. De par sa formulation simple et très générale, le rééchantillonnage (en particulier le bootstrap et la validation croisée) est désormais un outil statistique utilisé dans un grand nombre de domaines. Les évoquer tous dépasserait largement le cadre de cette introduction, nous nous limiterons à citer quelques exemples en relation avec le cadre de cette thèse. Nous renvoyons à Young [**You94**] et à [**Cas03**] pour un large éventail d'applications du bootstrap. La validation croisée, quant à elle, s'applique au delà du monde de la statistique, en particulier depuis l'essor du «machine learning» (voir par exemple Hastie, Tibshirani et Friedman [**HTF01**]).

À l'origine, l'objectif d'Efron se limitait à l'estimation du biais et de la variance d'un estimateur, suivant en cela les travaux antérieurs sur le jackknife. Le bootstrap a ensuite été utilisé fructueusement pour construire des intervalles de confiance (DiCiccio et Efron [**DE96**]), calculer des p-valeurs pour des statistiques de test (voir notamment Boos [**Boo03**], Beran [**Ber03**] et Ge, Dudoit et Speed [**GDS03**]), estimer une erreur de prédiction (Wu [**Wu86**], Efron et Tibshirani [**ET97**] et Molinaro, Simon et Pfeiffer [**MSP05**]), faire de la sélection de modèles (Efron [**Efr83**], Shao [**Sha96**]), agréger des classifieurs (c'est le «bagging», contraction de «bootstrap aggregating» ; voir Bühlmann et Yu [**BY02**]), *etc.*

Notons que l'on peut utiliser d'autres types de rééchantillonnage, en particulier le sous-échantillonnage, pour construire des intervalles de confiance ou des tests (Politis, Romano et Wolf [**PRW99**]), et pour bien d'autres applications encore. La validation (croisée ou non) ne s'applique donc pas que pour évaluer l'erreur de prédiction ou sélectionner des modèles.

1.1.4. Avantages et limites. Comme nous l'avons déjà remarqué, le principal avantage du rééchantillonnage est sa simplicité. Le bootstrap et la validation croisée peuvent être décrits très intuitivement, et sont d'autant plus faciles à mettre en œuvre aujourd'hui que les capacités de calcul des ordinateurs ont très significativement augmenté. C'est notamment pour cela que le rééchantillonnage a de nombreux utilisateurs, bien au-delà du monde de la statistique.

Un deuxième point fort est qu'il n'est pas nécessaire de faire des hypothèses fortes *a priori* sur la distribution des observations. Par exemple, on peut utiliser le rééchantillonnage pour construire un intervalle de confiance sur une quantité qui est loin d'être gaussienne, par exemple parce que le nombre d'observations est insuffisant pour faire une telle approximation.

Enfin, utiliser plusieurs rééchantillons permet de «stabiliser» ou de «régulariser» un algorithme. Ceci est particulièrement utile en classification, où nombre d'algorithmes sont très sensibles à une perturbation par un petit nombre de données. Une partie des observations étant absente dans chaque rééchantillon, on peut ainsi obtenir un algorithme dont la sortie varie très peu si l'on supprime un petit nombre de données. Citons ici les forêts aléatoires (notamment utilisées pour stabiliser l'algorithme de classification CART, introduit par Breiman *et al.* [**BFOS84**]) et le bagging, entre autres algorithmes «stabilisateurs».

La simplicité du rééchantillonnage le rend cependant facile à utiliser abusivement. Tout d'abord, rééchantillonner ne fournit pas de nouvelles observations, mais simplement plus d'informations sur les mêmes observations. Cette constatation peut sembler évidente, mais elle implique notamment une difficulté à estimer les queues de distributions lorsque celles-ci sont plus lourdes

que gaussiennes (voir par exemple Hall [**Hal90**]). Cette difficulté peut notamment être contournée en remplaçant le bootstrap d'Efron par le bootstrap «m out of n» pour un $m \ll n$ bien choisi (Bretagnolle [**Bre83**]).

De plus, dans sa formulation non-paramétrique[14], le rééchantillonnage (à poids échangeables) donne le même poids à toutes les données. Il est donc important que celles-ci soient *échangeables*. Si ce n'est pas le cas, il faut alors faire très attention à la manière d'appliquer l'heuristique de rééchantillonnage. Par exemple, dans le cadre de la régression sur un plan d'expérience (design) fixe, les données sont de la forme $(X_1, Y_1), \ldots, (X_n, Y_n)$ avec $X_{1\ldots n}$ déterministes, $Y_i = s(X_i) + \epsilon_i$ et les ϵ_i sont i.i.d. On est donc amené à *rééchantillonner les résidus*. L'idée est la suivante : on estime les résidus $\epsilon_1, \ldots, \epsilon_n$ par $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$ puis on «rééchantillonne»[15] $(\widehat{\epsilon}_i)_{1\leq i\leq n}$ en construisant $(\widehat{\epsilon}_i^\star)_{1\leq i\leq n}$. Enfin, on définit le rééchantillon $(X_i, s(X_i)+\widehat{\epsilon}_i^\star)_{1\leq i\leq n}$. Dans le cas hétéroscédastique, le niveau de bruit σ_i dépend de l'indice i et ce sont les ϵ_i/σ_i qui sont i.i.d., avec $(\sigma_i)_{1\leq i\leq n}$ inconnu. Le problème devient alors encore plus délicat (Wu [**Wu86**]), et il devient nécessaire d'utiliser le wild bootstrap. L'alternative «bootstrapping pairs»[16] contre «bootstrapping residuals»[17] est également considérée à la Sect. 9.5 du livre d'Efron et Tibshirani [**ET93**].

Lorsque les données sont *dépendantes*, l'inconsistance du bootstrap a rapidement été prouvée (Singh [**Sin81**]), mais une telle astuce ne peut plus être appliquée. Le cadre le plus étudié jusqu'à présent est celui des séries temporelles (Politis [**Pol03**]), et en particulier les méthodes de rééchantillonnage par blocs («block resampling», où l'on ne rééchantillonne plus des données individuelles, mais des groupes de données, à l'intérieur desquels la dépendance des observations est conservée).

Notons également que le jackknife ne permet pas d'obtenir des intervalles de confiance satisfaisants (ce en quoi il a très vite été surpassé par le bootstrap). Pour le même genre de raisons, il est très périlleux d'utiliser le leave-one-out (ou jackknife) en faisant une approximation Monte-Carlo, ne serait-ce que parce que l'on n'utilise pas toutes les données séparément. Cela conduit en général à perdre l'effet stabilisateur du rééchantillonnage. On préfère alors utiliser la validation croisée V-fold (avec un V notamment adapté au temps de calcul disponible), ou bien faire une approximation Monte-Carlo avec le bootstrap ou le bootstrap échangeable avec des poids «Rademacher» ou «Random hold-out» (définis Sect. 6.3.3).

Ceci n'est évidemment pas une liste exhaustive de limitations du rééchantillonnage. Pour le cas du bootstrap, nous renvoyons en particulier à Mammen [**Mam92**] et à [**Cas03**].

1.1.5. Résultats théoriques. Tout comme pour ce qui concerne les applications, les études théoriques sur le rééchantillonnage se sont multipliées depuis les travaux d'Efron. On peut citer par exemple les livres d'Efron [**Efr82**], Efron et Tibshirani [**ET93**], Hall [**Hal92**], Shao et Tu [**ST95**] (sur le bootstrap), Barbe et Bertail [**BB95**] (sur le bootstrap à poids).

Pour la majorité des cas, on peut distinguer deux types de résultats : des calculs exacts (d'espérance, de variance, de l'estimateur bootstrap, *etc.*) et des résultats asymptotiques (consistance, théorème central limite, etc.). Les premiers — quoi que très instructifs sur les «bonnes» façons de rééchantillonner — sont souvent limités à des cas particuliers, et peu utilisables dans d'autres cadres. C'est pourquoi nous n'en mentionnerons pas ici.

[14] dans le cas paramétrique, tout dépend de l'estimateur $\widehat{\alpha}$.

[15] ceci ne peut pas être fait avec n'importe quel type de rééchantillonnage, car on a besoin d'un vrai rééchantillon de taille n.

[16] rééchantillonner les couples de données (X_i, Y_i).

[17] rééchantillonner les résidus.

Résultats asymptotiques. L'essentiel des résultats en dimension 1 sur le bootstrap peuvent être trouvés dans le livre de Hall [**Hal92**]. Celui-ci, grâce aux développements d'Edgeworth, met en évidence la propriété «stabilisatrice» du bootstrap. Dans diverses circonstances, en particulier pour construire des intervalles de confiance, le bootstrap fournit ainsi des quantités asymptotiquement correctes au second ordre. Des résultats similaires au sujet du bootstrap à poids échangeables peuvent être trouvés dans le livre de Barbe et Bertail [**BB95**].

Une approche du bootstrap fondée sur les processus empiriques est à trouver dans les articles de Giné et Zinn [**GZ90**] puis Arcones et Giné [**AG92**]. Dans le cas du bootstrap à poids échangeables, des énoncés similaires sont énoncés au Chap. 3.6 du livre de van der Vaart et Wellner [**vdVW96**] (voir aussi le cours de Saint-Flour de Giné [**Gin97**]). Ceux-ci se fondent sur les résultats de Mason et Newton [**MN92**] (consistance), Præstgaard et Wellner [**PW93**] (normalité asymptotique), Huškova et Janssen [**HJ93**] (cas des U-statistiques), ou encore Hall et Mammen [**HM94**] (propriétés du second ordre).

La clé de voûte de ces résultats est sans doute le «Conditional Multiplier Central Limit Theorem», que l'on trouve au Chap. 2.9 du livre de van der Vaart et Wellner [**vdVW96**]. On en déduit alors le résultat suivant (où l'on a omis des conditions de mesurabilité sur $\mathcal{F}$ pour simplifier l'énoncé) :

THÉORÈME 1.1 (Théorème 3.6.13, van der Vaart et Wellner [**vdVW96**]). *Soit $\mathcal{F}$ une classe de Donsker de fonctions mesurables. Pour tout $n \in \mathbb{N}$, soit $(W_{n,1}, \dots, W_{n,n})$ un vecteur aléatoire positif, échangeable, indépendant de $\xi_{1\dots n}$ tel que*

$$\sup_{n\in\mathbb{N}} \left\| W_{n,1} - \overline{W}_n \right\|_{2,1} < \infty \qquad \text{avec} \quad \overline{W}_n = n^{-1} \sum_{i=1}^{n} W_{n,i} \quad \text{et} \quad \|Z\|_{2,1} := \int_0^\infty \sqrt{\mathbb{P}\left(|Z| > t\right)} dt$$

$$n^{-1/2} \mathbb{E}\left[\max_{1\le i\le n} \left| W_{n,i} - \overline{W}_n \right| \right] \xrightarrow{(p)} 0 \qquad \text{et} \qquad n^{-1} \sum_{i=1}^{n} \left(W_{n,i} - \overline{W}_n \right)^2 \xrightarrow{(p)} c^2 > 0 \ .$$

Alors, lorsque n tend vers l'infini,

$$\sup_{h\in BL_1} \left| \mathbb{E}_W \left[h\left(\widehat{\mathbb{G}}_n \right) \right] - \mathbb{E}\left[h\left(c\mathbb{G} \right) \right] \right| \xrightarrow{(p)} 0$$

$$\text{avec} \quad \widehat{\mathbb{G}}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{n,i} \left(\delta_{\xi_i} - P_n \right) = \sqrt{n} \left(P_n^W - \overline{W}_n P_n \right) \ ,$$

$\mathbb{G}$ étant un processus gaussien de moyenne nulle et de fonction de covariance $\operatorname{cov}(f,g) = P(fg) - P(f)P(g)$ et BL_1 l'ensemble des fonctions 1-lipschitziennes et bornées par 1 (pour la norme $\|\cdot\|_\infty$).

Résultats non-asymptotiques. Si la théorie probabiliste asymptotique du bootstrap à poids échangeables est assez bien documentée, ce n'est pas le cas de son pendant non-asymptotique. Au sujet du sous-échantillonnage, on mentionnera les inégalités de concentration sur la somme de Serfling [**Ser74**]. Dans le cadre de la sélection de modèles, voir aussi Györfi *et al.* [**GKKW02**] et van der Laan, Dudoit et Keles [**vdLDK04**].

Pour ce qui est du bootstrap et du bootstrap à poids échangeables, plusieurs inégalités en espérance sont disponibles. Lorsque les poids $W_i - 1$ sont i.i.d. symétriques, on peut utiliser une inégalité de symétrisation classique (voir par exemple Giné et Zinn [**GZ84**], ou le lemme 1 de Fromont [**Fro07**]) :

LEMME 1.1 (Inégalité de symétrisation). *Soit $\mathcal{F}$ une famille de fonctions et $W_1, \dots, W_n$ une suite de variables aléatoires i.i.d. L^1, symétriques (*i.e. *$W_1 - 1 \sim 1 - W_1$), indépendantes de $\xi_{1\dots n}$.*

Alors,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P-P_n)(f)\right] \leq \frac{2}{\mathbb{E}\,|W_1-1|}\mathbb{E}\left[\sup_{f\in\mathcal{F}}(P_n-P_n^W)(f)\right] \ .$$

Lorsque les poids $W_1,\dots,W_n$ sont échangeables de somme constante égale à n, la même inégalité est prouvée par Fromont [**Fro07**], dans la preuve de sa Prop. 2. Mentionnons également l'inégalité de Poissonisation de Le Cam, qui permet de relier le bootstrap au bootstrap à poids Poisson (μ) (avec $\mu=1$), qui ont l'avantage d'être indépendants.

LEMME 1.2 (Lemme de Poissonnisation de Le Cam). *Soit N_n une variable de Poisson de moyenne n, indépendante d'une suite $(\xi_i)_{i\geq 1}$ de v.a.i.i.d $\xi_{1\dots n}$ de loi commune P. Alors, pour toute classe de fonctions $\mathcal{F}$,*

$$\left(1-\frac{1}{e}\right)\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\left(f(\xi_i)-P(f)\right)\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N_n}\left(f(\xi_i)-P(f)\right)\right] \ .$$

Par ailleurs, Fromont [**Fro07**] prouve plusieurs inégalités de concentration (Prop. 1 et 2) fondées sur l'inégalité de McDiarmid [**McD89**] (rappelée avec la Prop. 8.7, Sect. 8.5), d'où des termes de reste en $n^{-1/2}$:

PROPOSITION 1.3 (Fromont [**Fro07**], Prop. 1). *Soit $\mathcal{F}$ une famille dénombrable de fonctions $\Xi\mapsto[0;1]$ et $W_1,\dots,W_n$ une suite de variables aléatoires i.i.d. L^1, symétriques (*i.e. $W_1-1\sim 1-W_1$*), indépendantes de $\xi_{1\dots n}$. Alors, pour tout $x>0$, on a avec probabilité au moins $1-e^{-x}$:*

$$\sup_{f\in\mathcal{F}}(P-P_n)(f) \leq \frac{2}{\mathbb{E}\,|W_1-1|}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(P_n-P_n^W\right)(f)\,\middle|\,\xi_{1\dots n}\right]+3\sqrt{\frac{x}{2n}} \ .$$

La Prop. 2 de Fromont [**Fro07**] donne un résultat similaire lorsque les poids sont échangeables de somme n. Mais la majorité des résultats non-asymptotiques concernent les poids Rademacher, car ils correspondent aux complexités de Rademacher bien connues en théorie de l'apprentissage. Voir par exemple la revue de Boucheron, Bousquet et Lugosi [**BBL05**].

1.2. Sélection de modèles

Après l'idée de rééchantillonnage, le thème principal de cette thèse est la sélection de modèles, considérée d'un point de vue non-asymptotique. La principale référence de cette section est le cours de Saint-Flour de Massart [**Mas07**]. Nous commençons par présenter le problème de la prédiction, dans lequel les travaux de cette thèse s'inscrivent plus particulièrement (sans s'y restreindre). Nous décrivons ensuite la problématique de la sélection de modèles, différentes stratégies pour l'aborder, et quelques résultats théoriques sur ces stratégies.

1.2.1. Cadre de la prédiction. Le problème de la prédiction s'inscrit dans la *théorie statistique de l'apprentissage* (pendant statistique du «machine learning»), qui a notamment été initiée par les travaux de Vapnik [**Vap82, Vap98**]. On peut le décrire comme suit. On observe n réalisations indépendantes $\xi_1=(X_1,Y_1),\dots\xi_n=(X_n,Y_n)\in\Xi=\mathcal{X}\times\mathcal{Y}$ d'une variable aléatoire[18] $\xi\in(X,Y)$ de loi inconnue P. Étant donnée une nouvelle réalisation $\xi_{n+1}=(X_{n+1},Y_{n+1})$ de (X,Y), indépendante des précédentes, on aimerait pouvoir prédire Y_{n+1} à l'aide de X_{n+1} (et des n observations précédentes). Autrement dit, on cherche à construire un *prédicteur* $t:\mathcal{X}\mapsto\mathcal{Y}$. Typiquement, X est beaucoup plus facilement observable qu'une quantité d'intérêt Y, et un prédicteur permet d'y accéder à moindre coût, ou même de l'évaluer lorsque elle est inaccessible.

[18] La notation $\xi_i=(X_i,Y_i)$ indique que la prédiction s'inscrit dans le cadre présenté en Sect. 1.1. Dans la suite de cette introduction, par souci de généralité, nous utiliserons la notation générale ξ_i aussi souvent que possible.

Pour mesurer la qualité d'un prédicteur, on a besoin d'une mesure de «distance» entre $t(X_{n+1})$ et Y_{n+1}. Notons $\mathcal{S}$ l'ensemble des prédicteurs. Étant donné un *contraste* $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}^+$, la qualité d'un prédicteur t est mesurée par son *risque*

$$P\gamma(t,\cdot) := \mathbb{E}_{(X,Y)\sim P}\left[\gamma\left(t,(X,Y)\right)\right] = \mathbb{E}\left[\gamma\left(t,(X_{n+1},Y_{n+1})\right) \mid t\right] \ .$$

Dans la suite, on utilisera la notation fonctionnelle $Q\gamma(t)$ définie ci-dessus avec $Q = P$ pour différentes mesures de probabilités Q, et différents estimateurs t. En particulier, nous insistons ici sur le fait que si t est aléatoire (par exemple fonction de $\xi_{1\ldots n}$), alors $P\gamma(t)$ est lui-même aléatoire. Implicitement, l'intégration n'est réalisée que relativement à la nouvelle observation ξ_{n+1} (dans le membre de droite de la définition du risque).

Le risque minimal n'étant pas nul en général, on préfère au risque la notion d'*excès de risque*

$$l(s,t) := P\gamma(t) - \inf_{u\in\mathcal{S}}\left\{P\gamma\left(u\right)\right\} \geq 0 \ .$$

Le *prédicteur idéal* s, lorsqu'il existe, est appelé *prédicteur de Bayes*, est celui qui minimise le risque, c'est-à-dire

$$s \in \arg\min_{t\in\mathcal{S}}\left\{P\gamma(t)\right\} \ .$$

Notons qu'il n'est pas nécessairement unique.

Exemples. Avant d'aller plus loin, nous présentons deux exemples, la régression et la classification, dans lesquels se situent l'essentiel des résultats de cette thèse.

Exemple 1.1 (Régression). La variable d'intérêt Y est un scalaire et prend des valeurs continues (*i.e.* $\mathcal{Y}$ est un intervalle de $\mathbb{R}$). Par exemple, un indicateur du taux de pollution au centre de Paris. La variable explicative X peut être de nature assez générale, mais un cas typique est $X \in \mathcal{X} \subset \mathbb{R}^d$. Autrement dit, on cherche à prédire Y à l'aide de d paramètres scalaires. Dans l'exemple du taux de pollution, les coordonnées de X peuvent par exemple être l'intensité du trafic automobile en différents endroits, la vitesse du vent, la température, la pression atmosphérique, *etc.*

Une autre manière de formuler un problème de régression est la suivante :

$$Y = \eta(X) + \sigma(X)\epsilon \qquad \text{avec} \quad \eta(X) = \mathbb{E}\left[Y \mid X\right] \ .$$

La fonction $\eta : \mathcal{X} \mapsto \mathcal{Y}$ est la *fonction de régression*, ϵ est un terme de bruit (centré et de variance 1 conditionnellement à X, mais pas forcément indépendant de X), $\sigma : \mathcal{X} \mapsto \mathbb{R}^+$ est le niveau de bruit. Dans l'écriture ci-dessus, on a séparé σ et ϵ pour insister sur la possibilité d'avoir un niveau de bruit variable. Nous nous placerons dans un tel cadre, dit *hétéroscédastique*, dans plusieurs résultats de cette thèse.

Un contraste souvent utilisé en régression est le *contraste des moindre carrés*

$$\gamma : (t,(x,y)) \mapsto \left(t(x) - y\right)^2 \ .$$

Celui-ci est naturel car le prédicteur de Bayes est alors la fonction de régression $s = \eta$. En effet, pour tout prédicteur t,

$$\begin{aligned}\mathbb{E}(t(X) - Y)^2 &= \mathbb{E}\left(t(X) - \eta(X)\right)^2 + 2\mathbb{E}\left[(t(X) - \eta(X))\mathbb{E}\left[(\eta(X) - Y) \mid X\right]\right] + \mathbb{E}\left(\eta(X) - Y\right)^2 \\ &= \mathbb{E}\left(t(X) - \eta(X)\right)^2 + \mathbb{E}\left(\eta(X) - Y\right)^2 \geq \mathbb{E}\left(\eta(X) - Y\right)^2 \ .\end{aligned}$$

Par conséquent, l'excès de risque des moindre carrés est le carré de la distance L^2 à s :

$$l(s,t) = \mathbb{E}\left[\left(t(X) - s(X)\right)^2\right] \ .$$

Notons que le problème de régression décrit ici est aussi appelé *régression sur un plan d'expérience (design) aléatoire*, l'objectif étant de prédire Y_{n+1} avec X_{n+1} copie indépendante de X_1. On peut également considérer le cas du *design fixe*, où X_{n+1} est choisi uniformément parmi $X_1, \dots, X_n$. Ceci revient à considérer que $X_1, \dots, X_n$ sont déterministes, $\epsilon_1, \dots, \epsilon_n$ i.i.d. et $Y_i = \eta(X_i) + \sigma(X_i)\epsilon_i$. Pour la distinction entre ces deux cadres, on consultera l'introduction de Breiman [**Bre92**], ou encore Baraud [**Bar00, Bar02**].

EXEMPLE 1.2 (Classification binaire supervisée). En classification (supervisée[19]), la variable d'intérêt Y est discrète, c'est-à-dire $\mathcal{Y}$ est fini. Autrement dit, à chaque variable X, on associe une étiquette $Y \in \mathcal{Y}$. Pour simplifier, nous nous concentrerons sur le cas de la classification binaire, c'est-à-dire lorsque Y ne prend que deux valeurs : $\mathcal{Y} = \{0, 1\}$.

Ceci correspond à de nombreux problèmes réels, par exemple :

- Aide au diagnostic médical : étant donnés les résultats (X) de différents examens, un patient est-il malade ou sain (Y) ? Et s'il est malade, quelle est sa maladie ? En particulier, on peut utiliser les données de puces à ADN pour répondre à ces questions (Tibshirani *et al.* [**THNC03**]).
- Bioinformatique : détection de gènes dans une séquence ADN ($\mathcal{X} = \{A, T, C, G\}^{\mathbb{N}}$), de sites actifs dans une protéine ($\mathcal{X} = \{\text{acides aminés}\}^{\mathbb{N}}$) ; catégorisation de protéines, *etc.*
- Reconnaissance et identification de caractères manuscrits : X est une image en niveaux de gris, *i.e.* $\mathcal{X} = [0, 1]^K$ où K est le nombre de pixels. L'étiquette Y indique si X représente ou non un caractère donné.
- Classification des e-mails entre spams et non-spams.
- Catégorisation de textes : Y indique alors si le texte X relève ou non d'une thématique donnée.
- Reconnaissance de paroles (X est une donnée sonore), de formes (X est une image digitale), *etc.*

En général, l'ensemble $\mathcal{X}$ est complexe ou de grande dimension, si bien qu'il n'est pas envisageable d'estimer directement la fonction de régression $\eta(X) = \mathbb{P}(Y = 1 \mid X)$. La «simplicité» de $\mathcal{Y}$ en classification est ainsi compensée par la grande complexité de $\mathcal{X}$.

Il existe différents contrastes classiques en classification. Nous ne considérons ici que le contraste 0-1, *i.e.*

$$\gamma : (t, (x, y)) \mapsto \mathbb{1}_{t(x) \neq y} \ ,$$

qui coïncide avec le contraste des moindre carrés car on a choisi $\mathcal{Y} = \{0, 1\}$. Ainsi, le risque de t est le nombre moyen d'erreurs commises par t.

On peut alors exprimer le prédicteur de Bayes à l'aide de la fonction de régression :

$$\forall x \in \mathcal{X}, \qquad s(x) = \mathbb{1}_{\eta(x) \geq \frac{1}{2}} \ .$$

En effet, pour tout prédicteur t,

$$\begin{aligned} P\gamma(t) &= P\gamma(s) + \mathbb{E}\left[\mathbb{1}_{t(X) \neq Y} - \mathbb{1}_{s(X) \neq Y}\right] \\ &= P\gamma(s) + \mathbb{E}\left[\mathbb{1}_{t(X) \neq s(X)} \mathbb{E}\left[\mathbb{1}_{s(X) = Y \neq t(X)} - \mathbb{1}_{t(X) = Y \neq s(X)} \,\middle|\, X\right]\right] \\ &= P\gamma(s) + \mathbb{E}\left[\mathbb{1}_{t(X) \neq s(X)} \, |2\eta(X) - 1|\right] \geq P\gamma(s) \ , \end{aligned}$$

[19] On distingue trois types de classifications : supervisée (lorsque l'on observe l'étiquette Y_i pour toutes les données X_i), non-supervisée (lorsque l'étiquette Y_i n'est jamais observée ; on cherche alors uniquement à constituer des groupes cohérents, c'est le *clustering*), et semi-supervisée (où l'on n'observe l'étiquette Y_i que pour une partie — en général petite — des données). Dans cette thèse, nous ne considérons que la classification supervisée.

d'où

$$l(s,t) = \mathbb{E}\left[|t(X) - s(X)|\,|2\eta(X) - 1|\right] \quad . \tag{1.1}$$

Notons que s n'est pas nécessairement unique, son risque ne dépendant pas des valeurs qu'il prend sur $\left\{ x \in \mathcal{X} \text{ t.q. } \eta(x) = \frac{1}{2} \right\}$.

Au sujet de la classification, on consultera notamment la revue de Boucheron, Bousquet et Lugosi [**BBL05**].

Minimisation du risque empirique. Pour construire un prédicteur à partir des données, une méthode naturelle est de remplacer[20] dans la définition de s la distribution P inconnue par la distribution empirique

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i} \quad , \qquad \text{soit} \qquad \widehat{s}_S \in \arg\min_{t \in S} P_n\gamma(t,\cdot) = \arg\min_{t \in S_m} \left\{ \frac{1}{n}\sum_{i=1}^{n} \gamma\left(t, \xi_i\right) \right\} \quad .$$

Remarquons que l'on a remplacé ici l'ensemble $\mathcal{S}$ de tous les prédicteurs par un sous-ensemble S arbitraire (appelé *modèle*). En effet, en conservant la minimisation sur $\mathcal{S}$ tout entier, le minimum est atteint en toutes les fonctions qui valent Y_i en chacun des X_i (si les X_i sont tous distincts), si bien que $s(x)$ n'est pas défini de manière unique sur $\mathcal{X} \backslash \{ X_1, \ldots, X_n \}$. Le modèle $S = \mathcal{S}$ ne fournit donc pas de solution satisfaisante au problème de prédiction.

Si $s \notin S$, on définit

$$s_S \in \arg\min_{t \in S} P\gamma(t)$$

un minimiseur du risque dans S, auquel l'estimateur $\widehat{s}_S$ se compare plus naturellement. En effet, le risque de $\widehat{s}_S$ est nécessairement supérieur à celui de s_S, si bien qu'on peut le décomposer en la somme de deux termes positifs :

$$l(s, \widehat{s}_S) = l(s, s_S) + \left(P\gamma(\widehat{s}_S) - P\gamma(s_S) \right) \quad .$$

Le premier terme est appelé *biais*, il mesure la distance de s au modèle S. Le second terme est un terme de *variance*, il quantifie la difficulté d'estimation de s_S à l'aide de l'échantillon $(X_i, Y_i)_{1 \leq i \leq n}$.

Nous revenons maintenant sur les deux exemples de la régression et de la classification.

Exemple 1.3 (Régression, suite). En régression, on considère souvent des modèles qui sont des sous-espaces vectoriels[21] de $\mathcal{S}$. Des exemples classiques sont :

- lorsque $\mathcal{X} = [0,1]^k$, l'espace engendré par les premiers vecteurs d'une base de Fourier, d'ondelettes, *etc.*
- lorsque $\mathcal{X} \subset \mathbb{R}^k$, l'espace engendré par les projections sur un sous-ensemble de coordonnées. On essaie alors d'exprimer Y comme une combinaison linéaire d'une partie des variables contenues dans X, c'est la *sélection de variables*.
- étant donnée une partition $(I_\lambda)_{\lambda \in \Lambda}$ de $\mathcal{X}$, on appelle «modèle d'*histogrammes* associé à la partition $(I_\lambda)_{\lambda \in \Lambda}$» l'espace des fonctions $\mathcal{X} \mapsto \mathbb{R}$ constantes sur chacun des I_λ. On dispose alors d'une base $(\mathbb{1}_{I_\lambda})_{\lambda \in \Lambda}$, qui a l'avantage d'être orthogonale dans $L^2(\mu)$ pour toute mesure de probabilité μ sur $\mathcal{X}$. Par conséquent, il est particulièrement simple d'exprimer s_S et $\widehat{s}_S$ dans cette base :

$$s_S = \sum_{\lambda \in \Lambda} \beta_\lambda \mathbb{1}_{I_\lambda} \qquad \text{avec} \qquad \beta_\lambda = \mathbb{E}\left[Y \mid X \in I_\lambda \right]$$

[20] On peut donc considérer la minimisation du risque empirique comme une application de l'heuristique de rééchantillonnage : si l'on note $s = F(P)$, on a $\widehat{s}_{\mathcal{S}} = F(P_n)$.

[21] Ce qui suppose implicitement que $\mathcal{Y} = \mathbb{R}^l$ pour un entier $l \geq 1$; c'est le cas le plus courant, car on ne peut pas déterminer le support de Y avec un nombre fini de données.

$$\widehat{s}_S = \sum_{\lambda\in\Lambda} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \text{avec} \qquad \widehat{\beta}_\lambda = \frac{1}{\operatorname{Card}\{X_i \in I_\lambda\}} \sum_{X_i\in I_\lambda} Y_i$$

$$\text{en notant} \qquad p_\lambda := P(I_\lambda) = \mathbb{P}(X \in I_\lambda) \qquad \widehat{p}_\lambda := P_n(I_\lambda) = \frac{\operatorname{Card} X_i \in I_\lambda}{n} \ .$$

Dans la suite de cette thèse, nous considérerons régulièrement cet exemple, dans lequel les calculs sont plus aisés.

Un modèle S étant fixé, il existe de nombreux résultats sur le minimiseur du risque empirique. Voir par exemple le livre de Györfi, Kohler, Krzyżak et Walk [**GKKW02**].

EXEMPLE 1.4 (Classification binaire, suite). En classification, les modèles sont de la forme $\{\mathbb{1}_A \text{ t.q. } A \in \mathcal{A}\}$ pour un certain ensemble $\mathcal{A}$ de parties de $\mathcal{X}$. Des exemples classiques d'ensembles $\mathcal{A}$ sont les demi-espaces de $\mathcal{X} = \mathbb{R}^k$, les parties convexes de $\mathcal{X}$. Lorsque $\mathcal{X} = [0,1]$, on peut également considérer les *segmentations de $\mathcal{X}$ en k morceaux*, c'est-à-dire

$$\mathcal{A} := \left\{ \bigcup_{0\le i\le (k-1-\varepsilon)/2} [a_{2i+\varepsilon}, a_{2i+1+\varepsilon}] \text{ t.q. } 0 = a_0 < a_1 < \cdots < a_k = 1 \text{ et } \varepsilon \in \{0,1\} \right\} \ .$$

Il est délicat de définir une mesure de la complexité des modèles en classification. Une première mesure est la dimension de Vapnik-Červonenkis [**VC74**], qui se définit de la manière suivante :

$$V := \sup\left\{ N \in \mathbb{N} \text{ t.q. } m_{\mathcal{A}}(N) = 2^N \right\} < \infty$$

$$\text{avec} \qquad m_{\mathcal{A}}(N) := \sup_{C\subset\mathcal{X},\, \operatorname{Card}(C)=N} \{\operatorname{Card}(A\cap C) \text{ t.q. } A \in \mathcal{A}\} \ .$$

Son principal inconvénient est d'être indépendante de la distribution, et donc souvent trop pessimiste. D'autres mesures ont été introduites depuis, en termes d'entropie (entropie métrique, entropie à crochets, *etc.* ; voir notamment Tsybakov [**Tsy04**]) ou de processus de Rademacher (voir par exemple Koltchinskii [**Kol01**], Bartlett, Boucheron and Lugosi [**BBL02**], Bartlett et Mendelson [**BM02**]).

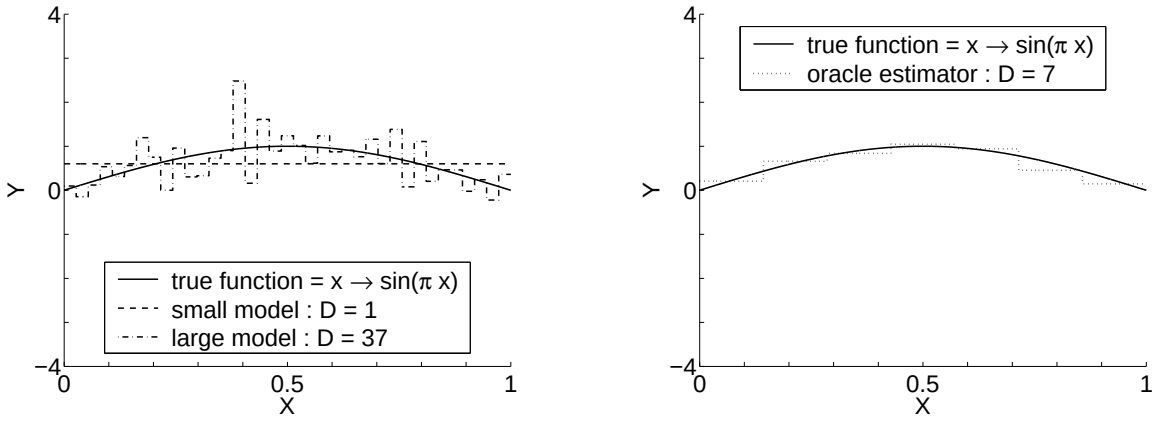
Comme en régression, de nombreux résultats existent sur le minimiseur du risque empirique pour un modèle donné. On consultera entre autres le livre de Lugosi [**Lug02**], et la revue de Boucheron, Bousquet et Lugosi [**BBL05**] pour les contributions les plus récentes. En particulier, on peut prouver dans différents cadres que le minimiseur du risque empirique a un risque optimal (à constante près), au sens du minimax, pour un choix adéquat de S (voir par exemple Massart et Nédélec [**MN06**]). C'est donc un bon candidat pour construire des procédures adaptatives.

Un cadre qui a reçu une attention particulière ces dernières années est celui où pour tout prédicteur t, la variance du processus $\gamma(t,\cdot)-\gamma(s,\cdot)$ peut être majorée en fonction de son espérance, *i.e.* de l'excès de risque $l(s,t)$. Une telle inégalité, appelée *condition de marge*, a été introduite par Mammen et Tsybakov [**MT99**] et peut s'écrire

$$\operatorname{var}_P\left(\gamma(t,\cdot) - \gamma(s,\cdot)\right) \le w\left(l(s,t)\right)$$

pour une fonction $w : (0,\infty) \mapsto (0,\infty)$, croissante, avec $x \mapsto w(x)/x$ décroissante. Par exemple, lorsque $|2\eta(X) - 1| \ge h > 0$ p.s., la condition de marge est satisfaite avec $w(\epsilon) = h^{-1/2}\epsilon$ (en utilisant (1.1)). Sous des hypothèses moins restrictives, Tsybakov [**Tsy04**] considère des fonctions w de la forme $w(\epsilon) = c\epsilon^\theta$ avec $\theta \in (0,1]$.

Lorsqu'une telle condition est satisfaite, on peut prouver que le risque du minimiseur du risque empirique décroît avec une *vitesse rapide*, *i.e.* plus rapide que la vitesse minimax globale en $n^{-1/2}$ (Massart et Nédélec [**MN06**] ; Tsybakov [**Tsy04**]). Une telle analyse repose sur l'idée de *localisation*, qui tient compte du fait que la variance du processus $\gamma(t,\cdot) - \gamma(s,\cdot)$ est faible en



(a) À éviter : $D = 1$ et $D = 37$ (b) Choix optimal : $D = 7$

FIG. 1.2. Un échantillon de taille $n = 200$ avec $s(x) = \sin(\pi x)$ et $\sigma = 1$: différents histogrammes réguliers pour estimer s.

$t = \widehat{s}_m$. On mesure alors la complexité du modèle S à l'aide d'une majoration du type

$$\forall u \in S, \forall \sigma > 0 \text{ t.q. } \phi_m(\sigma) \leq \sqrt{n}\sigma^2, \quad \sqrt{n}\mathbb{E}\left[\sup_{t \in S,\, \mathrm{var}_P[\gamma(u,\cdot)-\gamma(t,\cdot)] \leq \sigma^2} \{(P_n - P)(\gamma(u) - \gamma(t))\}\right] \leq \phi(\sigma) \ ,$$

où ϕ est croissante, positive, avec $x \mapsto \phi(x)/x$ décroissante (ce cadre, qui est celui de [**MN06**], est exposé plus en détails à la Sect. 7.2). La quantité importante est alors l'unique solution strictement positive $\epsilon^\star$ de l'équation $\sqrt{n}\epsilon^2 = \phi(w(\epsilon))$. Massart et Nédélec [**MN06**] ont ainsi montré que l'on peut majorer l'excès de risque du minimiseur du risque empirique sur S à l'aide de $\epsilon^\star$:

$$\forall x > 0, \quad \mathbb{P}\left(l(s, \widehat{s}_S) \leq 2l(s, S) + \kappa x (\epsilon^\star)^2\right) \geq 1 - e^{-x} \ ,$$

pour une constante absolue $\kappa > 0$.

1.2.2. Sélection de modèles.

Principe. D'après les résultats précédemment évoqués, le minimiseur du risque empirique est un bon candidat pour l'adaptativité, à condition que l'on soit capable de choisir convenablement le modèle S.

Par exemple, en régression, si l'on se limite aux modèles d'histogrammes réguliers (*i.e.* tels que les éléments I_λ de la partition sont tous de même taille), la question du choix de modèles revient à choisir le nombre D d'éléments de la partition. Ce choix est crucial. En effet, choisir $D = 1$ fournit un estimateur très peu sensible aux erreurs de mesures, mais de très faible qualité dès lors que s est loin d'être constant. Le défaut d'un tel modèle est son grand *biais*, mesuré par $l(s, S) = l(s, s_S)$ (autrement dit, la distance de s à S). À l'inverse, en choisissant D de l'ordre de n, on obtient un estimateur très sensible au bruit, et certainement très mauvais (si le niveau de bruit est non-nul) en prédiction, même si $s \in S_D$. Pour visualiser ce problème, on a représenté les minimiseurs du risque empirique sur des modèles d'histogrammes réguliers à la Figure 1.2. Les choix extrémaux $D = 1$ et $D = 37$ sont clairement à éviter, en comparaison avec le choix $D = 7$ (appelé *oracle*, avec la terminologie détaillée ci-dessous).

On peut mettre en lumière ce problème en calculant l'espérance du risque de $\widehat{s}_S$ dans le cas de la régression sur un design fixe (pour simplifier, mais le même phénomène se produit lorsque le design est aléatoire), avec le contraste des moindre carrés. En supposant le niveau de bruit σ

constant, et S un espace vectoriel de dimension D, on a

$$\mathbb{E}\left[l(s,\widehat{s}_S)\right] = l(s,S) + \frac{\sigma^2 D}{n} \ . \tag{1.2}$$

Le premier terme est le biais, le second terme est appelé *variance*. Il montre pourquoi choisir un modèle de grande dimension fournit en général un mauvais prédicteur. En d'autres termes, choisir un bon modèle S revient à trouver un bon *compromis entre le biais et la variance*.

Une description plus générale de la problématique de la sélection de modèles peut être trouvée dans le cours de Saint-Flour de Massart [**Mas07**]. En particulier, celle-ci ne se limite pas au cadre de la prédiction, mais concerne aussi l'estimation de densité ou encore le problème de l'identification. Étant donnée une famille d'estimateurs[22] $(\widehat{s}_m)_{m\in\mathcal{M}_n}$ (ou de résultats d'algorithmes), la question posée est celle du choix de m. On aimerait déterminer un bon modèle m à l'aide des observations uniquement, c'est-à-dire construire un estimateur $\widehat{m}(\xi_{1\dots n})$. Les $\widehat{s}_m$ peuvent être obtenus en minimisant le risque empirique sur une famille de modèles $(S_m)_{m\in\mathcal{M}_n}$, mais la problématique du choix de modèle ne se limite pas à ce cadre. Par exemple, en classification, on peut considérer la famille $(\widehat{s}_k)_{k\geq 1}$ la famille des estimateurs des k plus proches voisins (voir notamment Devroye et Wagner [**DW77, DW78**]), ou encore la famille des estimateurs par Support Vector Machines (*cf.* Scholkopf et Smola [**SS01**]) avec différents noyaux $(K_m)_{m\in\mathcal{M}_n}$. On consultera également le Chap. 7 du livre de Hastie, Tibshirani et Friedman [**HTF01**] à propos de la sélection de modèles en apprentissage. Pour le cadre plus spécifique de la régression, voir le livre de Györfi, Kohler, Krzyżak et Walk [**GKKW02**].

Sélection de modèles idéale.

Prédiction. Une procédure idéale pour la prédiction serait celle qui choisirait *l'oracle*

$$m^\star \in \arg\min_{m\in\mathcal{M}_n} \left\{ l(s,\widehat{s}_m) \right\} \ .$$

Remarquons que l'oracle $m^\star$ dépend de la vraie distribution P et des données. Ceci à deux conséquences principales. D'une part, l'oracle est inaccessible en pratique, c'est le choix *idéal* que l'on tente seulement d'approcher. D'autre part, l'oracle n'est pas forcément le «vrai» modèle $\widetilde{m}$, si celui-ci existe. Par exemple, si $s \in S_{\widetilde{m}}$ un modèle de grande dimension, avec une taille d'échantillon n petite ou un niveau de bruit σ élevé, $\widetilde{m}$ ne réalisera pas le compromis biais-variance à cause de son terme de variance[23]. Insistons également ici sur le fait qu'en prédiction, *on ne suppose pas que l'un des modèles est exact*, c'est-à-dire que $s \in \bigcup_{m\in\mathcal{M}_n} S_m$.

Une procédure de sélection de modèles est donc bonne lorsqu'elle a des performances comparables à celles l'oracle. Pour valider théoriquement une procédure, on cherche donc à montrer :

– L'*optimalité asymptotique* de $\widehat{m}$:

$$\mathbb{P}\left(\frac{l(s,\widehat{s}_{\widehat{m}})}{\inf_{m\in\mathcal{M}_n}\left\{ l(s,\widehat{s}_m) \right\}} \xrightarrow[n\to\infty]{} 1 \right) = 1 \ . \tag{1.3}$$

C'est le critère asymptotique le plus classique.

– Une *inégalité oracle* (non-asymptotique) :

$$\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right] \leq C \inf_{m\in\mathcal{M}_n} \left\{ \mathbb{E}\left[l(s,\widehat{s}_m) + R(m,n) \right] \right\} \ , \tag{1.4}$$

[22] La famille des indices $\mathcal{M}_n$ peut dépendre de n en toute généralité. C'est une des principales motivations pour le point de vue non-asymptotique que nous considérons ici.

[23] C'est pourquoi il ne faut pas surinterpréter le choix d'un modèle $\widehat{m}$ dans une procédure visant à une prédiction optimale. C'est simplement un modèle qui utilise au mieux les données disponibles pour prédire de nouvelles données.

pour une constante $C \geq 1$ (aussi proche de 1 que possible), et un terme de reste $R(m,n) \geq 0$ éventuellement aléatoire, dans la mesure du possible petit devant $l(s, \widehat{s}_m)$. Notons qu'une telle inégalité compare $\widehat{m}$ au meilleur choix déterministe de m, si bien qu'on pourrait trouver un cadre où elle serait satisfaite avec $C < 1$. C'est pourquoi nous préférons dans cette thèse l'inégalité oracle plus forte[24] suivante, qui compare $\widehat{m}$ à l'oracle :

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \leq C\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{l(s, \widehat{s}_m) + R(m,n)\right\}\right] \quad . \tag{1.5}$$

Notons toutefois que (1.5) est plus rarement considérée que (1.4), principalement parce qu'elle est plus difficile à démontrer.

– Une *inégalité oracle «trajectorielle»* : avec grande probabilité (par exemple $1 - Ln^{-2}$, où L est une constante),

$$l(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \left\{l(s, \widehat{s}_m) + R(m,n)\right\} \quad . \tag{1.6}$$

La différence avec (1.5) est que nous comparons ici $\widehat{m}$ à l'oracle sur un événement de grande probabilité. C'est donc une version non-asymptotique de l'optimalité asymptotique (1.3). Du point de vue de la prédiction, il est plus naturel de considérer (1.6), qui donne un résultat pour presque tout jeu de données, plutôt que (1.5) qui se limite à une comparaison en moyenne[25]. Remarquons également que (1.6) implique (1.5) lorsque la fonction de perte $l(s,t)$ est uniformément bornée par $B < \infty$, au prix d'un terme de reste supplémentaire BLn^{-2} (qui majore simplement l'espérance de $l(s, \widehat{s}_{\widehat{m}})$ sur l'événement défavorable).

Adaptativité. L'adaptation est également une qualité recherchée pour un estimateur $\widehat{s}_{\widehat{m}}$ obtenu par choix de modèles (voir par exemple Birgé et Massart [**BM97**]). De manière générale, on peut décrire l'objectif de l'adaptation comme suit. Supposons que la vraie distribution $P \in \mathcal{P} = \bigcup_{\alpha \in \mathcal{A}} \mathcal{P}_\alpha$, le paramètre inconnu α_0 tel que $\mathcal{P} \in \mathcal{P}_{\alpha_0}$ représentant une propriété de P (par exemple, l'ordre de la régularité hölderienne de s en régression). On dit alors qu'un estimateur $\widehat{s}_{\widehat{m}}$ est *adaptatif au paramètre* α s'il n'utilise pas la connaissance de α_0 tout en étant aussi bon (par exemple en termes d'erreur de prédiction) que tout estimateur $\widehat{s}_{\alpha_0}$ qui utiliserait α_0.

On peut notamment évaluer l'adaptativité d'un estimateur en comparant son risque $\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right]$ avec le *risque minimax*. Nous rappelons ici rapidement la définition de ce dernier. Étant donnée une famille $\mathcal{P}$ de lois de probabilités, le risque minimax de la famille $\mathcal{P}$ est

$$\mathcal{R}_{\min\max}(\mathcal{P}) := \inf_{\widetilde{s}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[l(s, \widetilde{s})\right] \quad ,$$

où l'inf est pris sur tous les estimateurs. Le risque minimax mesure donc le pire cas sur la classe $\mathcal{P}$, si bien qu'un estimateur minimax[26] pour une très grande famille $\mathcal{P}$ n'est pas nécessairement un bon estimateur en pratique. On dira donc d'un estimateur qu'il est *adaptatif au sens du minimax* si pour tout $\alpha_0 \in \mathcal{A}$, pour toute vraie distribution $P \in \mathcal{P}_{\alpha_0}$,

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \leq K\mathcal{R}_{\min\max}\left(\mathcal{P}_{\alpha_0}\right)$$

pour une constante[27] $K > 0$.

En classification, le risque minimax global (avec $\mathcal{P}$ la famille des lois de probabilité sur $\mathcal{X} \times \mathcal{Y}$ telles que $s \in S$ un modèle de dimension de Vapnik fini) décroît en $n^{-1/2}$, alors qu'avec une condition de marge supplémentaire sur les $P \in \mathcal{P}$, ce même risque peut atteindre des vitesses

[24] On a en effet l'implication (1.5) $\Rightarrow$ (1.4), la réciproque étant fausse en général.

[25] et donc ne peut pas détecter une éventuelle sous-optimalité de $\widehat{m}$ si celle-ci n'a lieu que pour des échantillons tels que $l(s, \widehat{s}_{m^\star})$ est bien plus petite que son espérance.

[26] *i.e.* un estimateur dont le risque est majoré par $L\mathcal{R}_{\min\max}(\mathcal{P})$ pour une constante numérique L.

[27] idéalement, une constante absolue ; en général, L dépend de α_0, mais jamais de P ni de la taille n de l'échantillon.

en $n^{-\alpha}$ avec $\alpha \in (1/2; 1]$ (Tsybakov [**Tsy04**], Massart et Nédélec [**MN06**]). C'est pourquoi l'on parle de «vitesses rapides». Construire un estimateur s'adaptant à la condition de marge est un problème de recherche de grand intérêt actuellement.

En régression, les quantités auxquelles on aimerait s'adapter son notamment la régularité de la fonction de régression et le niveau de bruit $\sigma : \mathcal{X} \mapsto [0, \infty)$. Ainsi, supposons que $\mathcal{X} = [0,1]^k$, σ est constante et s appartient à une boule de Hölder d'ordre α, *i.e.*

$$s \in \mathcal{H}(\alpha, R) := \{ f : \mathcal{X} \mapsto \mathbb{R} \text{ t.q. } \forall x_1, x_2 \in \mathcal{X}, \quad |f(x_1) - f(x_2)| \leq R \, \|x_1 - x_2\|_\infty^\alpha \} \ .$$

Alors, le risque minimax est donné par (Stone [**Sto80**])

$$\mathcal{R}_{\min\max} \left(\{ P \text{ t.q. } s \in \mathcal{H}(\alpha, R) \text{ et } \sigma \equiv \sigma_0 \} \right) = L_1(\alpha) \sigma_0^{\frac{4\alpha}{2\alpha+k}} R^{\frac{2k}{2\alpha+k}} n^{-2\alpha/(2\alpha+k)} \ ,$$

pour une constante $L_1(\alpha) > 0$. Par ailleurs, si $k = 1$ et σ n'est plus constante mais régulière, alors Galtchouk et Pergamenschikov [**GP05**] ont montré que

$$\mathcal{R}_{\min\max} \left(\left\{ P \text{ t.q. } s \in \mathcal{H}(\alpha, R), \, \sigma \text{ régulière et } \|\sigma\|_{L^2(\text{Leb})} \leq \sigma_0 \right\} \right) = L_2(\alpha) \sigma_0^{\frac{4\alpha}{2\alpha+1}} R^{\frac{2}{2\alpha+1}} n^{-2\alpha/(2\alpha+1)} \ ,$$

pour une constante absolue $L_2(\alpha) > 0$.

Identification. Déterminer le «vrai» modèle peut également être un objectif d'une procédure de sélection de modèles. On suppose alors que $s \in \bigcup_{m \in \mathcal{M}_n} S_m$, et l'on note $m^\star_{\text{ident}}$ le «vrai modèle», c'est-à-dire le modèle m le moins complexe[28] tel que $s \in S_m$.

Dans ce cadre, l'objectif est de déterminer $\widehat{m}$ tel que $\mathbb{P}(\widehat{m} = m^\star_{\text{ident}})$ soit maximale. Le pendant de l'optimalité asymptotique est alors la *consistance* :

$$\mathbb{P} \left(\widehat{m} = m^\star_{\text{ident}} \right) \xrightarrow[n \to \infty]{} 1 \ . \tag{1.7}$$

Incompatibilité. Ces trois objectifs ne sont pas toujours compatibles. En effet, Yang [**Yan05**] a montré qu'un estimateur ne peut pas être simultanément asymptotiquement minimax (à constante multiplicative près) et consistant. Autrement dit, on ne peut pas utiliser les mêmes méthodes pour l'adaptation et l'identification. Le plus souvent, on prouve l'adaptativité d'une procédure par le biais d'une inégalité-oracle (par exemple avec AIC ou le C_p de Mallows), en prenant une famille de modèles suffisamment riche. Dès lors, prédiction et identification sont incompatibles pour de telles familles de modèles. Des résultats similaires peuvent également être trouvés dans [**Shi86, FG94, LP05**].

Dans le cas de la sélection de modèles linéaires, Shao [**Sha97**] propose une classification des principales méthodes de sélection de modèles en trois classes. Ces méthodes fonctionnent asymptotiquement ou non (au sens de (1.3) ou (1.7)), selon qu'il existe ou non des modèles corrects de dimensions fixes. Notons toutefois qu'il existe des conditions dans lesquelles prédiction et identification reviennent au même (comme l'illustre la Prop. 1 de Shao [**Sha97**]). Il est également possible, dans de nombreuses situations, de combiner les avantages de AIC et BIC (Yang [**Yan03**], van Erven, Grünwald et de Rooij [**vEGdR07**], et les références citées par Yang [**Yan05**]).

Dans cette thèse, nous nous focaliserons essentiellement sur les objectifs de prédiction et d'adaptation, si bien que les procédures que nous établirons ne seront pas consistantes en termes d'identification. Nous verrons cependant dans la sous-section suivante que les critères de prédiction (tel AIC) et d'identification (tel BIC) sont reliés, si bien que l'on peut penser modifier un critère de prédiction pour obtenir un bon critère d'identification.

[28] Lorsque les modèles sont emboîtés, cette notion de complexité minimale ne pose pas de problème. En général, il se peut que $m^\star_{\text{ident}}$ ne soit pas unique.

Méthodes de choix de modèles. De même que l'on a remplacé la vraie loi P par la loi empirique P_n pour définir l'estimateur de minimisation du risque empirique, on pourrait utiliser la même substitution pour choisir un modèle :

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{ P_n \gamma (\widehat{s}_m) \} \quad .$$

Le défaut de ce critère (appelé *risque de resubstitution*, ou erreur d'entraînement) est qu'il sous-estime fortement le risque. Par exemple, si $\widehat{s}_m$ minimise dans S_m le risque empirique, cela conduirait systématiquement à choisir le plus grand modèle (celui qui explique le mieux les données, et non celui qui les prédit le mieux). Plus généralement, l'utilisation des mêmes données pour construire $\widehat{s}_m$ et mesurer son risque risque de conduire à une forte sous-estimation du risque.

Dans le cas de la régression homoscédastique sur un design fixe, on peut calculer l'espérance de ce critère (en notant D_m la dimension du modèle S_m comme espace vectoriel) :

$$\mathbb{E} [P_n \gamma (\widehat{s}_m)] = P \gamma (s_m) - \frac{\sigma^2 D_m}{n} \quad . \tag{1.8}$$

Celui-ci sous-estime donc le biais du modèle S_m, *a fortiori* le risque de $\widehat{s}_m$.

Validation. Pour éviter les écueils de l'erreur de resubstitution, une idée naturelle est de ne pas utiliser les même données pour construire les estimateurs $(\widehat{s}_m)_{m \in \mathcal{M}_n}$ et pour évaluer leurs risques. La manière la plus simple d'utiliser ce principe est de découper les données en deux échantillons. Avec le premier, appelé *échantillon d'entraînement*, on construit une famille d'estimateurs $\left(\widehat{s}_m^{(e)} \right)_{m \in \mathcal{M}_n}$ (par exemple en minimisant le risque empirique sur l'échantillon d'entraînement). Avec le second (l'*échantillon de validation*), on choisit le modèle qui minimise l'estimateur naturel du risque $P_n^{(v)} \gamma \left(\widehat{s}_m^{(e)} \right)$, où $P_n^{(v)}$ désigne la mesure empirique associée à l'échantillon de validation. Ainsi, le modèle

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n^{(v)} \gamma \left(\widehat{s}_m^{(e)} \right) \right\} \tag{1.9}$$

est obtenu en minimisant un estimateur sans biais du risque de $\widehat{s}_m^{(e)}$, qui doit être proche de celui de $\widehat{s}_m$.

Une telle méthode peut bien être généralisée à des découpages aléatoires successifs (qui relèvent tous de l'idée de sous-échantillonnage, donc de rééchantillonnage). Nous reviendrons sur ces procédures dans la Sect. 1.3.1.

Pénalisation. Une autre approche consiste à remarquer que l'erreur de resubstitution $P_n \gamma (\widehat{s}_m)$ ne sous-estime le risque qu'à cause d'un terme de complexité. Par exemple, en régression, ce biais est de l'ordre de $2\sigma^2 n^{-1} D_m$, où D_m désigne la dimension du modèle m. On peut donc chercher à obtenir un estimateur sans biais du risque $P \gamma (\widehat{s}_m)$ en ajoutant au risque empirique de $\widehat{s}_m$ «*pénalité*» :

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{ P_n \gamma (\widehat{s}_m) + \mathrm{pen}(m) \}$$

où $\mathrm{pen} : \mathcal{M} \mapsto \mathbb{R}$ mesure la complexité du modèle m. Ce dernier terme, le plus souvent positif, sert à pénaliser les grands modèles, qui s'adaptent «trop bien» aux données et sont donc inadaptés à la généralisation. On parle alors de minimisation d'un critère empirique pénalisé, ou encore de *pénalisation*.

Une telle procédure est également appelée *minimisation du risque structurel* par Vapnik [**Vap82**], la pénalité reflétant la structure du modèle m (*via* sa complexité). Cependant, une bonne pénalité doit également tenir compte des données (par exemple, en régression, à travers le nombre d'observations n et le niveau de bruit σ, étant donné (1.8)). En poussant ce raisonnement jusqu'au bout, ceci à conduit ces dernières années à l'introduction de *pénalités aléatoires*, calculée

en fonction des données (et pas uniquement à travers un estimateur de la variance $\widehat{\sigma^2}$). Une bonne partie des résultats de cette thèse étant reliés à des méthodes de pénalisation, nous reviendrons dessus plus en détails à la Sect. 1.2.3.

Agrégation. Une méthode concurrente de la sélection de modèles est l'*agrégation*, qui consiste à définir un estimateur agrégé de la forme

$$\widetilde{s}_{\text{agreg}} = \sum_{m \in \mathcal{M}_n} w_m(\xi_{1\dots n}) \widehat{s}_m \ ,$$

les poids $(w_m)_{m\in\mathcal{M}_n}$ étant de somme 1 et déterminés à partir des données. Voir par exemple Nemirovski [**Nem00**], et pour des résultats plus récents (notamment en classification, sous des hypothèses de marge) la thèse de Lecué [**Lec07a**].

L'inconvénient d'une telle méthode est qu'elle suppose généralement les estimateurs $\widehat{s}_m$ donnés, indépendants des observations $\xi_{1\dots n}$, si bien que les poids sont calculés avec des données indépendantes des estimateurs $(\widehat{s}_m)_{m\in\mathcal{M}_n}$. En pratique, ceci peut être fait en découpant préalablement l'échantillon en deux, la première moitié servant à construire les estimateurs, la seconde moitié à les agréger. Autrement dit, il faut recourir à la validation («hold-out»), et il ne semble pas évident de limiter l'influence du choix d'un découpage comme avec la validation croisée en sélection de modèles.

Ainsi, s'il peut être préférable d'agréger plutôt que de sélectionner un modèle (Lecué [**Lec07b**]), ce n'est que lorsque la famille $(\widehat{s}_m)$ est préalablement donnée. La nécessité de recourir à la validation laisse ouverte la question du choix optimal du découpage (au moins du point de vue pratique), et la comparaison avec une méthode de sélection de modèles n'est pas aisée. En effet, les résultats concernant l'agrégation comparent $\widetilde{s}_{\text{agreg}}$ au meilleur des $\widehat{s}_m^{(e)}$, qui sont construits avec seulement n_e données (la taille de l'échantillon d'entraînement), tout en supposant que n_v (la taille de l'échantillon de validation) est assez grande (proportionnelle à n). Il est alors difficile de savoir si une inégalité oracle avec constante 1 en agrégation est meilleure qu'une inégalité oracle avec constante $1+\epsilon_n$ en sélection de modèles, où l'on se compare au meilleur des $\widehat{s}_m$ (construits avec $n > n_e$ données, donc plus performants).

Notons toutefois que l'agrégation permet d'obtenir des procédures adaptatives à la condition de marge en classification binaire (Lecué [**Lec07a**]), chose qui n'a pu être prouvée pour une procédure de choix de modèles que dans le cas du hold-out (Blanchard et Massart [**BM06d**]). Aux constantes multiplicatives près, l'agrégation permet d'obtenir des procédures optimales dans un tel cadre.

L'objectif principal de cette thèse est précisément de décrire des procédures pour lesquelles ces constantes multiplicatives sont proches de 1, au moins d'un point de vue pratique. Nous ne nous contenterons donc pas de procédures de type validation, dont les faiblesses sont bien connues des praticiens (en particulier la sensibilité au choix d'un découpage de l'échantillon). Nous ne reviendrons pas sur les méthodes d'agrégation dans la suite, mais nous soulignons ici que la compréhension fine des méthodes de sous-échantillonnage devrait aider à élaborer des stratégies de découpage d'échantillons dans les procédures d'agrégation.

1.2.3. Pénalisation. Rappelons que l'idée de la pénalisation est de choisir un modèle

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \{ P_n\gamma(\widehat{s}_m) + \text{pen}(m) \} \tag{1.10}$$

où pen : $\mathcal{M} \mapsto \mathbb{R}$ est une «pénalité», en général positive, qui mesure la complexité du modèle m. Nous décrivons dans cette section la manière dont pen doit être choisie, et nous indiquons quelques pénalités classiques. Ceci dépendant de l'objectif à atteindre, nous distinguons les cas

de la prédiction et de l'identification (qui souffrent d'incompatibilités, comme nous l'avons déjà remarqué ; *cf.* Shao [**Sha97**] et Yang [**Yan05**]).

Prédiction. En prédiction, le critère idéal (que l'oracle $m^\star$ minimise) est $P\gamma(\widehat{s}_m)$. Il existe donc une *pénalité idéale*, qui est la différence entre ce critère et l'erreur de resubstitution :

$$\mathrm{pen}_{\mathrm{id}}(m) := P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m) \quad . \tag{1.11}$$

L'*heuristique de Akaike et Mallows* est qu'une bonne pénalité doit être *sans biais*, *i.e.* telle que

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{E}[\mathrm{pen}(m)] = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)] \quad .$$

Ceci a conduit notamment à l'introduction des critères FPE[29] (Akaike [**Aka70**]), AIC[30] (Akaike [**Aka73**]), SURE[31] (Stein [**Ste81**]) et C_p de Mallows (Mallows [**Mal73**]). Ce dernier, dans le cas de la régression homoscédastique sur un design fixe, est la somme du risque de resubstitution et de la pénalité

$$\mathrm{pen}_{\mathrm{Mallows}}(m) := \frac{2\sigma^2 D_m}{n} = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)] \quad . \tag{1.12}$$

La seconde égalité provient de la combinaison de (1.2) et (1.8). De tels critères sont alors asymptotiquement optimaux sous différentes hypothèses, comme cela a été prouvé successivement par Shibata [**Shi81**], Li [**Li87**] et Baraud [**Bar00, Bar02**]. On peut également considérer des pénalités plus générales, de la forme $\lambda_n \widehat{\sigma^2} D_m n^{-1}$, où $\widehat{\sigma^2}$ est un estimateur de σ^2. Shao [**Sha97**] les appelle GIC_{λ_n}[32] et montre que $\lambda_n \equiv 2$ fonctionne (asymptotiquement) en prédiction, $\lambda_n \to \infty$ fonctionne (asymptotiquement) en identification, et $2 < \lambda_n < \infty$ est un compromis entre ces deux méthodes. Les pénalités AIC, C_p et SURE peuvent également entrer dans le cadre général des *pénalités covariance* (Efron [**Efr04**]), qui estiment $2\,\mathrm{cov}\,(g(\widehat{s}_m(X_i)), Y_i)$ pour une fonction g dépendant du contraste utilisé pour mesurer le risque. Il est donc possible de les adapter à des contrastes autre que celui des moindre carrés ou la log-vraisemblance.

Depuis la mise en évidence du phénomène de *concentration de la mesure* (voir notamment Ledoux et Talagrand [**LT91**], Talagrand [**Tal96**] et Ledoux [**Led01**]), une analyse non-asymptotique des procédures de pénalisation a pu être développée, en particulier avec les travaux de Massart [**Mas07**]. Le point de départ en est le calcul suivant, qui utilise uniquement les définitions (1.10) de $\widehat{m}$ et (1.11) de $\mathrm{pen}_{\mathrm{id}}$. Pour tout modèle $m \in \mathcal{M}_n$,

$$\begin{aligned}
l(s, \widehat{s}_{\widehat{m}}) &= P_n\gamma(\widehat{s}_{\widehat{m}}) + \mathrm{pen}_{\mathrm{id}}(\widehat{m}) - P\gamma(s) \\
&= P_n\gamma(\widehat{s}_{\widehat{m}}) + \mathrm{pen}(\widehat{m}) + (\mathrm{pen}_{\mathrm{id}} - \mathrm{pen})(\widehat{m}) - P\gamma(s) \\
&\leq P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) + (\mathrm{pen}_{\mathrm{id}} - \mathrm{pen})(\widehat{m}) - P\gamma(s) \\
&= l(s, \widehat{s}_m) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(m) + (\mathrm{pen}_{\mathrm{id}} - \mathrm{pen})(\widehat{m}) \quad ,
\end{aligned}$$

soit

$$l(s, \widehat{s}_{\widehat{m}}) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(m) \} \quad . \tag{1.13}$$

Par conséquent, si $\mathrm{pen} \geq \mathrm{pen}_{\mathrm{id}}$ uniformément sur $m \in \mathcal{M}_n$, alors on a l'inégalité oracle

$$l(s, \widehat{s}_{\widehat{m}}) \leq \inf_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(m) \}$$

[29] Final Prediction Error : erreur de prédiction finale.
[30] Akaike Information Criterion : critère d'information d'Akaike.
[31] Stein's Unbiased Risque Estimate : estimateur non biaisé du risque.
[32] Generalized Information Criterion.

Ensuite, si la pénalité pen n'est pas beaucoup plus grande que $\mathrm{pen}_{\mathrm{id}}$, le terme de reste de cette inégalité oracle est négligeable. L'intérêt de ce calcul est de mettre en relief l'ordre des priorités pour la calibration d'une pénalité :

(1) Ne pas sous-pénaliser : $\mathrm{pen} \geq \mathrm{pen}_{\mathrm{id}}$ uniformément en $m \in \mathcal{M}_n$ (ou du moins pour les modèles susceptibles d'être sélectionnés).

(2) Dans le mesure du possible, ne pas trop sur-pénaliser : $\mathrm{pen} \leq (1+\epsilon)\,\mathrm{pen}_{\mathrm{id}}$ avec $\epsilon > 0$ petit.

Une stratégie naturelle est donc de prendre un estimateur sans biais de $\mathrm{pen}_{\mathrm{id}}$ comme pénalité, et de montrer qu'il est proche de $\mathrm{pen}_{\mathrm{id}}$ uniformément sur $m \in \mathcal{M}_n$, à l'aide d'inégalités de concentration. Ceci tend donc à justifier l'heuristique de Mallows dans un cadre assez général.

Cependant, la hiérarchie entre les inégalités $\mathrm{pen} \geq \mathrm{pen}_{\mathrm{id}}$ et $\mathrm{pen} \leq (1+\epsilon)\,\mathrm{pen}_{\mathrm{id}}$ indique les limites d'une telle heuristique.

D'une part, lorsqu'il y a *beaucoup de modèles*[33], une comparaison uniforme de pen et $\mathrm{pen}_{\mathrm{id}}$ est délicate. Il devient alors en général nécessaire de prendre une pénalité telle que $\mathbb{E}\left[\mathrm{pen}(m)\right] > \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$, afin de compenser les fluctuations de $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ uniformément en $m \in \mathcal{M}_n$. En régression, ceci a conduit à l'introduction de formes de pénalités plus générales que celle de Mallows. On consultera à ce sujet les travaux de Baron, Birgé et Massart [**BBM99**], Birgé et Massart [**BM01**] et Sauvé [**Sau06**]. Par exemple, lorsque la famille $\mathcal{M}_n$ contient $\binom{n}{D}$ modèles de dimension D, Baron, Birgé et Massart [**BBM99**] proposent d'utiliser une pénalité de la forme

$$\mathrm{pen}(m) = \frac{K_1 D_m}{n}\left(1 + K_2 \log\left(\frac{n}{D_m}\right)\right) \quad , \tag{1.14}$$

où K_1 et K_2 sont des constantes absolues à déterminer. Les travaux de Birgé et Massart [**BM06c**] ont montré que ce terme supplémentaire en $\log(n/D_m)$ est inévitable. Cette nécessité d'augmenter la pénalité peut être interprétée de la façon suivante. Tous les modèles de même dimension D ont une complexité similaire, et doivent donc être pénalisés de la même manière[34]. En écrivant $\mathrm{pen}(m) = \mathrm{pen}(D_m)$, la procédure (1.11) revient donc à minimiser le risque empirique sur chaque modèle agrégé

$$\widetilde{S}_D := \bigcup_{D_m = D} S_m \qquad \Rightarrow \qquad \widehat{s}_D \in \arg\min_{t \in \widetilde{S}_D} \left\{ P_n \gamma(t) \right\} \quad ,$$

puis pénaliser chaque modèle $\widetilde{S}_D$ en fonction de sa complexité. Lorsque le nombre de modèles de dimension D est grand, la complexité de $\widetilde{S}_D$ est clairement plus grande que celle de l'un des modèles S_m de dimension D. C'est pourquoi la pénalité (1.12) n'est plus suffisamment grande.

D'autre part, même lorsque la famille $m \in \mathcal{M}_n$ est de petite taille, il se peut qu'une pénalité donne de moins bons résultats qu'une pénalité légèrement plus grande (voir par exemple les simulations des Sect. 5.4 et 6.5). Ceci provient du fait que les fluctuations de $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ peuvent être importantes lorsque n est petit et σ grand, si bien qu'une pénalité sans biais risque de choisir un modèle de trop grande dimension avec une probabilité non négligeable. À notre connaissance, un tel phénomène n'a jamais été étudié en profondeur, alors qu'il nous semble important d'en tenir compte d'un point de vue non-asymptotique. Nous reviendrons sur cette question en conclusion.

Pénalités minimales et heuristique de pente. Récemment, Birgé et Massart [**BM01, BM06c**] se sont intéressés à la question de la calibration optimale d'une pénalité. Se plaçant dans le cas de la régression sur un design fixe avec un bruit gaussien homoscédastique, ils montrent qu'il existe une *pénalité minimale* $\mathrm{pen}_{\min}(m)$, satisfaisant les propriétés suivantes.

[33] par exemple, plus qu'un nombre polynomial cn^{α}, pour toutes constantes $c, \alpha > 0$.

[34] Notons que ceci n'est pas valable en général dans un cadre hétéroscédastique, voir Chap. 4.

(1) Si pen : $\mathcal{M}_n \mapsto [0,\infty)$ vérifie $\mathrm{pen}(m) \leq (1-\epsilon)\,\mathrm{pen}_{\min}(m)$ pour tout modèle $m \in \mathcal{M}_n$, avec $\epsilon > 0$, alors le modèle $\widehat{m}$ défini par (1.10) est avec grande probabilité de «grande dimension», et de risque bien plus grand que l'oracle, même si la cible s appartient à un modèle de petite dimension.

(2) Si $\mathrm{pen}(m) \approx 2\,\mathrm{pen}_{\min}(m)$ pour tout modèle $m \in \mathcal{M}_n$, alors la procédure (1.10) satisfait une inégalité oracle avec constante presque 1.

La conséquence du premier point est une illustration supplémentaire de la hiérarchie des inégalités dans la comparaison entre pen et $\mathrm{pen}_{\mathrm{id}}$. Sous-pénaliser peut avoir des conséquences dramatiques, alors que surpénaliser ne fait perdre qu'une constante multiplicative. En particulier, Birgé et Massart justifient la forme (1.14) de la pénalité lorsque la famille $\mathcal{M}_n$ est riche. D'autres minorations de la pénalité dans le cas où la variance σ^2 est inconnue ont été obtenus par Baraud, Giraud et Huet [**BGH07**].

Le second point (comparé au premier) a une conséquence encore bien plus intéressante pour la calibration de pénalités à l'aide des données. Se fondant sur l'*heuristique de pente*[35] selon laquelle $2\,\mathrm{pen}_{\min}$ est une pénalité optimale, Birgé et Massart proposent l'algorithme suivant (on s'est restreint ici au cadre où la pénalité minimale est linéaire en D_m ; Birgé et Massart [**BM06c**] proposent une forme de pénalité plus générale, mais toujours avec $\mathrm{pen}(m) = F(D_m)$) :

(1) Calculer $\widehat{m}(K) \in \arg\min_{m\in\mathcal{M}_n} \{ P_n\gamma(\widehat{s}_m) + KD_m \}$ pour tout $K > 0$.

(2) Déterminer $K_{\min} = \widehat{K}$ telle que $D_{\widehat{m}(K)}$ est «grande» lorsque $K < K_{\min}$, et $D_{\widehat{m}(K)}$ est «raisonnable» lorsque $K > K_{\min}$.

(3) Choisir $\widehat{m} = \widehat{m}\left(2\widehat{K}\right)$.

Une telle méthode peut également être appliquée lorsque la pénalité minimale est de la forme (1.14). Lebarbier [**Leb05**] a ainsi pu l'utiliser avec succès en détection de rupture, afin de calibrer les constantes K_1 et K_2. Dans le cas où la pénalité de Mallows (1.12) fonctionne, l'heuristique de pente permet d'estimer la variance σ^2, à travers le choix de la constante $\widehat{K} = \widehat{\sigma^2} n^{-1}$.

Le résultat théorique à la base de cette heuristique est la proximité des deux quantités suivantes :

$$p_1(m) := P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right) \approx P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right) =: p_2(m) \ . \tag{1.15}$$

Le terme de droite représente (en espérance) l'écart entre le biais $P\gamma\left(s_m\right)$ et l'erreur de resubstitution $P_n\gamma\left(\widehat{s}_m\right)$. En pénalisant moins que $p_2(m)$, on choisit $\widehat{m}$ suivant un critère inférieur au biais de S_m (en espérance). Le modèle sélectionné a donc tendance a être de grande dimension, d'où une explosion du risque. Le terme de gauche est l'écart entre le critère idéal (le risque) et le biais. La pénalité idéale (celle qui conduit à une sélection de modèles optimale) est donc — au moins en espérance — égale à $p_1(m) + p_2(m)$. L'heuristique (1.15) implique donc l'heuristique de pente $\mathrm{pen}_{\mathrm{opt}} = 2\,\mathrm{pen}_{\min}$, quelle que soit la forme de la pénalité minimale. C'est ainsi qu'au chapitre 3, nous avons étendu les résultats de Birgé et Massart dans un cadre hétéroscédastique, où la pénalité minimale ne dépend pas nécessairement de la dimension.

Prédiction en classification. Nous nous sommes pour l'instant concentrés sur le cas de la régression, où les méthodes de pénalisation sont plus simples à comprendre qu'en classification. Dans ce dernier cadre, on distingue deux types de pénalités.

[35] Le nom de cette heuristique est lié au cas où $\mathrm{pen}_{\min}(m)$ est linéaire en la dimension D_m. La constant $\widehat{K}$ obtenue par l'algorithme de Birgé et Massart peut alors être vue comme la valeur absolue de la pente de l'erreur de resubstitution $P_n\gamma\left(\widehat{s}_m\right)$, vue comme une fonction de D_m, qui est quasi-linéaire lorsque D_m est grand.

Les pénalités globales. sont des estimations de

$$\mathrm{pen}_{\mathrm{id,g}}(m) := \sup_{t \in S_m} \{ (P - P_n)(\gamma(t)) \} \geq (P - P_n)(\gamma(\widehat{s}_m)) = \mathrm{pen}_{\mathrm{id}}(m) \ .$$

Le terme «global» renvoie au fait que l'on considère un sup sur S_m tout entier, sans chercher à exploiter le fait que $\widehat{s}_m$ n'est pas situé n'importe où dans S_m.

De façon générale, lorsque $\mathrm{pen}(m) \geq \mathrm{pen}_{\mathrm{id},g}(m)$ pour tout modèle $m \in \mathcal{M}_n$, on peut montrer (Massart [**Mas07**], Thm. 8.1) que la procédure de choix de modèles qui en résulte satisfait une inégalité oracle de la forme

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right] \leq \inf_{m \in \mathcal{M}_n} \{ l(s, s_m) + \mathrm{pen}(m) \} + \frac{K}{\sqrt{n}} \ .$$

Lorsque S_m est une classe de Vapnik de dimension V_m, on peut par exemple prendre

$$\mathrm{pen}(m) = K \sqrt{\frac{V_m}{n}} \ ,$$

pour une constante absolue $K > 0$, auquel cas on peut obtenir une procédure adaptative à V du point de vue minimax global.

Ceci peut être amélioré en utilisant des mesures de complexité prenant en compte l'échantillon $\xi_{1\ldots n}$, par exemple l'entropie combinatoire de S_m (voir Massart [**Mas07**], Sect. 8.2.1) ou la *complexité de Rademacher* (introduite indépendamment par Koltchinskii [**Kol01**] et Bartlett, Boucheron et Lugosi [**BBL02**])

$$\widehat{R}_n(m) := \frac{1}{n} \mathbb{E}\left[\sup_{t \in S_m} \sum_{i=1}^{n} \epsilon_i \gamma(t, \xi_i) \,\middle|\, \xi_{1\ldots n} \right] \ ,$$

où $\epsilon_1, \ldots, \epsilon_n$ sont des variables de Rademacher (valant ± 1 avec même probabilité) indépendantes entre elles, et indépendantes de $\xi_{1\ldots n}$. D'autres pénalités globales ont également été introduites, notamment la complexité gaussienne, la contradiction maximale (ou maximal discrepancy, *cf.* Bartlett et Mendelson [**BM02**]), et les pénalités bootstrap globales (Fromont [**Fro04**]). Dans la mesure où celles-ci (de même que les complexités de Rademacher) relèvent d'une forme de rééchantillonnage, nous reviendrons dessus en Sect. 1.3.2.

Notons que malgré le fait que ces pénalités globales prennent en compte les données, elles ne permettent pas d'obtenir les vitesses de convergence rapides qui découlent de la condition de marge. Il est pour cela nécessaire d'utiliser l'approche locale.

Une pénalité locale. cherche à l'inverse à approcher directement $\mathrm{pen}_{\mathrm{id}}(m)$, en tenant compte de la position de $\widehat{s}_m$ dans S_m. Jusqu'à présent, les pénalités locales en classification sont toutes du type «*complexités de Rademacher locales*» (Bartlett, Mendelson et Philips [**BMP04**] ; Lugosi et Wegkamp [**LW04**] ; Bartlett, Bousquet et Mendelson [**BBM05**] ; Koltchinskii [**Kol06**]). En général, elles estiment par rééchantillonnage (avec des poids Rademacher i.i.d.) une majoration de $\mathrm{pen}_{\mathrm{id}}$ qui tient compte de la localisation (*cf.* Sect. 1.3.2). Au Chap. 7, nous proposons et étudions les propriétés de pénalités locales en classification, ne faisant pas intervenir de constantes inconnues. Nous conjecturons que de telles pénalités induisent une procédure optimale en termes de prédiction, qui serait donc adaptative à la condition de marge.

Identification. Lorsque l'on utilise la pénalisation pour identifier le «vrai» modèle $m^\star_{\mathrm{ident}}$ contenant la cible s, la pénalité idéale n'est bien sûr plus $\mathrm{pen}_{\mathrm{id}}$. En effet, l'objectif étant d'avoir $\widehat{m} = m^\star_{\mathrm{ident}}$ avec probabilité presque 1, une bonne pénalité doit empêcher de sélectionner un modèle m' de dimension $D_{m^\star_{\mathrm{ident}}} + 1$ et contenant également s. Avec un critère de type C_p, l'espérance

de la différence $\mathrm{crit}(m') - \mathrm{crit}(m^\star_{\mathrm{ident}})$ est $\sigma^2 n^{-1}$, qui est de l'ordre des fluctuations de cette même différence.

En revanche, en utilisant une pénalité telle que BIC[36] (Schwarz [**Sch78**]), qui s'écrit

$$\mathrm{pen}_{\mathrm{BIC}}(m) = \frac{\ln(n)\sigma^2 D_m}{n}$$

en régression, on obtient une procédure d'identification consistante. Voir notamment à ce sujet Shao [**Sha97**], Burnham et Anderson [**BA02**] ou encore Yang [**Yan06**] (et les références qu'ils contiennent). Parmi les autres pénalités classiques pour l'identification, on peut notamment citer MDL[37] (Rissanen [**Ris78**]).

La construction de pénalités pour identifier le vrai modèle n'est pas le thème central de cette thèse, et le résultat d'incompatibilité de Yang [**Yan05**] montre qu'en visant à l'adaptation, on obtient des procédures inconsistantes. En revanche, les pénalités BIC et et Mallows ne différant que par un facteur $\ln(n)/2$, on peut raisonnablement penser qu'une méthode de pénalisation fonctionnant dans le cadre de la prédiction peut être modifiée pour identifier le vrai modèle en multipliant la pénalité par un facteur proportionnel à $\ln(n)$. Une telle flexibilité n'est ici possible qu'avec la formulation en termes de pénalités, alors qu'une méthode directe de type validation croisée ne permet pas une telle modification.

1.2.4. Contributions de la thèse. Si cette thèse est principalement consacrée à l'étude de méthodes de rééchantillonnage, notamment en sélection de modèles, plusieurs résultats concernent la sélection de modèles de manière plus générale. Ils sont développés dans les Chap. 3 et 4.

Heuristique de pente. Le premier résultat (Chap. 3) concerne l'*heuristique de pente*, énoncée par Birgé et Massart [**BM06c**] sur la base de résultats limités au cas de la régression sur un design fixe, avec un bruit homoscédastique. Nous nous plaçons dans le cadre de la régression avec un design aléatoire et un *bruit hétéroscédastique*, et nous montrons que l'heuristique de pente reste valide dans le cas de modèles d'histogrammes. Ce résultat indique en particulier que cette heuristique ne se restreint pas à des pénalités de la forme $\widehat{K} D_m$, ni même à des fonctions de la dimension D_m.

Plus précisément, nous définissons la «pénalité minimale» suivante :

$$\mathrm{pen}_{\mathrm{min}}(m) = \mathbb{E}\left[P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right)\right] \quad .$$

Lorsque $\mathrm{pen} \leq (1-\epsilon)\,\mathrm{pen}_{\mathrm{min}}$ uniformément sur $m \in \mathcal{M}_n$, la dimension sélectionnée est de l'ordre de $n\ln(n)^{-1}$, tandis qu'elle est inférieure à $n^{1-\eta}$ avec $\eta > 0$ dès lors que $\mathrm{pen} \geq (1+\epsilon)\,\mathrm{pen}_{\mathrm{min}}$ (Thm. 3.2). Il y a donc un *saut de dimension* au voisinage de $\mathrm{pen} = \mathrm{pen}_{\mathrm{min}}$. De plus, un choix de pénalité $\mathrm{pen} \approx 2\,\mathrm{pen}_{\mathrm{min}}$ suffit pour obtenir une inégalité oracle (non-asymptotique) avec constante presque 1 (Thm. 3.1).

Si l'on connaît (ou si l'on a estimé) la forme $\mathrm{pen}_0(m)$ de la pénalité optimale à une constante multiplicative près, nos résultats suggèrent un algorithme pour choisir $\widehat{K}$ de telle sorte que $\mathrm{pen}(m) = \widehat{K}\,\mathrm{pen}_0(m)$ fournisse une procédure de choix de modèle optimale en prédiction (algorithme 3.1) :

(1) Estimer la forme optimale de pénalité $\mathrm{pen}_0(m)$

(2) Pour tout $K > 0$, calculer

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + K\,\mathrm{pen}_0(m) \right\}$$

[36] Bayesian Information Criterion.

[37] Minimum Description Length.

(3) Trouver $\widehat{K}_{\min} > 0$ correspondant au «saut de dimension»

(4) Choisir le modèle $\widehat{m} = \widehat{m}\left(2\widehat{K}_{\min}\right)$.

La seconde étape de cet algorithme peut être réalisée en temps polynomial en $\mathrm{Card}(\mathcal{M}_n)$ (voir Sect. 3.4), si bien que cette méthode peut réellement être utilisée en pratique.

Quant à l'estimation de la forme optimale de pénalité, les Chap. 5 à 8 nous suggèrent d'utiliser l'une des pénalités par rééchantillonnage (V-fold ou échangeable). En particulier, les théorèmes 5.1 et 6.1 justifient leur usage dans le cas des histogrammes. Nous conjecturons que cela reste possible dans un cadre bien plus général. Notons également que si l'on dispose d'un autre estimateur de pen_0, ou d'informations spécifiques indiquant quelle doit être la forme de la pénalité, les résultats du Chapitre 3 restent applicables, quelle que soit cette forme de pénalité.

Limites des pénalités linéaires. Le second résultat (Chap. 4) souligne l'une des difficultés induites par l'hétéroscédasticité du bruit. Nous montrons, en considérant un exemple particulier (mais où s et σ ont des formes très simples), qu'*une pénalité linéaire en la dimension D_m ne peut pas être asymptotiquement optimale en termes de prédiction* (Prop. 4.1). Ce résultat est fort, puisqu'il concerne également les pénalités de la forme $\widehat{K}(P_n, P)D_m$ qui utiliseraient les données et leur vraie distribution P.

Une manière d'éclairer les raisons de cette limite est de calculer explicitement l'espérance de la pénalité idéale, dans le cas d'un modèle d'histogrammes (Sect. 4.2) :

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(2+\delta_{n,p_\lambda}\right)\left(\mathbb{E}\left[\left(\sigma(X)\right)^2\,\middle|\, X\in I_\lambda\right] + \mathbb{E}\left[\left(s(X)-s_m(X)\right)^2\,\middle|\, X\in I_\lambda\right]\right)$$

avec $\lim_{np\to\infty}\delta_{n,p} = 0$. Celle-ci n'est donc pas linéaire en D_m lorsque le bruit est *hétéroscédastique*, ou lorsque la cible s est suffisamment *irrégulière* pour être éloignée de son approximation s_m par un histogramme (ce dernier terme n'apparaissant que parce que les X_i sont aléatoires).

Les résultats du Chapitre 4 soulignent ainsi l'intérêt d'avoir étendu l'heuristique de pente à des formes de pénalités quelconques, au Chapitre 3. Ils montrent surtout le besoin réel de pénalités capables de s'adapter à un bruit hétéroscédastique ou à l'aléa des X_i avec s irrégulière. Une solution possible à ce dernier point est apportée par les pénalités par rééchantillonnage (Chap. 5 à 8), en particulier à travers l'algorithme 11.1.

1.3. Sélection de modèles par rééchantillonnage

1.3.1. Validation croisée. Le principe de la sélection de modèles par validation — que nous avons déjà décrit, voir (1.9) — est le suivant. On découpe les données en un échantillon d'entraînement (e) de taille n_e et un échantillon de validation (v) de taille n_v, puis l'on choisit

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n}\left\{P_n^{(v)}\gamma\left(\widehat{s}_m^{(e)}\right)\right\} \quad .$$

Blanchard et Massart [**BM06d**] (voir aussi Massart [**Mas07**], Sect. 8.5) ont montré assez simplement que *la validation permet d'obtenir l'adaptation à la condition de marge*, fait qui est très délicat à prouver pour d'autres procédures. Ce résultat souligne l'écart entre la théorie (qui ne peut pas distinguer la validation de méthodes plus élaborées) et la pratique. En effet, la validation est connue pour être très variable, car elle repose sur le choix arbitraire d'un découpage (qui n'a pas de raison d'être dès lors que les données sont échangeables).

Pour réduire cette variabilité, un méthode consiste à réaliser successivement plusieurs découpages. On parle alors de *validation croisée*. Du point de vue théorique, il est alors beaucoup plus difficile d'étudier les performances d'une telle procédure, puisque l'on n'a plus deux jeux de données indépendants servant à deux tâches entièrement distinctes.

Nous décrivons dans cette section différentes stratégies relevant de ce même principe. S'il y a peu de résultats théoriques à leur sujet, en revanche de nombreuses comparaisons expérimentales ont été réalisées. On consultera par exemple Efron [**Efr83, Efr86**], Efron et Tibshirani [**ET97**], Zhang [**Zha93**] et Molinaro, Simon et Pfeiffer [**MSP05**]. L'un des objectifs de cette thèse est d'améliorer la compréhension théorique de ces procédés.

Découpages exhaustifs. Dans sa version initiale, la validation croisée (Allen [**All74**], Stone [**Sto74**], Geisser [**Gei75**]) ou *leave-one-out* est proche du jackknife. Elle consiste à retirer successivement chacune des données ξ_i, et l'utiliser comme échantillon de validation, le reste des données servant à l'entraînement :

$$\widehat{m}_{\mathrm{loo}} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \gamma\left(\widehat{s}_m^{(-i)}; \xi_i\right) \right\} \qquad \text{où} \quad \widehat{s}_m^{(-i)} \in \arg\min_{t\in S_m} \left\{ \frac{1}{n-1}\sum_{j\neq i} \gamma\left(t; \xi_j\right) \right\} \ .$$

On peut ensuite généraliser au *leave-p-out*, aussi appelé «delete-p CV», de même que le «delete-p jackknife» généralise le jackknife :

$$\widehat{m}_{\mathrm{lpo}} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{\binom{n}{p}} \sum_{I\subset\{1,\dots,n\},\,\mathrm{Card}(I)=p} P_n^{(I)} \gamma\left(\widehat{s}_m^{(I^c)}\right) \right\} \tag{1.16}$$

où $P_n^{(I)} := \frac{1}{\mathrm{Card}(I)}\sum_{j\in I}\delta_{\xi_j}$ et $\widehat{s}_m^{(I^c)} \in \arg\min_{t\in S_m}\left\{ P_n^{(I^c)}\gamma\left(t\right)\right\}$.

Dans le cas de modèles linéaires, Craven et Wahba [**CW79**] ont défini la *validation croisée généralisée*[38], qui est une version invariante par rotation de la validation croisée classique. En réalité, malgré son nom, la validation croisée généralisée est plus proche des critères C_p et C_L de Mallows que de la validation croisée elle-même (Efron [**Efr86**]).

Suivant la classification proposée par Shao [**Sha97**], on peut distinguer trois types de comportements asymptotiques pour le leave-p-out en fonction de p :

- lorsque $p \ll n$, le leave-p-out est asymptotiquement optimal (en prédiction), mais inconsistant pour l'identification. Lorsque $p = 1$, on trouve également ce résultat dans un article de Li [**Li87**], qui considère également la validation croisée généralisée ; les deux procédures sont asymptotiquement équivalentes au critère C_p de Mallows.
- lorsque $p \sim \lambda n$ avec $\lambda \in (0,1)$, le leave-p-out est asymptotiquement équivalent au critère FPE_α

$$\mathrm{crit}_{\mathrm{FPE}_\alpha} := P_n\gamma\left(\widehat{s}_m\right) + \alpha\frac{\widehat{\sigma^2} D_m}{n} \qquad \text{avec} \qquad \alpha = (2-\lambda)/(1-\lambda) > 2 \ ,$$

  $\widehat{\sigma^2}$ étant un estimateur du niveau de bruit (Zhang [**Zha93**]). Le leave-p-out surestime donc l'erreur de prédiction, sans toutefois être consistant pour l'identification. La raison en est que l'estimateur leave-p-out estime le risque d'un estimateur construit avec $n - p < n$ données, qui est donc plus grand que celui d'un estimateur construit avec n données. En estimation de densité, le résultat de van der Laan, Dudoit et Keles [**vdLDK04**] indique le même comportement.
- Lorsque $p \sim n$ et $n - p \to \infty$, le leave-p-out est consistant pour l'identification (Shao [**Sha93, Sha97**]).

Méthodes moins coûteuses. Lorsque le temps de calcul est pris en compte, il est rarement possible de réaliser un leave-one-out exhaustif, et encore moins un leave-p-out avec $p > 1$. Plusieurs autres méthodes ont alors été proposées, la première étant la validation croisée «V-fold» (Geisser

[38] Generalized cross-validation.

[**Gei75**]). L'idée est de découper l'échantillon en V blocs[39], et d'utiliser successivement chacun de ces blocs comme échantillon de validation. Autrement dit, $(B_j)_{1\leq j\leq V}$ étant une partition fixée de $\{1,\dots,n\}$, on choisit le modèle

$$\widehat{m}_{\mathrm{VFCV}} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{V}\sum_{j=1}^{V} P_n^{(B_j)}\gamma\left(\widehat{s}_m^{(-j)}\right) \right\} \qquad \text{où} \quad \widehat{s}_m^{(-j)} \in \arg\min_{t\in S_m} \left\{ P_n^{\left(B_j^c\right)}\gamma\left(t\right) \right\} \ .$$

Dans le cadre de l'estimation de densité, Celisse et Robin [**CR06**] ont montré que le choix arbitraire d'une partition $(B_j)_{1\leq j\leq V}$ induit une variabilité supplémentaire par rapport au V-fold (lorsque $p = n/V$), dont ils donnent une expression explicite.

Une telle procédure souffrant du même défaut que le leave-p-out pour $p \sim n/V$, Burman [**Bur89, Bur90**] a proposé une correction du biais de la validation croisée V-fold. Elle consiste à remplacer le critère V-fold classique (noté $\mathrm{crit}_{\mathrm{VFCV}}$) par

$$\mathrm{crit}_{\mathrm{corr.VFCV}}(m) := \mathrm{crit}_{\mathrm{VFCV}}(m) + P_n\gamma\left(\widehat{s}_m\right) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right) \ ,$$

qui est asymptotiquement non biaisé (à un terme d'ordre $(V-1)n^{-2}$ près). Notons que cette correction ressemble à une correction du biais par rééchantillonnage (*cf.* Hall [**Hal92**], Sect. 1.3 et 3.10).

Une autre approche est celle de méthodes d'apprentissage-test répété[40] (Breiman, Friedman, Olshen et Stone [**BFOS84**]). L'idée est d'utiliser (1.16) en ne considérant que B sous ensembles $I_1,\dots,I_B \subset \{1,\dots,n\}$ aléatoires indépendants, choisis uniformément parmi les sous-ensembles de taille p. En d'autres termes, on utilise une approximation de type Monte-Carlo pour le processus de sous-échantillonnage. Dans le cas de la régression linéaire, Zhang [**Zha93**] montre que cette méthode est asymptotiquement équivalente au leave-p-out, pourvu que $B \gg n^2$. Enfin, de même que la validation croisée V-fold, cette méthode peut être corrigée pour son biais (Burman [**Bur89**]). Sous certaines hypothèses (Burman [**Bur90**]), elle fournit alors une bonne estimation de l'erreur de prédiction pourvu que $n/(pB)$ reste borné quand $n \to \infty$.

Choix de V, p, n_v, etc. Une question importante reste celle du choix de la taille n_v de l'échantillon de validation (ou, de façon équivalente, le choix de p ou de V).

Pour la prédiction, au vu de [**Zha93, Sha97, vdLDK04**], il est nécessaire d'avoir $n_v \ll n_e$ pour obtenir l'optimalité asymptotique d'un estimateur par validation ou validation croisée, à moins d'utiliser l'une des corrections proposées par Burman [**Bur89**] (mais pour lesquelles aucun résultat en sélection de modèles n'a encore été prouvé, mis à part le Chap. 5 de cette thèse).

Malgré cela, on reproche souvent au leave-one-out sa variabilité (Hastie, Tibshirani et Friedman [**HTF01**], Sect. 7.10 ; voir aussi Breiman [**Bre96**]). En particulier, lorsqu'il est utilisé pour évaluer l'erreur d'algorithmes instables (k plus proches voisins, CART, minimisation du risque 0-1 sur des modèles très riches en classification, *etc.* ; voir Breiman et Spector [**BS92**]), le leave-one-out fournit un estimateur variable de l'erreur de prédiction. D'après Molinaro, Simon et Pfeiffer [**MSP05**], un tel défaut disparaît lorsque les algorithmes utilisés sont suffisamment stables. Pour corriger la variabilité du leave-one-out, plusieurs méthodes fondées sur le bootstrap ont été proposées par Efron [**Efr83**], auxquelles nous consacrons le paragraphe suivant.

Le choix de p repose donc sur un compromis entre le biais (qui est petit quand p est petit) et la variabilité (qui semble plus grande lorsque $p = 1$). En estimation de densité, un calcul explicite

[39] de tailles comparables, et choisis aléatoirement pour éviter tout biais lié à un ordre particulier des données.
[40] Repeated learning testing methods.

permet à Celisse et Robin [**CR06**] de définir un critère de choix de p, qui minimise la somme d'un terme de biais et d'un terme de variance. Voir aussi le Chap. 9 du livre de Politis, Romano et Wolf [**PRW99**] à ce sujet.

À l'inverse, pour l'identification, Shao [**Sha93**] a montré (dans un cadre de régression paramétrique) que la consistance du leave-p-out nécessitait $p \sim n$ et $n - p \to \infty$. Autrement dit, il faut prendre un échantillon d'entraînement de taille n_e négligeable devant celle de l'échantillon de validation, ce qui semble hautement contre-intuitif. Les simulations de Zhang [**Zha93**] indiquent le même phénomène dans le cas V-fold, puisque le vrai modèle est choisi plus souvent lorsque V est petit. C'est ainsi que Dietterich [**Die98**] puis Alpaydin [**Alp99**] ont proposé d'utiliser une version répétée de la validation croisée 2-fold pour identifier le meilleur de deux algorithmes. Ce phénomène est appelé le *paradoxe de la validation croisée* par Yang [**Yan06, Yan07**], qui l'étudie en régression et en classification lorsqu'il n'y a que deux procédures à comparer. Il donne alors des conditions suffisantes plus générales sur n_e et n_v pour que la validation (ou la validation croisée) soit consistante. Lorsque les deux procédures ont une vitesse de convergence non-paramétrique, il s'avère que les conditions énoncées par Shao ne sont plus toujours nécessaires, et l'on peut avoir $n_e > n_v$.

Dans le cas de la validation croisée V-fold, une complication supplémentaire vient du fait que V gouverne simultanément la taille de $n_v = n/V$, la variabilité de l'estimateur (lorsque V est petit, la variabilité diminue avec V qui correspond au B des méthodes d'apprentissage-test répété ; ce n'est plus toujours le cas lorsque $V \approx n$, où l'on peut retrouver les défauts déjà évoqués du leave-one-out) et le temps de calcul. D'après les calculs asymptotiques de la variance de l'estimateur V-fold (corrigé ou non) de Burman [**Bur89**], il faut cependant choisir V aussi grand que possible, du point de vue asymptotique. La valeur optimale de V (si le temps de calcul ne pose pas de problème) doit donc tendre vers l'infini avec n, mais sans être trop proche de n. L'estimation précise de la variance du critère V-fold $\mathrm{crit}_{\mathrm{VFCV}}$ est cependant délicate, Bengio et Grandvaler [**BG04**] ayant montré qu'il n'en existe pas d'estimateur universellement non-biaisé.

D'un point de vue pratique, le temps de calcul doit rentrer en ligne de compte. Il semble que bien des praticiens considèrent «$5 \leq V \leq 10$ est optimal» comme une règle quasi-universelle[41], en particulier en classification où le leave-one-out demande un temps de calcul plus important au risque d'une plus grande variabilité.

Le bootstrap .632. De même que l'on peut utiliser le bootstrap pour stabiliser un algorithme[42], on peut l'utiliser pour diminuer la variabilité du leave-one-out en classification. C'est le *leave-one-out bootstrap*, défini de la façon suivante. Pour tout $i \in \{1, \dots n\}$, en retirant la donnée i de l'échantillon $\xi_{1\dots n}$, on définit $\left(W_j^{(i)}\right)_{j \neq i}$ un vecteur de poids bootstrap indépendant de $\xi_{1\dots n}$ et on note $\widehat{s}_m^{W(i)}$ le minimiseur du risque empirique bootstrap correspondant. L'estimateur leave-one-out bootstrap s'écrit alors

$$\mathrm{crit}_{\mathrm{loo\,boot}}(m) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\gamma\left(\widehat{s}_m^{W(i)}; \xi_i \right) \,\middle|\, \xi_{1\dots n} \right] \quad .$$

[41] La recommandation $V = 2$ de Dietterich [**Die98**] n'a pas été suivie par les praticiens, car elle conduit à un choix trop conservatif, son erreur de première espèce étant basse (l'identification du meilleur de deux algorithmes est en effet une procédure de test). Ainsi, elle ne permet que trop rarement de montrer la supériorité d'une nouvelle procédure par rapport aux méthodes préexistantes...

[42] en remplaçant la sortie $A(\xi_{1\dots n})$ par la moyenne des sorties $A(\xi^\star_{1\dots n})$ obtenues avec différents échantillons bootstrap.

C'est une version régularisée de l'estimateur leave-one-out

$$\mathrm{crit}_{\mathrm{loo}}(m) := \frac{1}{n}\sum_{i=1}^{n} \gamma\left(\widehat{s}_m^{(-i)}; \xi_i\right) \quad ,$$

puisque l'on a remplacé chaque terme par une moyenne sur les rééchantillons bootstrap de $(\xi_j)_{j\neq i}$. L'inconvénient du leave-one-out bootstrap est qu'il estime sans biais le risque d'un estimateur construit avec environ $(1-e^{-1})n$ données en moyenne, donc $\mathrm{crit}_{\mathrm{loo\,boot}}(m)$ surestime le risque $P\gamma(\widehat{s}_m)$. Efron [**Efr83**] a donc proposé l'*estimateur* .632 qui corrige ce biais en utilisant le fait que l'erreur de resubstitution sous-estime le risque :

$$\mathrm{crit}_{.632}(m) := \omega\,\mathrm{crit}_{\mathrm{loo\,boot}}(m) + (1-\omega)P_n\gamma(\widehat{s}_m) \quad , \tag{1.17}$$

avec $\omega = 1-e^{-1} \approx 0.632$. Il apparaît cependant que l'estimateur .632 sous-estime légèrement le risque. C'est pourquoi Efron et Tibshirani [**ET97**] ont introduit l'estimateur .632+, qui a la forme (1.17) avec $\omega > .632$ calculé à l'aide du «taux d'erreur non-informatif»[43]. Le principal inconvénient de ces deux estimateurs est, comme le reconnaissent Efron et Tibshirani, la faiblesse des arguments théoriques qui sous-tendent leur construction. Des études de simulation montrent en revanche la qualité des performances de l'estimateur .632+ dans différents cadres (Efron et Tibshirani [**ET97**], Molinaro, Simon et Pfeiffer [**MSP05**]).

1.3.2. Pénalités par rééchantillonnage. Notons que la distinction entre critères pénalisés et critères non-pénalisés est un peu artificielle, puisque l'on peut toujours écrire

$$\mathrm{crit}(m) = P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \qquad \text{avec} \qquad \mathrm{pen}(m) = \mathrm{crit}(m) - P_n\gamma(\widehat{s}_m) \quad .$$

Par exemple, les critères .632 et .632+ peuvent être vus comme des critères pénalisés[44], avec la pénalité

$$\mathrm{pen}_{.632}(m) := \omega\left(\mathrm{crit}_{\mathrm{loo\,boot}}(m) - P_n\gamma(\widehat{s}_m)\right) \quad .$$

Le choix $\omega > .632$ du critère .632+ s'interprète alors comme une forme de surpénalisation.

Dans le cas de critères par rééchantillonnage, cette distinction peut être fondée sur les quantités que l'on cherche à estimer par rééchantillonnage. D'une part, les critères de type «validation croisée» de la Sect. 1.3.1 cherchent à estimer directement l'erreur de prédiction $P\gamma(\widehat{s}_m)$. Utiliser le bootstrap de cette manière conduirait à minimiser le critère

$$\mathbb{E}_W\left[P_n\gamma\left(\widehat{s}_m^W\right)\right] \quad ,$$

mais celle-ci donne d'assez mauvais résultats d'après Efron [**Efr83**] (Sect. 8), qui les explique par un fort biais.

D'autre part, les pénalités par rééchantillonnage cherchent à estimer la pénalité idéale $(P-P_n)\gamma(\widehat{s}_m)$ (ou son espérance, ou son majorant $\mathrm{pen}_{\mathrm{id},g}$). Une quantité très semblable[45] est également appelée «optimisme» par Efron [**Efr83**]. Cette seconde approche tient donc plus de l'idée de rééchantillonnage itéré (Hall [**Hal92**], Sect. 1.4), qui permet de diminuer le biais d'un ordre de grandeur.

Pénalités bootstrap. Efron [**Efr83**] propose ainsi la *pénalité bootstrap*

$$\mathrm{pen}_{\mathrm{Efron}}(m) := \mathbb{E}_W\left[P_n\gamma\left(\widehat{s}_m^W\right) - P_n^W\gamma\left(\widehat{s}_m^W\right)\right] \quad ,$$

[43] no-information error rate. Voir la section 3 de [**ET97**] pour une définition précise de ω.
[44] c'est d'ailleurs de cette manière que l'introduit Efron [**Efr83**], Sect. 6.
[45] $\mathbb{E}\left[P\gamma(\widehat{s}_m)\right] - P_n\gamma(\widehat{s}_m)$

qui a une faible variabilité, mais peut sous-estimer fortement la pénalité idéale. Le même type de pénalité a été proposée, lorsque le contraste est mesuré par la log-vraisemblance, par Ishiguro, Sakamoto et Kitagawa [**ISK97**], qui l'appellent EIC (ce critère généralisant WIC proposé précédemment par Ishiguro et Sakamoto [**IS91**]). Toujours avec la log-vraisemblance, pour la sélection de modèles à espaces d'états, Cavanaugh et Shumway [**CS97**] ont proposé le critère AICb suivant :

$$\mathrm{pen}_{\mathrm{AICb}}(m) := 2\mathbb{E}_W \left[P_n \left(\gamma(\widehat{s}_m^W) - \gamma(\widehat{s}_m) \right) \right] \ .$$

Il s'agit d'une estimation par rééchantillonnage de $2P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right)$, qui est proche de $\mathrm{pen}_{\mathrm{id}}(m)$ d'après l'heuristique de pente (Birgé et Massart [**BM06c**]). L'équivalence asymptotique de AIC et des pénalités bootstrap, WIC, EIC et AICb a été prouvée par Shibata [**Shi97**], dans le cas de la log-vraisemblance. En particulier, toutes ces procédures sont alors asymptotiquement optimales.

Pour corriger le biais de $\mathrm{pen}_{\mathrm{Efron}}$, Efron [**Efr83**] a proposé une *pénalité bootstrap double*, qui s'écrit

$$\mathrm{pen}_{\mathrm{Efron\,double}}(m) = \mathrm{pen}_{\mathrm{Efron}}(m) + \mathrm{corr}(m) \ ,$$

où $\mathrm{corr}(m)$ est une estimation bootstrap de $\mathrm{pen}_{\mathrm{id}}(m) - \mathrm{pen}_{\mathrm{Efron}}(m)$. Il s'agit ici de bootstrap itéré proprement dit. Le principal inconvénient d'une telle méthode est évidemment le temps de calcul, puisque pour chaque échantillon bootstrap simulé, il faut considérer B sous-échantillons bootstrap, avec B aussi grand que possible.

D'autres pénalités bootstrap, de type «pénalité covariance», ont été proposées par Efron [**Efr04**]. Celles-ci apparaissent comme des versions «stabilisées» du leave-one-out, ce qui propose une nouvelle illustration de la variabilité de ce dernier.

Pour l'identification, Shao [**Sha96**] a montré que la pénalité bootstrap $\mathrm{pen}_{\mathrm{Efron}}$ est inconsistante, dans le cadre de la régression avec un critère des moindre carrés. En revanche, si l'on utilise le bootstrap «m out of n» avec $n \gg m \to \infty$, la même pénalité se révèle alors consistante pour l'identification, résultat que l'on peut rapprocher de celui sur le leave-p-out (Shao [**Sha93**]). Dans le cadre de la sélection de variable (en régression sur un design fixe), on consultera également Breiman [**Bre92**], qui définit une procédure appelée «little bootstrap».

Dans la mesure où le jackknife est une approximation du bootstrap (Efron [**Efr79**]), on peut également construire une *pénalité jackknife*, qui est alors une approximation quadratique de $\mathrm{pen}_{\mathrm{Efron}}$. Efron [**Efr83**] la définit comme suit :

$$\mathrm{pen}_{\mathrm{jackknife}}(m) := \frac{1}{n} \sum_{i=1}^{n} \left[\gamma\left(\widehat{s}_m^{(-i)}; \xi_i \right) - P_n \gamma\left(\widehat{s}_m^{(-i)} \right) \right] \ .$$

Le critère qui en résulte est alors très proche du critère leave-one-out classique.

L'un des principaux apports de cette thèse consiste dans une étude non-asymptotique d'une famille de pénalités qui généralise la pénalité bootstrap d'Efron. C'est l'objet des Chap. 5 à 8. Nous décrivons ces pénalités et les résultats prouvés à leur sujet dans la sous-section suivante.

Classification. Dans le cadre de la classification binaire, les pénalités définies ci-dessus peuvent bien évidemment être utilisées. C'est d'ailleurs dans ce cadre qu'Efron les introduit [**Efr83**]. En particulier, l'estimateur .632 a été construit pour résoudre le problème de la variabilité du leave-one-out (voir notamment Efron [**Efr83**], remarque F).

Il est également possible de définir une *pénalité bootstrap randomisée* (Efron [**Efr83**]), comme suit. Au lieu de rééchantillonner à partir de l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ avec des poids uniformes, on rééchantillonne à partir de $(X_1, Y_1), (X_1, 1 - Y_1), \ldots, (X_n, Y_n), (X_n, 1 - Y_n)$, en

donnant un poids $n^{-1}\pi(X_i, \xi_{1\dots n}) < n^{-1}$ à (X_i, Y_i), et un poids $n^{-1}(1 - \pi(X_i, \xi_{1\dots n})) > 0$ pour $(X_i, 1 - Y_i)$. On peut soit choisir des poids $\pi(X_i, \xi_{1\dots n}) \equiv 0.9$, soit les choisir à partir des données (*cf.* [**Efr83**]).

D'autres pénalités par rééchantillonnage ont été introduites depuis la fin des années 90, dans le cadre de la théorie de l'apprentissage statistique. Celles-ci ont notamment vocation à s'appliquer au problème de la classification binaire.

Ainsi, Koltchinskii [**Kol01**] et Bartlett, Boucheron, Lugosi [**BBL02**] ont introduit indépendamment la *pénalité de Rademacher globale*

$$\mathrm{pen}_{\mathrm{Rad,g}}(m) := \widehat{R}_n(m) = \frac{1}{n}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n}\epsilon_i\gamma(t,\xi_i)\,\middle|\,\xi_{1\dots n}\right] ,$$

avec $\epsilon_1, \dots, \epsilon_n$ des variables de Rademacher (*i.e.* uniformes sur $\{-1, 1\}$) i.i.d. indépendantes de $\xi_{1\dots n}$. Comme l'a remarqué Koltchinskii [**Kol01**], cette pénalité est un cas particulier de pénalité par rééchantillonnage bootstrap à poids échangeables. Suivant cette remarque, Fromont [**Fro03, Fro07**] a construit des *pénalités bootstrap globales* de la forme

$$\mathrm{pen}_{\mathrm{boot,g}}(m) := \frac{1}{n}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n}Z_i\gamma(t,\xi_i)\,\middle|\,\xi_{1\dots n}\right] = \mathbb{E}_W\left[\sup_{t\in S_m}\left\{(P_n - P_n^W)\gamma(t)\right\}\right]$$

avec $Z_i = 1 - W_i$. La deuxième formulation souligne que de telles pénalités ne sont rien d'autre que des estimations par rééchantillonnage de la pénalité globale idéale $\mathrm{pen}_{\mathrm{id},g}(m)$. Fromont considère deux cas :

- $Z_1, \dots, Z_n$ sont i.i.d. symétriques, indépendantes de $\xi_{1\dots n}$ (par exemple, des variables de Rademacher)
- $Z_1, \dots, Z_n$ échangeables de somme p.s. égale à 0. Par exemple, le bootstrap, où les poids $W_1, \dots, W_n$ sont multinomiaux.

Fromont [**Fro03, Fro07**] prouve alors des inégalités oracle non-asymptotiques pour les procédures fondées sur de telles pénalités, qui généralisent les résultats précédemment prouvés sur les pénalités de Rademacher globales. En particulier, ces pénalités sont adaptatives à la dimension de Vapnik du point de vue minimax global.

La grande attention portée aux poids Rademacher plutôt qu'à des poids de rééchantillonnage plus généraux a pour cause la possibilité d'utiliser des outils de symétrisation tels que le Lemme 1.1. Les résultats de Fromont reposent sur une généralisation du Lemme 1.1 pour les deux types de poids qu'elle considère[46]. La constante 2 est alors remplacée par $1/\mathbb{E}(Z_1)_+$.

Les résultats théoriques ne justifient donc l'usage de telles pénalités que si on les multiplie par ce facteur $1/\mathbb{E}(Z_1)_+ > 1$, alors que celui-ci semble inutile en pratique (Lozano [**Loz00**], Fromont [**Fro07**]). Nous montrons au Chap. 9 que ce facteur est nécessaire du point de vue théorique, si bien que la calibration de telles constantes peut poser problème dans certains cas extrêmes. Cependant, les contre-exemples considérés au Chap. 9 laissent penser que ces limitations ne se produisent que dans les cas «raisonnables». Un résultat théorique avec de bonnes constantes pourrait donc être prouvé, au prix de l'ajout d'hypothèses minimes (par exemple, le meilleur classifieur du modèle n'est pas excessivement bon : $P\gamma(s_m) > n^{-1}$).

Approche locale. Cependant, comme ces pénalités globales estiment un majorant large de la pénalité idéale, elles ne permettent pas d'obtenir l'adaptation dans un cadre où une condition de marge (décrite dans l'exemple 1.4) est vérifiée. Pour tenir compte du fait que le minimiseur

[46] Le deuxième ingrédient dans les preuves de Fromont est l'inégalité de concentration de McDiarmid, qui fonctionne car le contraste γ est borné.

du risque empirique $\widehat{s}_m$ n'est pas situé n'importe où dans S_m, diverses pénalités locales ont été introduites (Bartlett, Mendelson et Philips [**BMP04**] ; Lugosi et Wegkamp [**LW04**] ; Bartlett, Bousquet et Mendelson [**BBM05**] ; Koltchinskii [**Kol06**]). On peut généralement les définir[47] à l'aide de points fixes de fonctions de la forme

$$\widehat{f}(r) = \mathbb{E}\left[\sup_{t\in S_m,\, c_1 r \leq P_n(\gamma(t)-\gamma(\widehat{s}_m))\leq c_2 r} \left\{ \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \gamma(t;\xi_i) \right\} \middle|\ \xi_{1\ldots n} \right] \ .$$

Il s'agit encore une fois d'estimation par rééchantillonnage (avec des poids Rademacher (1/2), en utilisant la terminologie introduite en Sect. 1.1) du point fixe $r^\star$ de

$$f(r) = \sup_{t\in S_m,\, c_1 r \leq l(s,t)\leq c_2 r} \left\{ (P-P_n)\left(\gamma(t)-\gamma(s)\right) \right\} \ ,$$

qui est une mesure de complexité locale (avec $c_1 \geq 0$, $c_2 > 0$ et au moins une constante multiplicative devant $r^\star$ à calibrer[48] ; éventuellement, $c_1 = 0$). Une vision unifiée de ces *pénalités de Rademacher locales* est donnée par l'article de Koltchinskii [**Kol06**] et la discussion qui le suit.

Ce n'est cependant pas la seule stratégie envisageable pour construire des pénalités tenant compte du phénomène de localisation. Dans la discussion de [**Kol06**], van de Geer [**vdG06**] a ainsi proposé une *pénalité construite par sous-échantillonnage* (de type «validation» simple), et utilisant l'expression exacte de la fonction w intervenant dans la condition de marge. Cette pénalité ne peut donc pas être utilisée en pratique, mais elle ouvre une nouvelle piste en direction de pénalités adaptatives à la condition de marge.

Rappelons également que la validation, telle quelle, permet d'obtenir l'adaptation à la condition de marge (Blanchard et Massart [**BM06d**]). Ce sont principalement les défauts de cette méthode en pratique qui rendent nécessaire la recherche de pénalités locales précises. En l'état actuel des recherches, aucune des deux stratégies évoquées ci-dessus ne semble suffisamment performante pour un praticien.

Enfin, on peut considérer la pénalité bootstrap $\mathrm{pen}_{\mathrm{Efron}}$ introduite par Efron comme un bon candidat à l'adaptation à la condition de marge. Elle tient en effet compte de la position de $\widehat{s}_m$ dans le modèle S_m en considérant le processus $(P_n - P_n^W)\gamma(t)$ en $t = \widehat{s}_m^W$. En comparaison avec les pénalités de Rademacher locales, elle est non seulement *beaucoup plus simple à calculer* (sans parler des problèmes de calibration, largement en faveur de la pénalité bootstrap), et en principe plus précise puisqu'elle cherche à estimer directement $\mathrm{pen}_{\mathrm{id}}$, et non son majorant $r^\star$.

L'objet des Chap. 5 à 8 de cette thèse est d'étudier les propriétés d'une grande classe de pénalités, dont $\mathrm{pen}_{\mathrm{Efron}}$ n'est qu'un exemple. Le Chap. 7 considère en particulier le cas de la classification. Le résultat de concentration qui y est prouvé encourage à poursuivre les recherches dans cette direction.

1.3.3. Contributions de la thèse.

Un nouveau regard sur la validation croisée V-fold. Tout d'abord, le Chap. 5 propose un nouveau regard sur la validation croisée V-fold. Comme nous l'avons indiqué dans la Sect. 1.3.1, l'un des défauts de celle-ci est son biais, car elle estime l'erreur de prédiction d'un estimateur construit avec $n(V-1)/V$ données au lieu de n. Dans le cadre de la régression sur des histogrammes, on

[47] voir [**Kol06**] pour une définition précise et rigoureuse.

[48] dans de «bonnes» situations (*e.g.* Koltchinskii [**Kol06**], Sect. 6.1, où l'on suppose que s appartient à l'un des modèles), on dispose de très grandes bornes supérieures sur ces constantes ; en général, la pénalité fondée sur $r^\star$ dépend de quantités inconnues telles que la fonction w intervenant dans la condition de marge (*e.g.* Koltchinskii [**Kol06**], Sect. 5.2).

peut alors calculer explicitement l'espérance du critère correspondant (Prop. 5.1) :

$$\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m)\right] = P\gamma(s_m) + \frac{V}{(V-1)n} \sum_{\lambda \in \Lambda_m} \left(1 + \delta_{n,p_\lambda}^{(V)}\right) \left(\mathbb{E}\left[\sigma(X)^2 + (s - s_m)(X)^2 \,\middle|\, X \in I_\lambda\right]\right) \quad ,$$

avec $\lim_{np \to \infty} \delta_{n,p}^{(V)} = 0$. Au vu des résultats de concentration prouvés à la fin du Chap. 5, ce calcul en espérance gouverne le comportement de la validation croisée V-fold dans ce cadre.

En comparant ce calcul avec celui du critère idéal (5.3), il apparaît donc que la validation croisée V-fold présente des propriétés d'adaptation à l'hétéroscédasticité du bruit. En revanche, le terme de variance est surestimé car $V/(V-1) > 1$. Considérant $\mathrm{crit}_{\mathrm{VFCV}}$ comme un *critère pénalisé*, le calcul ci-dessus montre que la pénalité correspondante vaut en espérance

$$\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m) - P_n\gamma\left(\widehat{s}_m\right)\right] = \left(1 + \frac{1}{2(V-1)} + \epsilon(n,m)\right) \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$$

avec $\lim_{\min_{\lambda \in \Lambda_m}\{np_\lambda\} \to \infty} \epsilon(n,m) = 0$. Ainsi, *la validation croisée V-fold avec V borné surpénalise*, et ce facteur de surpénalisation est d'autant plus grand que V est petit.

Cette vision des choses est cohérente avec les résultats cités à la Sect. 1.3.1, notamment ceux de Burman [**Bur89**] et Zhang [**Zha93**]. Mais elle permet également de mettre en relation le rôle de V avec l'idée qu'*il est parfois avantageux de surpénaliser, en particulier lorsque la taille de l'échantillon est petite ou le niveau de bruit élevé.* Nous avons expliqué en Sect. 1.2.3 que les raisons de ce fait sont essentiellement la variabilité du critère utilisé pour le choix de modèles. Dans le cas V-fold, celle-ci dépend certes de V ($V = 2$ donnant toujours un critère variable, et $V = n$ pouvant l'être, en particulier en classification), mais aussi de n est σ. C'est pourquoi il n'est pas surprenant que les simulations de la Sect. 5.4 indiquent que *le meilleur choix de V pour la prédiction est parfois $V = 2$*. Ce phénomène n'a — à notre connaissance — jamais[49] été mis en avant dans la littérature V-fold, ce qui s'explique notamment par sa nature non-asymptotique.

La conclusion de notre analyse est donc que le choix de V est encore plus difficile qu'il n'y paraît. La règle «$5 \leq V \leq 10$ est optimal», que bien des praticiens considèrent comme quasi-universelle, est donc à remettre en cause dans des situations «non-asymptotiques».

Une autre option : la pénalisation V-fold. En réponse à cette analyse quelque peu pessimiste de la validation croisée V-fold, nous proposons à la Sect. 5.3 la *pénalisation V-fold* comme méthode concurrente de la validation croisée V-fold.

Cette procédure peut être vue comme une version «validation croisée V-fold» de la pénalisation bootstrap $\mathrm{pen}_{\mathrm{Efron}}$. De façon générale, on définir la pénalité V-fold comme suit. Soit $(B_j)_{1 \leq j \leq V}$ une partition de $\{1, \ldots, n\}$, $P_n^{\left(B_j^c\right)}$ et $\widehat{s}_m^{(-j)}$ définis comme en Sect. 1.3.1. Alors,

$$\mathrm{pen}_{\mathrm{VFCV}}(m) := \frac{C}{V} \sum_{j=1}^{V} \left(P_n - P_n^{\left(B_j^c\right)}\right) \gamma\left(\widehat{s}_m^{\left(B_j^c\right)}\right) \quad ,$$

avec $C \geq V - 1$.

Lorsque $C = V - 1$, on retrouve le critère de validation croisée V-fold corrigé par Burman [**Bur89**] (Remarque 5.2). L'intérêt de laisser C variable est de permettre une surpénalisation, dans les situations où cela semble nécessaire, indépendamment de la valeur de V.

De plus, nous prouvons (Thm. 5.1) une *inégalité oracle trajectorielle non-asymptotique, avec constante presque 1*, et sans terme de reste, dans le cadre de la régression sur des histogrammes.

[49] si ce n'est dans un cadre de test (Dietterich [**Die98**], Alpaydin [**Alp99**]), qui s'apparente donc plus à l'identification qu'à la prédiction ; voir aussi Zhang [**Zha93**].

Malgré la force de la restriction aux histogrammes (uniquement motivée par la possibilité de calculer explicitement la pénalité V-fold et la pénalité idéale), nous pouvons raisonnablement penser que le même type de résultat reste valide dans un cadre beaucoup plus large. Cette conjecture est en particulier supportée par le fait que nos résultats ne nécessitent que très peu d'hypothèses sur le bruit. En particulier, celui-ci peut être *fortement hétéroscédastique* et *non-gaussien*, sans que cela ne perturbe la qualité de la procédure.

D'un point de vue pratique, les pénalités V-fold semblent bénéficier des avantages de la validation croisée V-fold (rapidité de calcul, simplicité et généralité de la définition, adaptation à une grande classe de problèmes, robustesse, *etc.*), tout en étant *plus flexible*. En effet, le paramètre V ne gouverne plus que le temps de calcul et la variabilité de l'estimation, tandis que le paramètre C permet de choisir un critère sans biais ou légèrement surpénalisant. Les simulations de la Sect. 5.4 illustrent l'intérêt de ces pénalités dans diverses situations «difficiles» : bruit hétéroscédastique, fonction de régression présentant des sauts, *etc.*

Pénalisation par rééchantillonnage échangeable. Au Chap. 6 (puis au Chap. 8 pour des résultats et commentaires annexes), nous étudions une famille de pénalités par rééchantillonnage dont un cas particulier est la pénalité bootstrap proposée par Efron [**Efr83**]. Étant donnés des poids de rééchantillonnage $W_1, \dots, W_n$ échangeables, on définit la pénalité

$$\mathrm{pen}(m) = C\mathbb{E}_W \left[\left(P_n - P_n^W \right) \gamma \left(\widehat{s}_m^W \right) \right] \quad ,$$

où $C \geq C_{W,\infty}$, cette dernière ne dépendant que de la variabilité des poids W_i. Par exemple, dans le cas des poids Efron (n), Random hold-out (n), Rademacher $(1/2)$ et Poisson (1), on a $C_{W,\infty} = 1$. Dans le cas des poids Leave-one-out, on a $C_{W,\infty} = n - 1$. Lorsque $C = C_{W,\infty}$, nous montrons que cette pénalité est non biaisée au premier ordre (dans le cas des histogrammes), tandis que le choix de $C > C_{W,\infty}$ permet de surpénaliser si cela est nécessaire.

Cette famille de pénalités comprend des procédures déjà étudiées dans d'autres cadres, et fournit une explication simple aux résultats qui les concernent. Ainsi, avec des poids Efron (n), on retrouve la pénalité $\mathrm{pen}_{\mathrm{Efron}}$ proposée par Efron [**Efr83**] (avec un contraste de type log-vraisemblance). Avec des poids Efron (m), on retrouve la procédure considérée par Shao [**Sha96**], pour laquelle il ne multiplie pas la pénalité par C. Notre calcul de $C_{W,\infty} \approx n/m$ permet donc d'interpréter le choix $m \ll n$ comme une surpénalisation, de même que le critère BIC s'obtient à partir de AIC en multipliant la pénalité par $\ln(n)/2$. Nos calculs permettent également de retrouver un analogue du résultat de Shibata [**Shi97**] à propos de $\mathrm{pen}_{\mathrm{Efron}}$ et $\mathrm{pen}_{\mathrm{AICb}}$.

Au delà d'un simple calcul d'espérance, des inégalités de concentration (dérivées notamment des inégalités de moments de Boucheron, Bousquet, Lugosi et Massart [**BBLM05**]) nous permettent de prouver deux résultats théoriques sur ces pénalités, toujours dans le cadre de la régression sur des histogrammes. Le premier (Thm. 6.1) est une *inégalité oracle trajectorielle non-asymptotique*, analogue à celle concernant les pénalités V-fold. Au prix d'une preuve légèrement plus longue, nous proposons plusieurs jeux d'hypothèses sous lesquels ce résultat reste valide, soulignant ainsi la robustesse de la procédure.

Le second résultat (Thm. 6.2) est *l'adaptation à la régularité hölder* de la fonction de régression s. En choisissant la famille des histogrammes réguliers, nous prouvons que l'estimateur sélectionné par notre procédure atteint (à constante multiplicative près) le *risque minimax non-asymptotique*

$$L_{\alpha,k} R^{\frac{2\alpha}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \|\sigma\|_\infty^{\frac{4\alpha}{2\alpha+k}} \quad ,$$

lorsque $\mathcal{X} \subset \mathbb{R}^k$ et s appartient à la boule de hölder $\mathcal{H}(\alpha, R)$ pour $\alpha \in (0, 1]$ et $R > 0$, avec $\sup_{\mathcal{X}} s - \inf_{\mathcal{X}} s \geq \epsilon > 0$. Un fait remarquable est que cette borne ne fait quasiment pas d'hypothèses sur

le niveau de bruit $\sigma(X)$. Lorsque ce dernier est supposé régulier, nous retrouvons (à constante multiplicative près) le *risque minimax hétéroscédastique* calculé par Galtchouk et Pergamenschikov [**GP05**], avec la bonne dépendance en σ :

$$L_{\alpha,k} R^{\frac{2\alpha}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \|\sigma\|^{\frac{4\alpha}{2\alpha+k}}_{L^2(\mathrm{Leb})} \quad .$$

Comme l'illustre une étude de simulations (Sect. 6.5), ces pénalités par rééchantillonnage échangeable montrent tout leur intérêt face à des problèmes difficiles. Comparées aux pénalités V-fold, on observe un *léger gain en précision à considérer des poids échangeables*, qui se traduit également par de meilleures bornes dans les résultats de concentration de pen dans le cas échangeable par rapport au cas V-fold. Nous ne savons pas toutefois si ces bornes sont du bon ordre de grandeur, si bien qu'il est difficile de tirer des conclusions d'une telle comparaison. Il n'est en revanche pas certain qu'en pratique l'utilisation de poids échangeables soit réellement bénéfique, son temps de calcul étant prohibitif (à l'exception peut-être du Leave-one-out). Il semble donc raisonnable d'utiliser une approximation Monte-Carlo, le nombre B de vecteurs de poids considérés étant alors à rapprocher du V du V-fold.

Nos simulations, ainsi que des calculs supplémentaires, permettent également de comparer les différents poids échangeables entre eux. Si toutes les pénalités sont équivalentes au premier ordre (lorsque $C = C_{W,\infty}$), il apparaît au second ordre[50] (6.15) que

$$\text{penRad } (1/2) \approx \text{penRho } (n/2) > \text{penLoo} >> \text{penEfr} \quad .$$

Ceci confirme la constatation d'Efron d'une légère sous-pénalisation avec $\text{pen}_{\text{Efron}}$. La pénalité leave-one-out s'avère être la moins biaisée, si bien qu'elle a des performances légèrement moins bonnes dans un cadre non-asymptotique (qu'une légère surpénalisation suffit à compenser, via une augmentation de C). Enfin, les pénalités Rademacher et Random hold-out sont celles qui ont montré les meilleures performances dans nos simulations.

Le bilan de cette étude est que *le choix des poids doit principalement être effectué en fonction du temps de calcul* (sachant qu'il faut exclure une approximation Monte-Carlo avec des poids leave-one-out). Ensuite, en fonction des poids choisis, *on ajuste la constante C pour surpénaliser au bon niveau.* En particulier, *si l'objectif est l'identification, il suffit de surpénaliser en prenant* $C \approx C_{W,\infty} \ln(n)/2$. L'utilisation de poids Efron (m) avec $m \ll n$ n'est donc en rien nécessaire pour l'identification.

L'utilisation de rééchantillonnage (V-fold ou échangeable) pour construire des pénalités ouvre également une nouvelle piste pour *déterminer à l'aide des données le bon niveau de surpénalisation.* Dans la discussion de la Sect. 6.6 (voir aussi Sect. 11.3.3), nous proposons de ne plus utiliser l'espérance $\mathbb{E}_W[\cdot]$ mais des quantiles conditionnellement à $\xi_{1\ldots n}$; il reste alors à choisir le bon «niveau de confiance» pour une prédiction optimale.

Premiers résultats en classification. Les résultat que nous venons de décrire sont limités à la régression sur un type particulier de modèles, mais les pénalités V-fold et par rééchantillonnage échangeable peuvent être définis dans un cadre beaucoup plus général. Un cas qui nous intéresse particulièrement est celui de la classification binaire, où la définition de pénalités adaptatives à la condition de marge, calculables en pratique et suffisamment peu variables, reste un problème largement ouvert. Le Chap. 7 indique ce que nous sommes capable de montrer pour le moment, et ce qu'il reste à faire.

[50] dans la comparaison, « > » signifie une petite différence et «>>» une différence plus grande. Il faut cependant se rappeler que ces quatre pénalités sont toutes égales au premier ordre.

Si le chemin restant est long, nous disposons d'ores-et-déjà d'un premier résultat encourageant (Thm. 7.1). Il s'agit d'une inégalité de concentration pour

$$\widehat{p}_2(m) := \mathbb{E}_W \left[P_n^W \left(\gamma\left(\widehat{s}_m\right) - \gamma\left(\widehat{s}_m^W\right) \right) \right] \ ,$$

valide dès lors que le contraste γ est borné et le processus de rééchantillonnage de type «sous-échantillonnage» (Random hold-out, Leave-one-out, V-fold). Ce résultat découle directement d'une inégalité de moments de Boucheron et Massart [**BM04**] pour $P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right)$.

Nous proposons ensuite deux pistes :

- Si l'heuristique (1.15) est valide, alors $2C_{W,\infty}\widehat{p}_2(m)$ pourrait être utilisée comme pénalité.
- Si l'on peut prouver une inégalité de moments pour $P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right)$ analogue à celle de Boucheron et Massart [**BM04**], alors, le même raisonnement fournirait un résultat de concentration sur la pénalité par rééchantillonnage complète.

Dans les deux cas, il reste à savoir comparer la pénalité par rééchantillonnage et la pénalité idéale en espérance, ce qui semble loin d'être simple.

Notons que la constante $C_{W,\infty}$ dans le cas de la classification n'a pas besoin d'être connue de manière précise, puisque l'on peut utiliser l'heuristique de pente (voir Chap. 3) pour la déterminer à partir des données (sous réserve que cette heuristique reste valide en classification).

Calibration des pénalités globales. Enfin, au Chap. 9, nous présentons des remarques sur la calibration des pénalités globales définies par Fromont [**Fro04**], dont les complexités de Rademacher globales sont un cas particulier.

Si ces résultats sont moins aboutis que ceux des autres chapitres, ils indiquent toutefois que la calibration de ces constantes peut poser problème dans certains cadres un peu extrêmes. Il apparaît en effet que la constante $1/\mathbb{E}\left(Z_1\right)_+$ exigée par la théorie est nécessaire en toute généralité (d'un point de vue non-asymptotique, Sect. 9.3), mais ne l'est plus sous une condition de symétrie (trop contraignante en classification), ni d'un point de vue asymptotique (Sect. 9.2). Les calibrations de Lozano [**Loz00**] et Fromont [**Fro07**] sont donc confirmées «en général», mais infirmées du point de vue théorique non-asymptotique.

Cependant, les contre-exemples que nous exhibons nous suggèrent que des bornes plus fines de comparaison en espérance pourraient être obtenues, au prix d'une hypothèse empêchant de considérer des modèles où le meilleur prédicteur s_m aurait un risque $P\gamma\left(s_m\right) < n^{-1}$.

Le phénomène qualitatif mis en lumière au Chap. 9 nous semble également être un argument supplémentaire en faveur de l'utilisation de l'heuristique de pente (voir Chap. 3) pour calibrer une pénalité. En effet, si la «bonne constante» devant la pénalité dépend de la loi inconnue P, il devient nécessaire de l'estimer à partir des données. Cette dépendance étant complexe, l'heuristique de pente présente l'avantage d'être assez simple à mettre en œuvre.

Outils probabilistes. Les résultats que nous venons de décrire reposent sur des inégalités de concentration et des comparaisons d'espérances entre pénalités idéales et pénalités par rééchantillonnage.

Dans le cas des histogrammes, le fondement de nos preuves est un *calcul explicite* de $\mathrm{pen}_{\mathrm{id}}$ et de sa version rééchantillonnée, lorsque les poids sont échangeables (Lemme 5.7, Sect. 5.7.2). Conditionnellement aux indices des X_i qui appartiennent à chacun des éléments I_λ de la partition associée au modèle S_m, ces deux quantités sont des U-statistiques d'ordre 2. Nous montrons qu'elles se concentrent autour de leurs espérances conditionnelles à l'aide d'*inégalités de moments* (Prop. 5.5, Sect. 5.6.3) qui découlent des résultats de Boucheron, Bousquet, Lugosi et Massart [**BBLM05**].

Pour des raisons principalement techniques (liées au fait que les X_i sont aléatoires), nous avons également besoin d'une inégalité de concentration pour $Z = \sum_{\lambda\in\Lambda_m} a_\lambda \left(\frac{1}{X_\lambda} \wedge T \right)$, avec $a_\lambda \geq 0$ et (X_λ) un vecteur multinomial (Lemme 5.4, Sect. 5.6.2). Nous l'obtenons en évaluant les moments exponentiels d'inverses de binomiales, puis en utilisant la notion d'association négative pour utiliser la méthode de Cramér-Chernoff (Dubhashi et Ranjan [**DR98**]).

Dans le cas de la classification (Chap. 7), les inégalités de concentration obtenues découlent d'une inégalité de concentration de Boucheron et Massart [**BM04**], qui repose sur des techniques de localisation (inégalités maximales, concentration de processus empiriques pondérés par peeling).

Les résultats en espérance proviennent de deux méthodes distinctes. D'une part, dans le cas des histogrammes, le calcul explicite nous ramène à évaluer $\mathbb{E}\left[1/Z \mid Z > 0\right]$ pour une variable aléatoire Z qui suit une loi binomiale (Lemme 5.3, Sect. 5.6.1), Poisson (Lemme 6.3, Sect. 6.7) ou Hypergéométrique (Lemme 6.2, Sect. 6.7).

D'autre part, dans le cas de pénalités globales (Chap. 9 ; voir aussi Sect. 10.2.2), nous nous appuyons sur une inégalité de symétrisation classique (dans le cas des poids i.i.d. symétriques) et la preuve d'une inégalité prouvé par Fromont dans le cas de poids multinomiaux [**Fro07**].

1.4. Régions de confiance et tests par rééchantillonnage

Le Chap. 10 de cette thèse étudie du point de vue non-asymptotique des régions de confiance et des procédures de test multiples construites par rééchantillonnage. Nous présentons ici rapidement ces problématiques, plusieurs motivations pratiques et quelques stratégies pour les aborder (en particulier celles qui se fondent sur le rééchantillonnage). Enfin, nous décrivons les contributions du Chap. 10.

1.4.1. Position du problème. Dans le cas de la régression, nous avons décrit dans les deux sections précédentes diverses méthodes aboutissant à la construction d'un estimateur $\widehat{s}$ de la moyenne s d'un signal bruité $Y = s(X)+\epsilon$. À part dans le cas déterministe, nous savons cependant que $\widehat{s} \neq s$, même si $\widehat{s}$ est «optimal». En effet, l'optimalité ici signifie juste que notre estimateur est *l'un des moins faux* parmi un ensemble d'estimateurs donné. Il nous est ensuite possible d'avoir une idée de l'erreur commise, c'est-à-dire d'estimer[51] $\mathbb{E}\left(s(X) - \widehat{s}(X)\right)^2$. Cependant, une telle mesure ne nous donne pas d'informations sur la répartition de $\widehat{s}(X)$ autour de $s(X)$.

On peut par exemple accepter d'exclure un événement de petite probabilité α, sur lequel $\widehat{s}$ est susceptible de se tromper beaucoup, mais vouloir savoir précisément dans quelle zone se situe s hors de cet événement. Une réponse à cette question est appelée *région de confiance pour s de niveau*[52] α.

Dans de nombreuses applications, on n'a pas besoin de tant d'informations, et l'on veut simplement savoir si le signal s est nul ou non. Par exemple, si s est la différence entre l'activité cérébrale du cortex visuel au repos et après une stimulation, un signal non-nul signifie que le cortex visuel est sollicité en réponse à cette stimulation. Il s'agit alors d'un problème de *test*.

Dans l'exemple de l'imagerie cérébrale (de même que dans la plupart des problèmes pratiques actuels, notamment en biologie), on dispose en général de données en de nombreux points du cortex. Il est alors tentant de réaliser un test en chaque point, et d'agglomérer ces résultats en une «carte» des points du cortex qui répondent à la stimulation. L'erreur commise ici est la suivante. Si

[51] Si $\widehat{s}$ minimise le risque empirique sur un modèle S_m de dimension D_m, $P_n\gamma\left(\widehat{s}_m\right) + 2\widehat{\sigma^2}D_m n^{-1}$ est un estimateur non-biaisé du risque. En général (notamment si $\widehat{s} = \widehat{s}_{\widehat{m}}$ a été obtenu par sélection de modèles), on peut estimer ce risque si l'on a pas utilisé toutes les données pour construire $\widehat{s}$. Les observations restantes sont alors appelées *échantillon test*.

[52] de manière équivalente, on parle de région de confiance de *probabilité de couverture* $1-\alpha$.

l'on réalise $K = 10\,000$ tests, chacun ne se trompant[53] qu'avec probabilité $\alpha = 0.05$ (par exemple), alors en moyenne au moins 500 réponses des tests seront positives[54], même si la stimulation n'a en réalité activé aucune zone du cortex. Pour ne pas commettre ce genre d'erreur, on réalise un *test multiple* plutôt que de juxtaposer les réponses de nombreux tests simples.

Régions de confiance. Soit $\theta \in \Theta$ une quantité d'intérêt reliée à la loi P qui a généré des observations $\xi_{1\dots n}$. Par exemple, la moyenne $\mu \in \mathbb{R}^K$ (avec $K = 1$, K grand ou K infini), la variance σ^2 (qui est un réel dans les cas unidimensionnel et homoscédastique, mais peut être de grande dimension en général), *etc.*

DÉFINITION 1.1. Soit $\alpha \in (0,1)$. Une région de confiance de probabilité de couverture $1-\alpha$ (ou encore de niveau α) pour θ est une partie $\mathcal{R}(\xi_{1\dots n}) \subset \Theta$, telle que

$$\mathbb{P}(\theta \in \mathcal{R}(\xi_{1\dots n})) \geq 1-\alpha \ . \tag{1.18}$$

Lorsque Θ est muni d'une distance d, on cherche souvent $\mathcal{R}$ sous la forme d'une boule centrée autour d'un estimateur $\widehat{\theta}$ de θ :

$$\mathcal{R}(\xi_{1\dots n}) := \left\{ x \in \Theta \text{ t.q. } d\left(x, \widehat{\theta}(\xi_{1\dots n})\right) \leq \widehat{r}_\alpha(\xi_{1\dots n}) \right\} \ . \tag{1.19}$$

Seul le rayon $\widehat{r}_\alpha(\xi_{1\dots n})$ est alors à déterminer. On parle de *boule de confiance*, ou encore d'*intervalle de confiance symétrique* si $\Theta \subset \mathbb{R}$.

Dans le cas réel, on considère également souvent des *intervalles de confiance unilatères*[55]

$$\mathcal{R}(\xi_{1\dots n}) := \left[\widehat{\theta} - \widehat{g}_\alpha(\xi_{1\dots n}), \infty\right) \ ,$$

ou des *intervalles de confiance bilatères asymétriques*

$$\mathcal{R}(\xi_{1\dots n}) := \left[\widehat{\theta} - \widehat{g}_\alpha(\xi_{1\dots n}), \widehat{\theta} + \widehat{d}_\alpha(\xi_{1\dots n})\right] \ ,$$

avec $\widehat{g}_\alpha(\xi_{1\dots n})$ et $\widehat{d}_\alpha(\xi_{1\dots n})$ à définir.

Lorsque $\Theta \subset \mathbb{R}^K$ avec $K > 1$, un choix classique est la distance induite par la norme L^p, $p \geq 1$, si les incertitudes sur les coordonnées $\theta_1, \dots, \theta_K$ sont du même ordre. Si ce n'est pas le cas, on définit alors des *ellipsoïdes de confiance* tenant compte des différences entre coordonnées. Quand $p = \infty$, on parle également d'*intervalles de confiance simultanés*, puisque $\mathcal{R}(\xi_{1\dots n})$ est le produit de K intervalles de confiance pour les coordonnées $\theta_1, \dots, \theta_K$ de θ, dont la validité est assurée *simultanément*.

Si la distance d est connue[56], construire une boule de confiance de la forme (1.19) revient à estimer le *quantile d'ordre* $1-\alpha$ de $d(\theta, \widehat{\theta}(\xi_{1\dots n}))$:

$$q_\alpha := \inf\left\{ t \geq 0 \text{ t.q. } \mathbb{P}\left(d\left(\theta, \widehat{\theta}(\xi_{1\dots n})\right) > t\right) \leq \alpha \right\} \ .$$

On distingue essentiellement trois types d'approches :

– l'approche *paramétrique* : on suppose la loi de $\widehat{\theta}(\xi_{1\dots n})$ connue (par exemple gaussienne ou poissonienne), avec un petit nombre de paramètres à estimer. On estime alors q_α par plug-in (et éventuellement par simulations si l'on n'a pas de formule pour q_α ; il s'agit alors de *rééchantillonnage paramétrique*).

[53] au sens de l'erreur de première espèce, voir la définition plus bas.

[54] *i.e.* 500 hypothèses nulles seront rejetées, voir plus bas.

[55] one-sided, en anglais, par opposition aux intervalles de confiance bilatères (two-sided).

[56] ce n'est pas forcément le cas, d pouvant par exemple dépendre des écarts-types $(\sigma_k)_{1\leq k\leq K}$ sur les coordonnées, qu'il faut alors estimer au préalable.

- l'approche *asymptotique* : on détermine la loi asymptotique de $d\left(\theta, \widehat{\theta}\left(\xi_{1\dots n}\right)\right)$ (convenablement renormalisée), et on utilise le quantile d'ordre $1-\alpha$ de cette distribution limite pour estimer q_α.
- l'approche par *rééchantillonnage* : on calcule $d\left(\widehat{\theta}(\xi_{1\dots n}), \widehat{\theta}(\xi^\star_{1\dots n})\right)$ pour différents rééchantillons $\xi^\star_{1\dots n}$, et on estime q_α à l'aide du quantile empirique d'ordre α. La justification de telles procédures est en général asymptotique (Hall [**Hal92**]).

Mentionnons également qu'une approche non-asymptotique (fondée sur la méthode de Lepski, qui n'est pas sans lien avec la sélection de modèles) a permis à Baraud [**Bar04**] de construire des boules de confiance dans un cadre de régression gaussienne.

Au Chap. 10, nous proposons une justification non-asymptotique de régions de confiance construites par rééchantillonnage. Ceci permet notamment de considérer un paramètre θ de grande dimension K, éventuellement plus grande que n.

Tests d'hypothèses. L'un des principaux objets des sciences expérimentales est d'établir des relations de cause à effet ou, à défaut, des corrélations entre différents phénomènes. L'objet de la théorie statistique des tests est de fonder mathématiquement une réponse à une question fermée telle que : «Y a-t-il ou non un lien significatif ?» Plus généralement, et plus formellement, un test cherche à répondre à une question de la forme : «la distribution P qui a généré les données $\xi_{1\dots n}$ possède-t-elle une propriété $\mathcal{P}$?»

Dans ce but, on formule deux hypothèses[57] :

- une *hypothèse nulle* H_0. En général, elle correspond au cas où il ne se passe rien, du moins rien de «nouveau». En un sens, elle modélise l'ensemble des connaissance scientifiques du présent.
- une *hypothèse alternative* H_1. Elle correspond aux «découvertes» que l'on cherche à établir.

Formellement, on peut voir H_0 et H_1 comme des parties de l'ensemble $\mathcal{M}$ des mesures de probabilité sur Ξ^n. Un cas typique est $H_1 = \mathcal{M}\backslash H_0$, mais d'autres choix sont possibles.

Un *test statistique* est alors une application mesurable $T : \Xi^n \mapsto \{0,1\}$. Lorsque $T(\xi_{1\dots n}) = 1$, on dit que l'*hypothèse nulle est rejetée* (on a fait une découverte). Lorsque $T(\xi_{1\dots n}) = 0$, on dit que l'*on accepte l'hypothèse nulle* (pas de découverte). Un test peut donc se tromper de deux manières :

- on parle d'*erreur de première espèce*[58] lorsque $T(\xi_{1\dots n}) = 1$ alors que $P \in H_0$, *i.e.* lorsque H_0 est rejetée à tort.
- on parle d'*erreur de seconde espèce*[59] lorsque $T(\xi_{1\dots n}) = 0$ alors que $P \in H_1$, *i.e.* lorsque H_0 est acceptée alors que H_1 était correcte.

L'approche de Neyman-Paerson consiste à contrôler *d'abord* la probabilité d'une erreur de première espèce :

$$\forall P \in H_0, \qquad \mathbb{P}_{\xi_{1\dots n}\sim P}\left(T\left(\xi_{1\dots n}\right) = 1\right) \leq \alpha \ , \tag{1.20}$$

[57] Notons la disymétrie entre H_0 (la connaissance établie) et H_1 (de nouvelles théories, candidates à remplacer les théories présentes). L'objectif d'un test est de déterminer si les connaissances présentes permettent d'expliquer les données observées, ou bien si ces données réfutent les théories présentes. On peut ainsi rapprocher cette formulation de la conception poppérienne de la science : «Une théorie qui n'est réfutable par aucun événement qui se puisse concevoir est dépourvue de caractère scientifique.» (Karl Popper, *Conjectures et réfutations*, Chap. 1, Sect. 1). Mais ceci est une autre histoire...

[58] type I error.

[59] type II error.

pour un certain $\alpha \in (0,1)$ (appelé le *niveau* du test). Ce n'est qu'une fois ce contrôle établi (pour un α fixé) que l'on cherche à minimiser la probabilité d'une erreur de seconde espèce :

$$\forall P \in H_1, \qquad \mathbb{P}_{\xi_{1\dots n}\sim P}\left(T\left(\xi_{1\dots n}\right)=0\right) \leq 1-\beta \ . \tag{1.21}$$

Lorsque (1.21) est satisfaite, on dit que T est de *puissance* $\beta \in (0,1)$. Cette disymétrie souligne que l'objectif d'un test est avant tout de garantir la validité d'un rejet. Dans un second temps seulement, on souhaite maximiser la puissance du test (sinon, $T \equiv 0$ pourrait convenir parfaitement). *Le fait d'accepter l'hypothèse nulle ne prouve donc en rien que celle-ci est valide*, mais simplement que soit elle est valide, soit les données ne permettent pas de l'invalider (en général parce que T n'est pas assez puissant, par exemple parce que n est trop petit, ou le niveau de bruit trop grand).

Mentionnons enfin la notion de *p-valeur*, souvent utilisée pour formuler le résultat d'un test sans avoir à choisir un niveau arbitraire (pourquoi $\alpha = 0.05$ plutôt que $\alpha = 0.01$?). Lorsque l'on dispose d'une famille $(T_\alpha)_{\alpha\in[0,1]}$ de tests de niveaux α, telle que $\alpha \mapsto T_\alpha$ est croissante et continue à droite (pour presque toute observation $\xi_{1\dots n}$), alors

$$T_\alpha\left(\xi_{1\dots n}\right) = \mathbb{1}_{p(\xi_{1\dots n})\leq\alpha} \qquad \text{avec} \qquad p\left(\xi_{1\dots n}\right) := \inf\left\{\alpha\in[0,1] \text{ t.q. } T_\alpha\left(\xi_{1\dots n}\right)=1\right\}$$

(voir [**Roq07**], Lemme 9.3). La variable aléatoire p est appelée *p-valeur*. On rejette l'hypothèse nulle lorsque $p \leq \alpha$ (le niveau peut donc être fixé *a posteriori*).

La formulation ci-dessus conduit à la stratégie générale suivante pour construire un test de niveau α :

(1) Définir une «statistique de test» $S : \Xi^n \mapsto \mathbb{R}$, telle que S a tendance à prendre de petites valeurs sous H_0 (c'est-à-dire des valeurs plus petites que celles qu'elle prend sous H_1).

(2) Évaluer le quantile $q_\alpha(P)$ d'ordre $1-\alpha$ de S sous $P \in H_0$, puis $q_\alpha := \sup_{P\in H_0} q_\alpha(P)$.

(3) Poser $T\left(\xi_{1\dots n}\right) := \mathbb{1}_{S(\xi_{1\dots n})\leq q_\alpha}$.

Par exemple, lorsque $\Xi = \mathbb{R}$, $H_0 = \left\{\mathcal{N}(\mu_0,\sigma^2)^{\otimes n}\right\}$ («$\xi_{1\dots n}$ sont i.i.d. gaussiens de moyenne μ_0 et de variance σ^2») et $H_1 = \left\{\mathcal{N}(\mu,\sigma^2)^{\otimes n} \text{ t.q. } \mu\neq\mu_0\right\}$ («la moyenne n'est pas μ_0»), une statistique de test possible est $S\left(\xi_{1\dots n}\right) = \sigma^{-1}n^{-1/2}\left|\sum_i\left(\xi_i-\mu_0\right)\right|$. Sous H_0, S est la valeur absolue d'une variable normale centrée réduite. Notons $\overline{\Phi}$ la queue de distribution supérieure d'une gaussienne standard. Le test qui rejette H_0 si et seulement si $\overline{\Phi}\left(\sigma^{-1}n^{-1/2}\sum_i(\xi_i-\mu_0)\right) \leq \alpha/2$ est donc de niveau α. Plus généralement, lorsque $H_0 = \left\{P \text{ t.q. } S(P)=0\right\}$ pour une fonctionnelle S régulière (ce qui correspond à de nombreux exemple, voir Bickel et Ren [**BR01**]), une statistique de test naturelle est $S(P_n)$ (notée $S\left(\xi_{1\dots n}\right)$ dans l'exemple précédent).

Lorsque H_0 est de la forme «$\theta = \theta_0$», on peut facilement construire un test T de niveau α à partir d'une région de confiance $\mathcal{R}$ pour θ de probabilité de couverture $1-\alpha$:

$$T\left(\xi_{1\dots n}\right) = \mathbb{1}_{\theta_0\in\mathcal{R}(\xi_{1\dots n})} \ .$$

Lorsque H_0 est de la forme «$\theta \in \Theta_0$», on peut appliquer le même principe en posant

$$T\left(\xi_{1\dots n}\right) = \mathbb{1}_{\Theta_0\cap\mathcal{R}(\xi_{1\dots n})\neq\emptyset} \ ,$$

mais T risque d'être très peu puissant si la forme de $\mathcal{R}$ n'est pas adaptée à celle de Θ_0. Notons que pour obtenir un test de niveau α, le contrôle de la probabilité de couverture de $\mathcal{R}$ peut n'être valide que lorsque $P \in H_0$. Ceci montre en quoi il peut être plus facile de construire un test qu'une région de confiance.

Le cadre restreint de cette introduction ne nous permettant pas de poursuivre notre chemin dans la théorie des tests (simples) d'hypothèses, nous renvoyons le lecteur intéressé au livre de

Lehmann et Romano [**LR05**]. Nous mentionnons cependant ici qu'il est possible d'utiliser la sélection de modèles pour construire un test. L'idée de la *sélection de tests* est de considérer toute une famille d'alternatives $(H_{1,m})_{m\in\mathcal{M}_n}$ à l'hypothèse nulle H_0, et de sélectionner un test $T_{\widehat{m}}$ parmi les plus performants de la famille $(T_m)_{m\in\mathcal{M}_n}$. Voir notamment à ce sujet les articles de Baraud, Huet et Laurent [**BHL03, BHL05**].

Tests d'hypothèses multiples. Le problème qui nous intéresse plus particulièrement ici est celui des *tests multiples*. Au lieu d'une seule hypothèse nulle, nous disposons désormais d'un ensemble $\mathcal{H}$ d'hypothèses nulles que l'on souhaite tester[60]. Nous supposerons toujours ici que $\mathcal{H}$ est fini, de cardinal K :

$$\mathcal{H} := \{ H_{0,1}, \dots, H_{0,K} \} \quad .$$

Par exemple, si $\Xi = \mathbb{R}^K$, on peut souhaiter tester pour chaque coordonnée $k \in \{1, \dots, K\}$ si $\mu_k = \mu_{0,k}$ ou non. En supposant les données gaussiennes de matrice de covariance Σ connue, cela correspond aux hypothèses nulles

$$H_{0,k} := \left\{ \mathcal{N}(\mu, \Sigma) \text{ t.q. } \mu \in \mathbb{R}^K \text{ et } \mu_k = \mu_{0,k} \right\} \qquad \text{pour} \quad 1 \leq k \leq K \quad .$$

Notons $\mathcal{H}_0$ l'ensemble des hypothèses nulles vraies. L'objectif d'un test multiple est alors de déterminer l'ensemble $\mathcal{H}_0^c$ des hypothèses nulles qui sont fausses. Une *procédure de test multiple* peut donc se formaliser comme une application $R : \Xi^n \mapsto \mathfrak{P}(\mathcal{H})$, l'ensemble $R(\xi_{1\dots n})$ désignant l'ensemble des hypothèses nulles rejetées. La qualité d'une telle procédure dépend donc, d'une part, de la loi du nombre d'erreurs de première espèce $(\mathrm{Card}\,(R(\xi_{1\dots n}) \cap \mathcal{H}_0)$, et d'autre part, du nombre d'erreurs de seconde espèce $(\mathrm{Card}\,(R(\xi_{1\dots n})^c \cap \mathcal{H}_0^c)$.

De même qu'en test simple, on impose d'abord à un test multiple un contrôle sur le nombre d'erreurs de première espèce, qui peut se mesurer de différentes manières :

- Le *Family-Wise Error Rate* (FWER) est la probabilité de *rejeter au moins une hypothèse nulle à tort* :
$$\mathrm{FWER}(R) := \mathbb{P}\left(\mathcal{H}_0 \cap R(\xi_{1\dots n}) \neq \emptyset\right) \quad . \tag{1.22}$$
- Le *False Discovery Rate* (FDR, introduit par Benjamini et Hochberg [**BH95**]) est la proportion moyenne de «fausses découvertes» (hypothèses nulles rejetées à tort) parmi l'ensemble des hypothèses rejetées :
$$\mathrm{FDR}(R) := \mathbb{E}\left[\frac{\mathrm{Card}\,(\mathcal{H}_0 \cap R(\xi_{1\dots n}))}{\mathrm{Card}\,(R(\xi_{1\dots n}))} \mathbb{1}_{\mathrm{Card}(R(\xi_{1\dots n}))>0} \right] \quad . \tag{1.23}$$

Notons que l'on a toujours $\mathrm{FDR}(R) \leq \mathrm{FWER}(R)$, si bien qu'un contrôle du FWER aboutit à une procédure plus conservative. En revanche, un contrôle du FDR ne donne pas une réponse définitive dans le cadre d'une étude scientifique rigoureuse, puisque parmi les «découvertes» mises en évidence, il reste en moyenne une fraction non-nulle de fausses découvertes. Une procédure contrôlant le FDR est donc avant tout utile pour mettre en évidence des hypothèses à tester plus particulièrement (par exemple, dans le cas de puces à ADN, un petit nombre de gènes à étudier d'une autre manière). Pour la validation finale d'un certain nombre de découvertes, un contrôle du FWER est nécessaire. D'autres mesures du nombre d'erreurs de première espèce (moins utilisées en pratique) existent, nous renvoyons à [**GDS03, Roq07**] pour leurs définitions.

On dit que la procédure R a un FWER (resp. un FDR) contrôlé au niveau α lorsque pour toute distribution $P \in \bigcap_{H_{0,k}\in\mathcal{H}_0} H_{0,k}$, $\mathrm{FWER}(R) \leq \alpha$ (resp. $\mathrm{FDR}(R) \leq \alpha$). Un tel contrôle est parfois appelé *contrôle fort*, par opposition au contrôle faible qui n'a lieu que lorsque $\mathcal{H} = \mathcal{H}_0$. Ce

[60] Par souci de simplicité, nous supposerons ici implicitement que les hypothèses alternatives sont toutes de la forme $H_{1,k} = H_{0,k}^c$.

contrôle étant imposé, on cherche une procédure R qui minimise le nombre d'erreurs de seconde espèce, *e.g.* dont la *puissance*

$$\mathbb{E}\left[\mathrm{Card}\left(\mathcal{H}_0^c \cap R\left(\xi_{1\dots n}\right)\right)\right] \tag{1.24}$$

est maximale.

L'approche naïve pour définir une procédure de test multiple consiste à définir K tests simples $T_{1,\alpha}, \dots, T_{K,\alpha}$, chacun de niveau α, et à poser

$$R\left(\xi_{1\dots n}\right) := \left\{ H_{0,k} \text{ t.q. } T_{k,\alpha}\left(\xi_{1\dots n}\right) = 1 \right\} \ .$$

Cependant, une telle procédure n'a en général[61] qu'un FWER borné par $K\alpha$, en utilisant la borne d'union

$$\mathrm{FWER}(R) = \mathbb{P}_{\xi_{1\dots n}\sim P}\left(\exists k,\, P \in H_{0,k} \text{ et } T\left(\xi_{1\dots n}\right) = 1\right) \leq \sum_{k=1}^{K} \mathbb{P}_{\xi_{1\dots n}\sim P\in H_{0,k}}\left(T\left(\xi_{1\dots n}\right) = 1\right) \leq K\alpha \ .$$

Pour obtenir un contrôle au niveau α, on peut alors utiliser la *correction de Bonferroni*, qui consiste à utiliser K procédures de test simple au niveau α/K :

$$R\left(\xi_{1\dots n}\right) := \left\{ H_{0,k} \text{ t.q. } T_{k,(\alpha/K)}\left(\xi_{1\dots n}\right) = 1 \right\} \ .$$

L'avantage de cette correction est qu'elle est valable quelle que soit la loi jointe des statistiques $T_{k,\alpha}$.

On peut également voir la correction de Bonferroni comme une manière d'«ajuster» des p-valeurs : on calcule tout d'abord les p-valeurs $p_1, \dots, p_K$ associés aux K tests simples de $H_{0,k}$ contre $H_{1,k}$. Ensuite, on définit les p-valeurs ajustées $\widetilde{p}_k = \min(Kp_k, 1)$. Enfin, l'ensemble des hypothèses rejetées est

$$R := \left\{ H_{0,k} \text{ t.q. } \widetilde{p}_k \leq \alpha \right\} \ .$$

Il existe d'autres manières d'ajuster des p-valeurs (par exemple *step-down* et *step-up*), dont les validités sont prouvées sous différentes hypothèses sur la dépendance entre les p-valeurs $p_1, \dots, p_K$. Pour un aperçu plus détaillé du monde des tests multiples, nous renvoyons à la revue de Ge, Dudoit et Speed [**GDS03**], ainsi qu'à l'introduction de la Partie II de la thèse de Roquain [**Roq07**]. Au sujet des méthodes step-down, on consultera également Romano et Wolf [**RW05**].

Notons enfin que l'on peut souvent construire une procédure de test multiple à partir de régions de confiance. Par exemple, si $H_{0,k}$ est de la forme «$\theta_k = \theta_{k,0}$» pour un paramètre θ_k, on peut tester simultanément $(H_{0,k})_{1\leq k\leq K}$ à l'aide d'intervalles de confiance simultanés $(I_k)_{1\leq k\leq K}$ sur les θ_k. Si leur probabilité de couverture simultanée est supérieure à $1-\alpha$, et si $R\left(\xi_{1\dots n}\right)$ est l'ensemble des $H_{0,k}$ tel que $\theta_{k,0} \notin I_k$, alors le FWER de R est majoré par α. Nous détaillons et utilisons cette méthode à la Sect. 10.4.1.

1.4.2. Motivations pratiques. Les applications nécessitant l'utilisation de régions de confiance ou de tests multiples sont nombreuses. Souvent, la dimension des paramètres d'intérêt ou le nombre d'hypothèses à tester est très grand, alors que le nombre de répétitions indépendantes est beaucoup plus modeste. Ceci motive l'approche *non-asymptotique* que nous avons empruntée au Chap. 10. Nous avons plus particulièrement en tête deux domaines de la biologie où de telles procédures sont particulièrement utiles. Nous les présentons dans cette section.

[61] il existe des cas où cette borne d'union est atteinte, voir **[Roq07]** Sect. 9.4.4.

Imagerie cérébrale. Nous décrivons ici un problème typique de magnétoencéphalographie (MEG). L'objectif est de comprendre quelles zones du cerveau jouent un rôle dans différentes activités, par exemple la lecture, le mouvement de la main [**JLN+07**], la vision consciente ou inconsciente [**SBD05, DBD07**]. Pour cela, on place un sujet dans un dispositif d'imagerie (par exemple, de MEG), qui mesure le champ magnétique en différents points de la surface de son crâne au cours d'une expérience faisant intervenir l'activité étudiée. Il s'agit généralement d'une tâche précise à accomplir : lire un texte, déplacer un curseur avec la main, regarder une série d'images puis répondre ensuite à une question à leur propos. Cette tâche est souvent répétée, et chacun de ces «essais» est séparé du précédent par une période de repos (qui fournit des mesures témoins, en l'absence de stimulation). Dans le cas de la MEG, la précision temporelle est très grande : il est ainsi possible d'obtenir une image par milliseconde, au cours d'expériences durant quelques secondes.

Les données subissent ensuite un pré-traitement (*cf.* la thèse de Baillet [**Bai98**]). À chaque instant, il faut d'abord résoudre un problème inverse pour reconstruire l'activité électrique à la surface du cerveau. C'est un problème difficile (Darvas *et al.* [**DRP+05**]) : à partir d'une centaine de senseurs, on cherche à déterminer l'activité électrique en environ 15 000 points. On veut alors comparer l'évolution temporelle de cette activité en chaque point entre les périodes de repos et les périodes de stimulation (généralement limitées à une fenêtre de 100 ms peu après le stimulus).

Une manière de réaliser ceci est de calculer la différence (en chaque point) entre les deux évolutions de l'activité mesurée, puis de chercher les points où celle-ci est significativement non-nulle. Étant donné le grand nombre de points et d'instants de mesures, on réduit souvent chaque courbe d'activité à une seule donnée intégrée. Dans les deux cas, on dispose désormais d'un vecteur $\xi \in \mathbb{R}^K$ avec $K = E \times T$, où $E \approx 15\,000$ est le nombre de points et $1 \leq T \leq 1\,000$ est le nombre de mesures par point. Ainsi, typiquement, $10^4 \leq K \leq 10^7$.

Lorsque l'expérience n'est réalisée qu'avec un seul sujet, les répétitions proviennent du nombre n_{ess} d'essais successifs. Ce nombre n_{ess} est alors limité pour éviter tout phénomène d'adaptation du sujet au dispositif. En général, on a $20 \leq n_{\text{ess}} \leq 100$.

L'expérience peut également être réalisée avec $n_{\text{suj}} > 1$ sujets (choisis de façon aussi homogène que possible, à moins que l'objectif ne soit de déterminer des différences entre deux groupes donnés). Dans ce cas, on réalise n_{ess} essais par sujet, et l'on fait la moyenne de ces répétitions pour n'obtenir qu'une observation ξ par sujet. De cette façon, on diminue le niveau de bruit, et l'on diminue fortement les corrélations temporelles. En effet, l'évolution de l'activité électrique en un point est souvent oscillante, si bien que les perturbations de la phase de ces oscillations induisent de fortes corrélations temporelles lorsque $n_{\text{ess}} = 1$. Avec plusieurs essais, celles-ci deviennent négligeables (si n_{ess} est assez grand). Dans ce second cas, on dispose donc de n_{suj} répétitions. En général, $15 \leq n_{\text{suj}} \leq 100$. Il peut se produire que l'on dispose de plus de sujets ($n_{\text{suj}} = 4\,000$, Waberksi *et al.* [**WGK+03**]), mais cela est malheureusement exceptionnel.

Les données prétraitées ont donc la forme suivante :

$$\xi^1, \dots, \xi^n \in \mathbb{R}^K \qquad \text{i.i.d.}$$

avec $n \leq 100 \ll 10^4 \leq K$. Nous nous trouvons donc dans un cadre hautement non-asymptotique. La loi de ξ est bien évidemment inconnue. Le problème posé est de déterminer

$$\{1 \leq k \leq K \text{ t.q. } \mu_k := \mathbb{E}[\xi_k] \neq 0\} \quad .$$

Il s'agit d'un problème de test multiple.

L'une des difficultés majeures de ce problème est que les coordonnées de chaque observation ξ sont fortement corrélées, et que ces corrélations sont totalement inconnues. Il n'est donc pas envisageable de définir un modèle paramétrique simple pour la loi de ξ. En effet, ces corrélations proviennent de facteurs multiples. D'une part, le prétraitement utilisant 150 valeurs pour en reconstruire 15 000, il est clair que le bruit de mesure en chacun des 150 senseurs induit un bruit fortement corrélé spatialement dans les observations ξ^i. De plus, le bruit provient également de facteurs environnementaux au cours des essais successifs, ou de différences entre sujets. Celui-ci est donc présent dans l'activité cérébrale réelle, et est fortement influencé par les connexions neuronales à l'intérieur du cerveau. Il y a donc de nombreuses corrélations «à distance» entre les différents points du cerveau. Enfin, lorsque l'expérience est limitée à un sujet, il est impossible de négliger les corrélations temporelles, qui sont alors certainement plus fortes que les corrélations spatiales. On ne peut même pas exclure des corrélations spatio-temporelles, l'activité en un point n'influençant directement l'activité en un autre point qu'avec un certain retard.

Dans de telles conditions, les méthodes couramment utilisées relèvent de la théorie des champs aléatoires ou de tests par permutations ou par rééchantillonnage. Les deux premières méthodes sont comparées par Pantazis *et al.* [**PNBL05**]. L'inconvénient de la troisième reste le caractère généralement asymptotique de sa justification. L'un des principaux intérêts des résultats du Chap. 10 est donc de fournir des résultats non-asymptotiques sur des procédures de test multiple par rééchantillonnage.

L'approche la plus commune en imagerie cérébrale est de chercher à contrôler la probabilité de faire au moins une fausse découverte (FWER, défini par (1.22)). Ceci est certainement dû au fait que le FDR a été introduit plus récemment, et est moins connu des neurobiologistes (bien que Perone Pacifico, Genovese et Verdinelli [**PPGVW04**] ont proposé un contrôle du FDR dans le cas de champs aléatoires). Cependant, nous pouvons donner un autre argument en faveur du FWER dans certains cas. En effet, la plupart des expériences font intervenir le système visuel ou le système moteur (voire les deux). Lorsque c'est le cas, de grandes zones du cortex sont activées fortement, si bien que l'on s'attend à ce qu'un grand nombre de moyennes μ_k soient non-nulles. Cependant, les régions du cortex visuel et du cortex moteur primaire sont bien connues des neurobiologistes, et les détecter n'apporte que peu d'informations[62]. En général, la question qui se pose porte surtout sur le reste du cortex : y a-t-il d'autres zones concernées par la lecture, le mouvement de la main, la vision consciente ? Si ces zones existent mais sont petites, une approche de type FDR n'apportera pas toujours d'informations fiables supplémentaires : on ne pourra pas savoir si la petite zone détectée fait partie des $\alpha = 5\%$ (par exemple) de «fausses découvertes» autorisées (en moyenne), ou si ce sont de vraies découvertes. En revanche, avec un contrôle du FWER, on obtient une procédure plus conservative, mais chaque rejet d'hypothèse nulle apporte de l'information, indépendamment de l'activité dans le reste du cerveau.

Analyse de données de puces à ADN. Une puce à ADN mesure le niveau d'expression d'un grand nombre de gènes (plusieurs milliers, et jusqu'à plus de 300 000, ce qui représente souvent la totalité du génome de l'organisme étudié) dans des conditions expérimentales données. Elle permet ainsi d'identifier[63] les gènes spécifiques d'un type de cellule, ou bien ceux qui ont un rôle dans une maladie. Dans ce dernier exemple, on compare les niveaux d'expression des gènes dans une cellule

[62] Si ce n'est une confirmation d'un fait attendu, et une information temporelle, raisons pour lesquelles on ne peut pas tout simplement «oublier» ces régions.

[63] En général, une fois ces gènes identifiés, les biologistes poursuivent leurs études en se focalisant spécifiquement sur ces gènes. C'est pourquoi un contrôle du FDR a un sens dans ce cadre. Il s'agit surtout de mettre en évidence quelques gènes à étudier pour comprendre un mécanisme ou une maladie.

malade et dans une cellule saine («le témoin»). Ces mesures étant naturellement variables[64], ces comparaisons relèvent d'une procédure de test multiple. Le nombre K d'hypothèses à tester est alors le nombre de gènes (donc de 10^3 à 10^5), alors qu'on dispose en général de peu de répétitions d'une même expérience (quelques dizaines en général, en raison du coût élevé des puces à ADN).

Une manière de formuler le problème est de considérer la différence ξ_k^i de niveau d'expression pour le gène k, entre les conditions expérimentales et le témoin, pour la i-ème puce à ADN. L'hypothèse nulle $H_{0,k}$ «le gène k s'exprime de la même manière dans les conditions expérimentales et dans le témoin» peut donc être reformulée : $\mathbb{E}\left[\xi_k^1\right] = 0$ (les $\xi^i \in \mathbb{R}^K$ étant i.i.d., $1 \leq i \leq n$). Comme dans le cas des neuroimages, les K coordonnées de ξ^1 sont corrélées, et ces corrélations ont une forme générale (il n'y a d'ailleurs même plus de notion de distance naturelle, puisque chaque coordonnée correspond ici à un gène), inconnue, et impossible à estimer.

Pour un aperçu des méthodes de tests multiples utilisées dans l'analyse de puces à ADN, nous renvoyons aux revues de Dudoit, Shaffer et Boldrick [**DSB03**] et de Ge, Dudoit et Speed [**GDS03**], ainsi qu'aux références qui y sont indiquées.

1.4.3. Méthodes par rééchantillonnage. Les principales utilisations du bootstrap sont la construction d'intervalles de confiance et le calcul de p-valeurs pour des statistiques de test (Boos [**Boo03**]). Il existe donc un tel nombre d'articles et de livres entiers à ce sujet que nous ne prétendons pas à l'exhaustivité dans cette introduction. Nous nous bornerons à esquisser quelques approches classiques, et à donner quelques références, plus particulièrement celles qui sont liées aux travaux du Chap. 10.

Régions de confiance. Pour plus de simplicité, nous nous focalisons ici sur le cas unidimensionnel, *i.e.* aux intervalles de confiance. La plupart des résultats classiques se limitent à ce cadre, et l'on peut noter qu'une fois une distance choisie sur Θ, construire une région de confiance se ramène au cas unidimensionnel puisque seul le rayon de la boule de confiance est à déterminer. Lorsque $\Theta \subset \mathbb{R}^K$, une autre approche est de construire des intervalles de confiance simultanés. À propos de bootstrap et régions de confiance en dimension plus grande que 1, voir notamment la section 4.2 du livre de Hall [**Hal92**], ainsi que l'article de Beran [**Ber03**].

Avant d'aller plus loin, citons quelques références sur les intervalles et régions de confiance. Avec le bootstrap : Hall [**Hal92**], Efron et Tibshirani [**ET93**], Shao et Tu [**ST95**], DiCiccio et Efron [**DE96**]. Avec le bootstrap à poids échangeables généraux, Hall et Mammen [**HM94**], Barbe et Bertail [**BB95**]. Par sous-échantillonnage : Politis, Romano et Wolf [**PRW99**], Chap. 7.

Il y a deux types d'intervalles de confiance classiques par rééchantillonnage : les intervalles «bootstrap-t» et les intervalles «percentile». Un *intervalle bootstrap-t* suppose connus un estimateur $\widehat{\theta}\left(\xi_{1\ldots n}\right)$ de θ, ainsi qu'un estimateur $\widehat{\sigma}^2\left(\xi_{1\ldots n}\right)$ de la variance de cet estimateur. L'idée d'un tel intervalle est d'estimer par rééchantillonnage la distribution de $(\widehat{\theta} - \theta)/\widehat{\sigma}$. On construit donc un intervalle de la forme

$$I_{\text{bootstrap-t}} := \left[\widehat{\theta}\left(\xi_{1\ldots n}\right) - \widehat{\sigma}\left(\xi_{1\ldots n}\right)\widehat{g}^{(\alpha)}\left(\xi_{1\ldots n}\right); \widehat{\theta}\left(\xi_{1\ldots n}\right) - \widehat{\sigma}\left(\xi_{1\ldots n}\right)\widehat{d}^{(\alpha)}\left(\xi_{1\ldots n}\right)\right] \ ,$$

où $\widehat{g}^{(\alpha)}$ et $\widehat{d}^{(\alpha)}$ sont des quantiles empiriques de

$$\frac{\widehat{\theta}\left(\xi^\star_{1\ldots n}\right) - \widehat{\theta}\left(\xi_{1\ldots n}\right)}{\widehat{\sigma}\left(\xi^\star_{1\ldots n}\right)} \ ,$$

[64] entre deux instants de la journée, entre deux individus, ou entre deux répétitions légèrement différentes d'une même expérience, sans même parler d'erreurs de mesure éventuelles ou de différences de sensibilités de deux puces à ADN.

conditionnellement à $\xi_{1\dots n}$ (en général calculés avec B échantillons indépendants). Les niveaux de ces quantiles sont fixés en fonction du type d'intervalle voulu : symétrique, avec une masse $\alpha/2$ de chaque côté, de longueur minimale, *etc.*

À l'inverse, un *intervalle bootstrap percentile* ne nécessite pas la connaissance de $\widehat{\sigma}^2(\xi_{1\dots n})$. Il repose sur l'estimation de la distribution de $(\widehat{\theta}-\theta)$:

$$I_{\text{percentile}} := \left[\widehat{\theta}(\xi_{1\dots n}) - \widehat{g}^{(\alpha)}(\xi_{1\dots n}) ; \widehat{\theta}(\xi_{1\dots n}) - \widehat{d}^{(\alpha)}(\xi_{1\dots n})\right]$$

où $\widehat{g}^{(\alpha)}$ et $\widehat{d}^{(\alpha)}$ sont des quantiles empiriques de $\widehat{\theta}(\xi^{\star}_{1\dots n}) - \widehat{\theta}(\xi_{1\dots n})$. De même, les niveaux de ces quantiles sont choisis en fonction du type d'intervalle demandé.

La justification de ces méthodes est asymptotique, *i.e.* les intervalles de confiance construits ont le niveau α requis lorsque la taille de l'échantillon n tend vers l'infini. En poussant à l'ordre suivant l'étude asymptotique, il s'avère que les intervalles bootstrap-t sont plus précis que les intervalles bootstrap percentile (Hall [**Hal92**]). La raison en est que le bootstrap estime mieux une *statistique pivotale*[65] telle que $n^{1/2}\widehat{\sigma}^{-1}\left(\widehat{\theta}-\theta\right)$ qu'une statistique non-pivotale comme $n^{1/2}\left(\widehat{\theta}-\theta\right)$.

Il est cependant possible de corriger les intervalles percentile pour leur biais, ce sont les *intervalles* BC_a[66]. Ils sont alors asymptotiquement aussi performants que les intervalles bootstrap-t, tout en étant plus robustes (en particulier invariants par transformation monotone de Θ). De manière générale, il est conseillé de passer par une statistique pivotale dans les cas «simples», et d'utiliser la correction du biais dans les cas plus difficiles.

Enfin, il est souvent possible d'approcher les intervalles BC_a analytiquement, ce qui réduit considérablement le temps de calcul de ceux-ci, en particulier lorsque θ est de grande dimension. C'est la *méthode* ABC[67].

Tests, tests multiples. Nombre de procédures de test ou de test multiple reposent sur la construction de régions de confiance, via l'argument mentionné en Sect. 1.4.1. Il est également possible de construire directement des tests, soit en estimant un quantile à l'aide de l'heuristique de rééchantillonnage (le niveau étant alors contrôlé asymptotiquement, le plus souvent), soit en utilisant un argument d'invariance de la loi de l'échantillon par certaines transformations sous l'hypothèse nulle. Dans ce dernier cas, on parle de *tests par randomisation* (ou encore *tests par permutations ou par symétrisation*, selon le type d'invariance en jeu), et le contrôle du niveau est en général exact.

Tests par rééchantillonnage direct. Considérons par exemple le cas d'un test où

$$H_0 = \{\, P \text{ t.q. } S(P) = 0 \,\} \quad ,$$

pour une fonctionnelle régulière S. Une procédure de test naturelle est alors

$$T(\xi_{1\dots n}) = \mathbb{1}_{S(P_n) \geq t_\alpha(\xi_{1\dots n})} \quad ,$$

où t_α est un seuil à déterminer. Le choix idéal serait le quantile d'ordre $1-\alpha$ de $S(P_n) = S(P_n) - S(P)$ sous H_0. L'heuristique de rééchantillonnage suggère de l'estimer à l'aide du quantile d'ordre $1-\alpha$ de $\mathcal{L}\left(S(P_n^W) - S(P_n) \,\middle|\, \xi_{1\dots n}\right)$:

$$\widehat{t}_\alpha^{\,W}(\xi_{1\dots n}) := \inf\left\{ t > 0 \text{ t.q. } \mathbb{P}\left(S\left(P_n^W\right) - S\left(P_n\right) > t \,\middle|\, \xi_{1\dots n}\right) \leq \alpha \right\} \quad .$$

[65] c'est-à-dire dont la loi ne dépend pas (au moins asymptotiquement) de paramètres inconnus. Ici, la variance σ^2 de $\widehat{\theta}$.

[66] «Bias Corrected and Accelerated», *i.e.* avec correction du biais et accélération.

[67] «Approximate Bootstrap Confidence intervals», intervalles de confiance bootstrap approchés.

Bickel et Ren [**BR01**] ont alors montré que cette stratégie fonctionne avec le bootstrap pour une certaine classe de tests. Elle peut également fonctionner avec des poids Efron (m), avec $n \gg m \to \infty$, dans des cas où le bootstrap ne fonctionne pas. À propos du choix de m à partir des données, voir également Bickel et Sakov [**BS05**].

Pour un aperçu des méthodes de test par rééchantillonnage, voir aussi Efron et Tibshirani [**ET93**], Chap. 16 (au sujet du bootstrap), Politis, Romano et Wolf [**PRW99**] (au sujet du sous-échantillonnage), et enfin Hall et Mammen [**HM94**] et Janssen et Pauls [**JP03**] (au sujet du bootstrap à poids échangeables).

Tests par randomisation. Lorsque l'échantillon $\xi_{1\dots n}$ possède certaines propriétés d'invariance sous H_0, on peut en tirer parti pour construire un test dont le niveau est contrôlé précisément. L'idée, qui est ancienne (elle remonte au moins aux années 1930 et R.A. Fisher [**Fis35**][68]), est la suivante. Soit $\mathcal{G}$ un groupe fini de transformations de Ξ^n qui laissent invariante la loi de l'échantillon, *i.e.*

$$\forall g \in \mathcal{G}, \qquad \mathcal{L}\left(\xi_{1\dots n}\right) = \mathcal{L}\left(g\left(\xi_{1\dots n}\right)\right) \ .$$

Alors, pour toute statistique de test $S : \Xi^n \mapsto \mathbb{R}$, pour toute distribution $P \in H_0$,

$$\mathbb{P}_{\xi_{1\dots n}\sim P}\left(\operatorname{Card}\left\{ g \in \mathcal{G} \text{ t.q. } S\left(\xi_{1\dots n}\right) < S\left(g\left(\xi_{1\dots n}\right)\right)\right\} < \alpha \operatorname{Card}(\mathcal{G})\right) \leq \alpha \ .$$

De plus, cette borne supérieure est exacte à $1/\operatorname{Card}(\mathcal{G})$ près, puisque $\lfloor \alpha \operatorname{Card}(\mathcal{G}) \rfloor / \operatorname{Card}(\mathcal{G})$ est une borne inférieure. On en déduit que la procédure de test

$$T\left(\xi_{1\dots n}\right) := \mathbb{1}_{\operatorname{Card}\left\{ g\in\mathcal{G} \text{ t.q. } S\left(\xi_{1\dots n}\right)<S\left(g\left(\xi_{1\dots n}\right)\right)\right\}<\alpha \operatorname{Card}(\mathcal{G})} \tag{1.25}$$

est de niveau α. On parle alors de «test exact» car le niveau de T est égal à α (à $1/\operatorname{Card}(\mathcal{G})$ près).

Une telle formulation générale recouvre de nombreux tests classiques, que l'on peut regrouper en deux classes principales :

- les *tests par permutation* : $\mathcal{G}$ est le groupe Σ_n des permutations de $(\xi_1, \dots, \xi_n)$. Par exemple, supposons que $\xi_1, \dots, \xi_m$ sont i.i.d. de loi Q_1 et $x_{m+1}, \dots, x_n$ dont i.i.d. de loi Q_2, pour un certain $1 \leq m < n$. Sous l'hypothèse nulle «$Q_1 = Q_2$», l'échantillon $\xi_{1\dots n}$ est invariant sous l'action de Σ_n. Le test (1.25) peut donc être utilisé pour tester l'*homogénéité*.
  Lorsque $\Xi = \mathcal{X} \times \mathcal{Y}$, on peut aussi considérer le groupe $\mathcal{G} = \Sigma_{n,\mathcal{Y}}$ qui agit sur $\xi_{1\dots n}$ en permutant les deuxièmes variables $Y_{1\dots n}$ uniquement. Par exemple, sous l'hypothèse nulle «$X_{1\dots n}$ est indépendant de $Y_{1\dots n}$», la loi de l'échantillon est invariante sous $\mathcal{G}$. On peut donc utiliser (1.25) pour tester l'*indépendance*.
- les *tests par symétrisation* : lorsque Ξ est un espace vectoriel (par exemple $\mathbb{R}$), on peut considérer le groupe $\mathcal{G} = \{-1, 1\}^n$ qui agit sur Ξ^n de la manière suivante :

  $$\forall g \in \mathcal{G}, \forall \xi_{1\dots n} \in \Xi^n, \quad (g_1, \dots, g_n) \cdot (\xi_1, \dots, \xi_n) := (g_1\xi_1, \dots, g_n\xi_n) \ .$$

  Sous l'hypothèse nulle[69]

  $$H_0 = \left\{ Q^{\otimes n} \text{ t.q. } Q \text{ distribution symétrique et de moyenne nulle sur } \Xi \right\} \ ,$$

  la loi de l'échantillon est invariante sous l'action de $\mathcal{G}$.
  On peut également utiliser un test de symétrisation pour tester H_0 : «$\mu = 0$», lorsque $\xi_1, \dots, \xi_n$ sont i.i.d. de loi symétrique et de moyenne μ.

Le contrôle du niveau étant non-asymptotique et quasi-exact, les tests par randomisation ont de bonnes chances d'être plus puissants que les tests fondés sur l'heuristique de rééchantillonnage. C'est pourquoi Efron et Tibshirani [**ET93**] (Chap. 15) conseillent d'utiliser les premiers lorsque

[68] voir aussi Oden et Wedel [**OW75**] à propos du test de Fisher.

[69] par «symétrique», nous entendons : si ξ est de loi Q, alors $2\mathbb{E}[\xi] - \xi$ est de loi Q.

cela est possible. En revanche, les seconds ont l'avantage de s'appliquer dans un cadre beaucoup plus général, où il n'y a rien à permuter ni à symétriser.

Toutefois, cette comparaison doit être modérée. D'une part, les tests par randomisation reposent plus fortement sur l'hypothèse d'invariance sous l'action de $\mathcal{G}$ que les tests bootstrap équivalents. Ces derniers sont donc plus robustes, par exemple lorsque l'on ne veut tester qu'un paramètre (*e.g.* la moyenne d'un échantillon), que les observations soient indépendantes ou non. Par ailleurs, dans un cadre de tests multiples, nous montrons au Chap. 10 qu'un test multiple rééchantillonnage peut présenter des avantages sur un test multiple par symétrisation.

Pour d'autres références sur les tests par randomisation, on consultera par exemple Romano [**Rom89, Rom90**], Westfall et Young [**WY93**], Janssen et Pauls [**JP03**].

Tests multiples. Les méthodes de test multiple par rééchantillonnage sont essentiellement de trois sortes :

- on construit une région de confiance par rééchantillonnage et on en déduit une procédure de test multiple (ceci est possible lorsque $H_{0,k}$ s'écrit «$\theta_k = \theta_{k,0}$», par exemple). À ce sujet, voir notamment Sect. 10.4.1 et Pollard et van der Laan [**PvdL03**].
- on construit K tests simples par rééchantillonnage, dont on ajuste ensuite les p-valeurs (par exemple par une méthode step-down ou step-up).
- on utilise le rééchantillonnage pour construire une méthode d'ajustement des p-valeurs. Ceci permet d'éviter d'avoir à faire trop d'hypothèses sur la loi jointe des p-valeurs. Voir par exemple Yekutieli et Benjamini [**YB99**], et Ge, Dudoit, Speed [**GDS03**].

Les deux premières approches ont déjà été évoquées dans le reste de cette section. La troisième demanderait de plus longs développements que cette introduction ne saurait accueillir. C'est pourquoi nous renvoyons le lecteur intéressé aux références précitées, auxquelles nous ajoutons Westfall et Young [**WY93**] et Romano et Wolf [**RW05, RW07**].

1.4.4. Contributions de la thèse. Le Chap. 10 est le fruit d'une collaboration avec Gilles Blanchard et Étienne Roquain. Une version courte de ces travaux a été publiée dans les actes de Colt 2007 [**ABR07**]. Nous y construisons des régions de confiance par rééchantillonnage (notamment pour en déduire des procédures de test multiple), avec un *contrôle non-asymptotique* de la probabilité de couverture.

Problème considéré. Nous considérons des observations i.i.d. $\xi_1 = \mathbf{Y}^1, \dots, \xi_n = \mathbf{Y}^n \in \Xi = \mathbb{R}^K$, de loi symétrique ou gaussienne, mais dont les coordonnées sont corrélées, et ces corrélations sont générales. Le paramètre qui nous intéresse est leur moyenne commune $\mu \in \mathbb{R}^K$, pour laquelle nous cherchons à construire des régions de confiance de la forme

$$R(\xi_{1\dots n}) = \left\{ x \in \mathbb{R}^K \text{ t.q. } \phi\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - x)\right) \le t_\alpha(\xi_{1\dots n}) \right\} ,$$

où ϕ est (par exemple) la norme L^p ($p \ge 1$) et t_α un seuil calculé à partir des données.

L'une des applications qui motive ce travail est l'imagerie cérébrale, où μ indique les régions du cortex qui répondent à une stimulation donnée (ce sont les positions k telles que $\mu_k \neq 0$), ainsi que l'intensité de cette réponse. Dans ce cadre comme dans beaucoup d'autres, la dimension K peut être beaucoup plus grande que le nombre de répétitions n. Une approche asymptotique est donc à exclure. De plus, les corrélations n'ont pas de forme particulière, si bien qu'une approche paramétrique n'est pas envisageable (il y aurait au moins K^2 paramètres, contre seulement nK données pour les estimer). Dans un cadre non-paramétrique, il est naturel d'estimer t_α par rééchantillonnage. La principale nouveauté de notre approche est que nous proposons un *contrôle non-asymptotique* de la probabilité de couverture.

Deux méthodes. Le seuil idéal est le quantile d'ordre $1-\alpha$ de $\phi\left(n^{-1}\sum_{i=1}^n \xi_i - \mu\right)$. Nous proposons deux méthodes, chacune fondée sur l'heuristique de rééchantillonnage, afin de l'approcher.

La première méthode repose sur plusieurs *inégalités de concentration* : $\phi\left(n^{-1}\sum_{i=1}^n \xi_i - \mu\right)$ satisfaisant une inégalité de concentration sous-gaussienne (Prop. 10.4 (i)), on peut majorer le seuil idéal par un terme d'espérance

$$\mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n (\xi_i - \mu)\right)\right]$$

plus un terme de déviation ne dépendant que de $\sigma = (\mathrm{var}(\xi_k))_{1\le k\le K}$.

En appliquant l'heuristique de rééchantillonnage, nous définissons t_α de la forme

$$t_{1,\alpha}(\xi_{1\dots n}) := B_W^{-1}\mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n\left(W_i - \frac{1}{n}\sum_{i=1}^n W_i\right)\xi_i\right)\,\middle|\,\xi_{1\dots n}\right] + R_n(\alpha,\sigma)\ ,$$

avec B_W explicite (voir (10.4)) et R_n un terme de reste (voir Thm. 10.1). Remarquons la présence de $n^{-1}\sum_i W_i$ au lieu de 1, qui découlerait d'une application directe de l'heuristique. On peut le voir comme une manière d'imposer $\sum_i W_i = n$, pour que le rééchantillon ressemble vraiment à un échantillon. De façon équivalente, ce terme revient à *recentrer empiriquement* les données :

$$\frac{1}{n}\sum_{i=1}^n\left(W_i - \frac{1}{n}\sum_{i=1}^n W_i\right)\xi_i = \frac{1}{n}\sum_{i=1}^n W_i\left(\xi_i - \frac{1}{n}\sum_{i=1}^n \xi_i\right)\ .$$

Le point important ici est que cela permet à $t_{1,\alpha}$ d'être *indépendant de la vraie moyenne* $\mu\in\mathbb{R}^K$. En particulier, la présence de très grande coordonnées ne le perturbe pas, ce qui est crucial en termes de puissance, pour une application en test multiple.

Nous montrons alors (Thm. 10.1) que les régions de confiance fondées sur le seuil $t_{1,\alpha}$ ont une probabilité de couverture contrôlée pour tout n fixé. Ceci est valable pour une grande classe de fonctions ϕ (dont toutes les normes L^p, $p\ge 1$), et pour des poids W échangeables ou de type «V-fold». Ce résultat repose notamment sur une inégalité de concentration (Prop. 10.4 (ii)) montrant que le seuil rééchantillonné $t_{1,\alpha}(\xi_{1\dots n})$ se concentre autour de son espérance plus rapidement[70] que $\phi\left(n^{-1}\sum_{i=1}^n \xi_i - \mu\right)$. Ce fait remarquable permet notamment de combiner $t_{1,\alpha}$ avec un seuil déterministe sans perdre de niveau (Cor. 10.1).

La deuxième approche est de type «*percentile*» : on estime directement par rééchantillonnage le quantile d'ordre $1-\alpha$ de $\phi\left(n^{-1}\sum_{i=1}^n \xi_i - \mu\right)$, d'où le seuil

$$t_{2,\alpha}(\xi_{1\dots n}) := \inf\left\{t>0 \text{ t.q. } \mathbb{P}\left(\phi\left(\frac{1}{n}\sum_{i=1}^n\left(W_i - \frac{1}{n}\sum_{i=1}^n W_i\right)\xi_i\right) > t\,\middle|\,\xi_{1\dots n}\right)\le\alpha\right\}$$

que nous n'étudions que dans le cas des poids Rademacher (1/2). Notre preuve repose en effet sur des outils de symétrisation, qui ne sont valables que dans ce cadre. La nouveauté réside ici dans le terme en $n^{-1}\sum_i W_i$, qui rend le seuil plus robuste à des moyennes μ élevées, mais rend la preuve plus ardue. C'est pourquoi le résultat que nous prouvons (Thm. 10.2) rend nécessaire l'ajout d'un terme de reste à $t_{2,\alpha}$. Ce dernier peut par exemple être obtenu à partir de $t_{1,\alpha}$, le point important étant qu'il est négligeable devant $t_{2,\alpha}$ lorsque n est raisonnablement grand.

D'après une étude de simulations (Sect. 10.5), cette seconde méthode est plus performante que la première (dans le cas $\phi = \|\cdot\|_\infty$), sauf lorsque les coordonnées de ξ sont presques indépendantes (car on peut alors combiner la première méthode avec un seuil déterministe qui fonctionne bien lorsqu'il y a peu de corrélations). Cependant, en termes de temps de calcul, l'approche percentile

[70] à l'échelle n^{-1} au lieu de $n^{-1/2}$ lorsque W est échangeable ; à l'échelle $V^{1/2}n^{-1}$ dans le cas des poids V-fold.

exige de considérer de nombreux rééchantillons, alors que la méthode par concentration garde une précision raisonnable avec une approximation Monte-Carlo ou des poids V-fold.

Application aux tests multiples. Nous étudions à la Sect. 10.4 comment utiliser ces régions de confiance pour construire des procédures de test multiple. Deux problèmes peuvent être résolus :

- test multiple bilatère : lorsque $H_{0,k}$ est «$\mu_k = 0$», on peut utiliser les seuils $t_{i,\alpha}$ avec $\phi(x) = \sup_k |x_k|$.
- test multiple unilatère : lorsque $H_{0,k}$ est «$\mu_k \leq 0$», on peut utiliser les seuils $t_{i,\alpha}$ avec $\phi(x) = \sup_k (x_k)_+$.

L'avantage d'avoir construit des seuils invariants par translation de la vraie moyenne μ est qu'ils ne sont pas sensibles aux grandes valeurs des coordonnées non-nulles de μ, contrairement aux tests par symétrisation classiques (*cf.* Cor. 10.11). L'utilisation de procédures step-down permet de remédier à ce problème, mais au prix d'un temps de calcul allongé. Dans les applications que nous avons en vue, ramener le temps de calculs de 48 heures à 24 heures est une amélioration considérable.

Une étude de simulations (Sect. 10.5) nous montre que nos deux approches sont compétitives avec les seuils classiques (tels que Bonferroni), lorsque les corrélations entre les coordonnées sont suffisamment importantes. Le prix de l'adaptation à de fortes corrélations est une légère sous-optimalité dans le cas indépendant.

Notations

Conventions.

- the generic constants (L, $L_{p_1,\ldots,p_k}$, $L_{(\mathbf{A})}$) can change from a line to another, or even within the same line.
- a null indicator function always has priority over other terms. For instance, $x^{-1}\mathbb{1}_{x=0}$ is equal to 0 when $x = 0$.

Abbreviations.

a.e.	almost every
a.s.	almost surely
(BA)(m,p)	Bounded assumption (Sect. 10.1.4, p. 248)
Bonf	Bonferroni threshold ((10.11) p. 250)
CV	Cross-validation (Sect. 2.2.2, p. 77)
Efr, penEfr	Efron's bootstrap weights, penalty (Sect. 6.3.3, p. 154)
FDR	False Discovery Rate (Sect. 10.6.2)
FWER	Family-Wise Error Rate (p. 258).
(GA)	Gaussian assumption (Sect. 10.1.4, p. 248)
Loo, penLoo	Leave-one-out weights, penalty (Sect. 6.3.3, p. 154)
LOO	Leave-one-out (Sect. 2.2.2, p. 77)
Lpo	Leave-p-out
Mal	Mallows' C_p penalty
penEfr+, penLoo+, Mal+, ...	penEfr, penLoo, Mal, ... multiplied by a factor 5/4
Poi	Poisson i.i.d. weights (Sect. 6.3.3, p. 154)
Rad, penRad	Rademacher i.i.d. weights, penalty (Sect. 6.3.3, p. 154)
Rho, penRho	Random hold-out weights, penalty (Sect. 6.3.3, p. 154)
RP	Resampling penalties (Sect. 6.2, p. 151)
(SA)	Symmetry assumption (Sect. 10.1.4, p. 248)
s.t.	such that
VF, penVF	V-fold weights, penalty (Sect. 5.3.1, p. 123)
VFCV	V-fold cross-validation (Sect. 2.2.2, p. 77)
w.r.t.	with respect to

Mathematical notations (Chap. 2 to 9).

L, $L(p_1,\ldots,p_k) = L_{p_1,\ldots,p_k}$, $L_{(\mathbf{A})}$	generic constants (resp. numerical, dependent from $p_1,\ldots,p_k$, dependent from the constants appearing in the assumption set ($\mathbf{A}$))
$a \vee b$, $(a)_+$	the maximum of a and b, the positive part of a ($= a \vee 0$)
$a \wedge b$, $(a)_-$	the minimum of a and b, the negative part of a ($= -a \vee 0$)
$a \propto b$	a is proportional to b (*i.e.* $L_1 a \leq b \leq L_2 a$)
$a \ll b$	a is negligible in front of b

$\log = \ln, \ln_2$	natural logarithm, binary logarithm
$\lfloor a \rfloor$	largest integer smaller or equal to a
$\mathbb{1}_E$, $\mathrm{Card}(E)$, $\mathrm{diam}(E)$	indicator function, cardinality, diameter of the set E
E^c, $\mathrm{conv}(E)$	complementary, convex hull of the set E
$\mathfrak{P}(E)$	set of all subsets of E
$\mathcal{H}(\alpha, R)$, $\mathcal{H}_\epsilon(\alpha, R)$	Hölderian balls (Sect. 6.4.2, p. 158)
$\mathrm{varia}_{\mathcal{X}}\, s$	variation of the function $s : \mathcal{X} \mapsto \mathbb{R}$ over $\mathcal{X}$ ((6.14), p. 159)
$\mathbb{P}$, $\mathbb{E}$, $\mathrm{var}_P(X)$, cov	probability, expectation, variance of $X \sim P$, covariance
$\mathcal{L}$, $\mathcal{D}$, $\overset{(d)}{=}$	law (=distribution), equality in distribution
$Q\gamma(t)$	short for $\mathbb{E}_{\xi\sim Q}\left[\gamma(t,\xi) \mid t\right]$. Notice that if t is random (*e.g.* data-dependent), then $\xi \sim Q$ has to be independent from t, so that $Q\gamma\left(t\right)$ is random itself.
$\left\|\cdot\right\|_q$	moment of order q
Leb, δ_ξ	Lebesgue measure, Dirac measure at ξ
$\mathcal{U}(E)$	Uniform distribution on the set E
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution
$\mathcal{B}(n, p)$	Binomial distribution
$\mathcal{M}(n; p_1, \ldots, p_k)$	Multinomial distribution
$\mathcal{P}(\mu)$	Poisson distribution
$\mathcal{H}(n, r, q)$	Hypergeometric distribution
e_Z^+, e_Z^0	expectations of inverses (Sect. 6.7, p. 170)
$\mathcal{X}$, $\mathcal{Y}$, Ξ	feature space, label space, observation space (often $\mathcal{X} \times \mathcal{Y}$)
$(X_i, Y_i)_{1\le i\le n}$, $\xi_{1\ldots n}$	samples
$\epsilon_1, \ldots, \epsilon_n$	error terms
$\mathcal{S}$, S, S_D, S_m, t	set of all predictors, 3 subsets of $\mathcal{S}$, generic predictor
γ, $l(s,t)$, $l(s,S)$	contrast function, excess loss at t, bias of S $(:= \int_{t\in S} l(s,t))$
η, s	regression function, Bayes predictor
$\sigma(x)$	noise level at point $X = x$
w, ϕ_m, $\varepsilon_{\star,m}$	margin condition and complexity measures in a general framework, including binary classification (Sect. 7.2.2, p. 196)
P, s	unknown data distribution, Bayes predictor
s_m	best predictor over S_m
$\widehat{s}$, $\widetilde{s}$	estimators
P_n, $\widehat{s}_m$	empirical distribution, empirical risk minimizer over S_m
$(B_j)_{1\le j\le V}$, I	partition of $\{1, \ldots, n\}$, subset of $\{1, \ldots, n\}$
$P_n^{(I)}$, $P_n^{(j)}$, $P_n^{(B_j)}$, $P_n^{(-j)}$, $P_n^{(B_j^c)}$	subsample empirical distributions
$P_n^{(v)}$, $P_n^{(t)}$	validation and training empirical distributions
$\widehat{s}_m^{(I)}$, $\widehat{s}_m^{(-j)}$, $\widehat{s}_m^{(v)}$	subsample empirical risk minimizers
W, $\epsilon_1, \ldots, \epsilon_n$	resampling weight vector, i.i.d. Rademacher variables
$P_n^\star$, P_n^W	bootstrap or weighted bootstrap empirical distributions
$\widehat{s}_m^W$	weighted bootstrap empirical risk minimizer
$\mathbb{E}_W\left[\cdot\right]$	expectation w.r.t. W only

$(S_m)_{m\in\mathcal{M}_n}$, D_m	family of models, dimension of the model m
$m^\star$	oracle model (Sect. 2.1.2, p. 71)
$m^\star_{\text{lin}}$	linear oracle model (Sect. 4.3, p. 109)
$\widehat{m}$, $\widehat{m}_{\text{VFCV}}$, $\widehat{m}_{\text{Loo}}$, ...	selected models
$\widehat{m}(K)$	model selected by penalization, according to the multiplicative factor K (with linear penalties, Sect. 4.4.1, p. 109 or general shapes, Sect. 11.3.2, p. 286)
C_{or}, $C_{\text{path}-\text{or}}$	model selection procedure benchmarks ((4.3), p. 110)
pen, pen_{VFCV}, $\text{pen}_{\text{Mallows}}$, ...	penalty functions
pen_{id}, $\text{pen}_{\text{id,g}}$	ideal penalty, ideal global penalty ((2.13), p. 76)
pen'_{id}	other ideal penalty (Sect. 5.7.3, p. 139)
pen_{min}	minimal penalty (Sect. 3.3.2, p. 92)
p_1, $\widetilde{p_1}$, p_2, $\widetilde{p_2}$, δ, $\overline{\delta}$	parts of the ideal penalty (Sect. 5.7.2, p. 135)
$\widehat{p}_1$, $\widehat{p}_2$	parts of the Resampling Penalty ((5.32) and (5.33) , p. 138)
$\widehat{C}_n(\mathcal{F})$, $\widehat{R}_n(\mathcal{F})$, $\widehat{G}_n(\mathcal{F})$, $\widehat{B}_n^{(Z)}(\mathcal{F})$	global complexities of the family of functions $\mathcal{F}$ (Chap. 233)
crit, crit$'$, $\text{crit}_{\text{VFCV}}$, ...	model selection minimization criteria
C, $C_{W,\infty}$	constants in front of Resampling Penalties (Sect. 6.2, p. 151)
$R_Z(\mathcal{F})$	constant in front of global resampling complexities ((9.5), p. 234)

Histogram framework (Sect. 6.3, p. 152 and Sect. 6.8.1, p. 171).

$(I_\lambda)_{\lambda\in\Lambda_m}$	partition of $\mathcal{X}$
$\mathbb{E}^{\Lambda_m}[\cdot]$	short for $\mathbb{E}\left[\cdot \mid (\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}\right]$
σ_λ^d, σ_λ^r, σ_λ	variability terms when $X\in I_\lambda$
p_λ, $\widehat{p}_\lambda$, $\widehat{p}_\lambda^W$	true, empirical, resampling empirical frequencies of X in I_λ
W_λ	mean of the weights W_i such that $W_i\in I_\lambda$
$A_n(m)$, $B_n(m)$	minimum of the empirical and true frequencies over $(I_\lambda)_{\lambda\in\Lambda_m}$
β_λ, $\widehat{\beta}_\lambda$, $\widehat{\beta}_\lambda^W$	coordinates of s_m, $\widehat{s}_m$, $\widehat{s}_m^W$ in the basis $(\mathbb{1}_{I_\lambda})_{\lambda\in\Lambda_m}$
$R_{1,W}$, $R_{2,W}$	resampling weights constants ((5.31), p. 138)
$\widehat{\mathcal{M}}$	family of "pre-selected" models
δ_{n,p_λ}, $\delta_{n,p_\lambda}^{(VF)}$, $\delta_{n,\widehat{p}_\lambda}^{(\text{penW})}$	quantities $\ll 1$ when $np_\lambda\to\infty$

Confidence regions and tests (Chap. 10).

$\|x\|_p$	L^p norm of $x\in\mathbb{R}^K$: if $1\leq p<\infty$, $\left(\sum_{k=1}^K \lvert x_k\rvert\right)^{1/p}$ if $p=\infty$, $\sup_{1\leq k\leq K}\lvert x_k\rvert$
$\overline{\Phi}$	standard Gaussian upper tail function
$\overline{\mathcal{B}}(n,\eta)$	Binomial upper tail function (p. 255)
$\mathbf{Y}$	$K\times n$ data matrix: each column $\mathbf{Y}^i$ is an individual observation
μ, Σ	mean and covariance matrix of $\mathbf{Y}^1$
$\overline{\mathbf{Y}}$	empirical mean $n^{-1}\sum_i \mathbf{Y}^i$
$\overline{W}$	mean of the weights $n^{-1}\sum_i W_i$
$\overline{\mathbf{Y}}_{[W]}$	weighted bootstrap empirical mean $n^{-1}\sum_i W_i\mathbf{Y}^i$
$\widetilde{\mathbf{Y}}^j$	mean of the block j (in the V-fold case) $\text{Card}(B_j)^{-1}\sum_{i\in B_j}\mathbf{Y}^i$
A_W, B_W, C_W, D_W, E_W	constants related to resampling weights (Sect. 10.2, p. 249

and Sect. 10.7.5)

$t_{\ldots,\alpha}$	some threshold, at level α
H_0	single null hypothesis
$[x]$	either $\lvert x\rvert$ (two-sided context) or x (one-sided context)
$H_{0,k}$	null hypotheses (in a multiple testing setting)
$R(\mathbf{Y})$	multiple testing procedure (*i.e.* data-dependent rejection set)
$\mathcal{H}$, $\mathcal{H}_0$	set of null hypotheses, set of true null hypotheses
$\mathbf{t}$	subset-based threshold (Sect. 10.4.2, p. 259)
$\mathcal{C}_i(\mathbf{Y}) := \{\sigma(j) \text{ s.t. } j \geq i\}$	the set which contains the $K - i + 1$ smaller coordinates of $\overline{\mathbf{Y}}$.

CHAPTER 2

Optimal model selection

RÉSUMÉ. Ce chapitre a pour but d'introduire et de motiver une partie essentielle de cette thèse, qui regroupe les chapitres 3 à 9, et dont l'objet est la calibration «optimale» de procédures de sélection de modèles. Mettant en évidence les différentes formes d'optimalité (théorique ou pratique, asymptotique ou non), nous montrons qu'il existe un réel fossé entre théoriciens et praticiens sur la manière de résoudre ce problème. Ce travail de thèse cherche à réduire cet écart, en étudiant une méthode de calibration de pénalités à l'aide des données (l'heuristique de pente, Chap. 3), ce qui est nécessaire dans certaines circonstances (Chap. 9) ; en proposant des méthodes d'estimation de la forme de la pénalité (Chap. 5 à 8), qui sont plus robustes que les simples pénalités linéaires (Chap. 4) ; en proposant une étude théorique non-asymptotique de la validation-croisée V-fold et d'une méthode de pénalisation apparentée (Chap. 5).

The aim of this chapter is to introduce and motivate the main part of this thesis, which goes from Chap. 3 to 9.

Our main motivation is to fill in a gap between theory and practice in model selection. On the one hand, theoretical results concern procedures that are either untractable (because they need a huge computation time, or — worse — because they make use of unknown parameters) or based upon a single data splitting. For instance, in binary classification, fast rates of convergence under margin conditions have only been obtained for local Rademacher complexities (Koltchinskii [**Kol06**]), the hold-out (Massart [**Mas07**], Blanchard and Massart [**BM06d**]) and some aggregation procedures (Lecué [**Lec07a**]). The first one makes use of unknown constants (and is computationally untractable), the last two ones rely on a single split of the data. Moreover, hold-out and aggregation are proved to be optimal compared to estimators built with a first part of the data, provided that the second part of the data is large enough. So, they may not be optimal compared with estimators built with all the data. The best choice of the splitting remains another crucial open problem.

On the other hand, practical users often prefer V-fold cross-validation (VFCV), which selects an estimator built with all the data. This procedure has several advantages: it is very simple to explain, computationally feasible (since V is mainly chosen between 5 and 10), quite stable (because it does not rely on a single split) and it does not rely on strong assumptions. Simulation studies always show that the hold-out has quite poor performances, whereas VFCV does generally much better. Surprisingly, there are very few theoretical results on VFCV, which is far less understood than the hold-out. The reason for this is that a single data split allows to make use of the independence between the two parts of the sample. With VFCV, all the data is used for both fit and selection, making theoretical results very hard to prove.

Our answer to this issue is two-fold. First, we suggest a practical way of tuning penalization procedures. This is based upon the "slope heuristics" from Birgé and Massart [**BM06c**], for

which we prove results in an heteroscedastic framework (Chap. 3 and 4). Secondly, we study the non-asymptotic performances of VFCV and define alternative penalization methods (V-fold or resampling penalties, Chap. 5 to 8). In particular, we prove sharp oracle inequalities for these procedures in a regression framework, with heteroscedastic noise (Chap. 5 and 6). These penalties do not involve unknown constants, except maybe a multiplicative one (like global resampling penalties, see Chap. 9), for which we can use the slope heuristics idea of Chap. 3.

The main drawback of our theoretical evidence is that they often assume a particular structure for the models, since explicit computations can only be made with histogram models. However, we are able to make part of the way towards an oracle inequality in a general framework, including bounded regression and binary classification (Chap. 7). We suggest to make the following use of our results. For theorists, our complete proofs for histograms could be a *guideline towards sharp oracle inequalities* in a general framework. At least, they enlighten conjectures that are likely to be true, and several difficulties that will have to be solved. For practical users, our accurate description of the histogram case shows *how resampling can be successfully used for model selection*, *what does not work in general* (without modification), and *which modifications could improve on the performance* of resampling in model selection.

2.1. Model selection for prediction

The common goal of users and theorists of model selection is to build *optimal model selection procedures*. In order to explain what this means, let us focus on the prediction problem, which is the main framework of this thesis.

2.1.1. The prediction framework. The prediction problem is part of statistical learning theory, initiated by the works of Vapnik [**Vap82, Vap98**]. It can be described as follows. We observe n independent realizations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ of a random variable (X, Y) of unknown distribution P. Given a new independent realization (X_{n+1}, Y_{n+1}) of (X, Y), we would like to predict Y_{n+1} (a quantity of interest) thanks to X_{n+1} (some feature parameters, easier to measure) and the past data $(X_i, Y_i)_{1 \leq i \leq n}$. In other words, we would like to build a data-dependent *predictor* $t : \mathcal{X} \mapsto \mathcal{Y}$. We have in mind several frameworks among which:

- regression: X contains several feature parameters (*e.g.* $\mathcal{X} \subset \mathbb{R}^k$), and Y is a signal of interest belonging to a continuous space (*e.g.* $\mathcal{Y} \subset \mathbb{R}$). We can then write $Y = \eta(X) + \sigma(X)\epsilon$ where $\eta(X) = \mathbb{E}[Y \mid X]$ is the regression function, and $\sigma(X)\epsilon$ a centered noise term (with variance $\sigma(X)^2$ conditionally to X; $\sigma(x)$ thus quantifies the noise-level at $X = x$).
- binary classification: Y is a label ($\mathcal{Y} = \{0, 1\}$), and X can be of various kinds (a DNA sequence, a digital image, a curve, to name but a few), generally high dimensional.

In order to define what is a good predictor, we need some measure of the "distance" between $t(X_{n+1})$ and Y_{n+1}. Let $\mathcal{S}$ be the set of predictors. Given a contrast function $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$, the quality of a predictor t is measured by the *prediction loss*

$$P\gamma(t) := \mathbb{E}_{(X,Y)\sim P}[\gamma(t, (X, Y)) \mid t] = \mathbb{E}[\gamma(t, (X_{n+1}, Y_{n+1})) \mid t] \quad .$$

In the following, we will often make use of the above functional notation $P\gamma(t)$ for expectations. The conditionning w.r.t. t means that the prediction loss of a data-dependent predictor is also data-dependent. It is often convenient to consider the *excess loss*

$$l(s, t) = P\gamma(t) - \inf_{t \in \mathcal{S}} \{P\gamma(t)\} \geq 0$$

instead of the loss. When this infimum is actually a minimum, we call *Bayes predictor* any predictor s of minimal prediction loss over $\mathcal{S}$. Remark that the quantity $l(s,t)$ is well-defined even if s is not. In the following, we focus on some frameworks where s exists.

For instance, in both regression and binary classification, a common contrast function is $\gamma(t,(x,y)) = (t(x)-y)^2$ (in classification, this is the 0-1 loss $\mathbb{1}_{t(x)\neq y}$). Then, in regression, the Bayes predictor s is equal to the regression function η and the excess loss can be written

$$l(s,t) = \mathbb{E}\left[\left(t(X)-s(X)\right)^2\right] \quad .$$

In binary classification, the Bayes predictor is $s : x \mapsto \mathbb{1}_{\eta(x)\geq\frac{1}{2}}$, and the excess loss is

$$l(s,t) = \mathbb{E}\left[\left|t(X)-s(X)\right|\left|2\eta(X)-1\right|\right] \quad .$$

A common way of defining a data-dependent predictor with a small excess loss is *empirical risk minimization*. Given a set S_m of predictors (called a *model*), the empirical risk minimizer is defined by

$$\widehat{s}_m \in \arg\min_{t\in S_m}\left\{P_n\gamma\left(t\right)\right\} \qquad \text{where} \qquad P_n\gamma\left(t\right) := \frac{1}{n}\sum_{i=1}^{n}\gamma\left(t,(X_i,Y_i)\right)$$

is the empirical risk. Of course, the excess loss of $\widehat{s}_m$ strongly depends on the choice of the model S_m. On the one hand, if one takes the whole set of predictors $\mathcal{S}$ as a model, there will be many predictors with an empirical risk equal to zero, all with a large excess loss because the data is noisy. On the other hand, if S_m is small, it generally does not contain s, so that

$$l(s,\widehat{s}_m) \geq \inf_{t\in S_m}\left\{l(s,t)\right\} := l(s,S_m) = l(s,s_m)$$

which can be large (s_m denotes a minimizer of the prediction loss over S_m, when it exists). This lower bound $l(s,S_m)$ is called the *bias of* S_m.

There is thus a *trade-off between bias and variance*, and one has to balance these two terms in order to choose a good model S_m:

$$l(s,\widehat{s}_m) = l(s,s_m) + P\left(\gamma\left(\widehat{s}_m\right)-\gamma\left(s_m\right)\right) \quad . \tag{2.1}$$

The first term is the bias of S_m, the second one is a *variance term*: it represents the difficulty of estimating s_m because of the noise. For instance, in the framework of homoscedastic regression on a fixed-design, its expectation is equal to $\sigma^2 D_m n^{-1}$, where D_m is the dimension of the model m.

2.1.2. Sharp oracle inequalities.

Model selection. We now come to the *model selection problem*: given a family of models $(S_m)_{m\in\mathcal{M}_n}$, which model m gives the best predictor $\widehat{s}_m$? That is, according to (2.1), which model achieves the bias-variance trade-off? Since our goal is prediction, an ideal procedure would select the *oracle model*

$$m^\star \in \arg\min_{m\in\mathcal{M}_n}\left\{l(s,\widehat{s}_m)\right\} \quad .$$

However, the oracle depends on the true distribution P, so that it can not be computed in practice. We would now like to make a few remarks:

- the Bayes predictor s is not assumed to belong to $\bigcup_{m\in\mathcal{M}_n} S_m$ (think for instance of the regression framework, with a smooth target s and the family $(S_D)_{1\leq D\leq n}$ of models, where S_D is the set of regular histograms of size D).

- the family $\mathcal{M}_n$ is allowed to depend on the sample size n, as in the above example. This motivates a *non-asymptotic approach*, in which $\mathrm{Card}(\mathcal{M}_n)$ (and sometimes the dimension of the data X) can be much larger than n, even if the sample size itself is large.
- even if $s \in S_{\widetilde{m}}$ for some $\widetilde{m} \in \mathcal{M}_n$, s does not necessarily belong to the oracle model $S_{m^\star}$. For instance, assume that $S_{\widetilde{m}}$ has a large dimension, n is small and the noise-level σ large. Then, $\widetilde{m}$ does not realize the bias-variance trade-off, and the oracle $m^\star$ is a much smaller model (with very high probability). This shows that identification (*i.e.* find $\widetilde{m}$ with high probability) and prediction are quite different goals (sometimes conflicting, see for instance Yang [**Yan05**]).

A common model selection technique is the so-called *structural risk minimization* (Vapnik [**Vap82**]). Starting from the fact that the resubstitution error $P_n\gamma(\widehat{s}_m)$ underestimates the prediction loss $P\gamma(\widehat{s}_m)$ (and would lead to always choose the largest model), the idea is to "penalize" the models for their complexity. More precisely, we choose the model

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \right\} \quad , \tag{2.2}$$

where $\mathrm{pen} : \mathcal{M}_n \mapsto [0,\infty)$ is a data-dependent *complexity measure* of S_m. For instance, in homoscedastic regression, Mallows' C_p corresponds to the penalty $2\sigma^2 D_m n^{-1}$. More on penalization can be found in Massart's Saint-Flour lecture notes [**Mas07**].

Oracle inequalities. The goal of model selection is to choose $\widehat{m}\left((X_1,Y_1),\ldots,(X_n,Y_n)\right) \in \mathcal{M}_n$ such that $\widehat{s}_{\widehat{m}}$ performs almost as well as the oracle $m^\star$, while using only the data. There are at least three theoretical ways of measuring the performance of such a model selection procedure:

- *Asymptotic optimality*:

$$\mathbb{P}\left(\frac{l(s,\widehat{s}_{\widehat{m}})}{\inf_{m\in\mathcal{M}_n}\left\{ l(s,\widehat{s}_m) \right\}} \xrightarrow[n\to\infty]{} 1 \right) = 1 \ . \tag{2.3}$$

- *A non-asymptotic oracle inequality*:

$$\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}}) \right] \le C \inf_{m\in\mathcal{M}_n} \left\{ \mathbb{E}\left[l(s,\widehat{s}_m) + R(m,n) \right] \right\} \quad , \tag{2.4}$$

  for some constant $C \ge 1$ (as close to 1 as possible), and a remainder term $R(m,n)$ that should be small in front of $l(s,\widehat{s}_m)$. Remark that (2.4) compares $\widehat{m}$ to the best deterministic choice of m, which performs worse than the oracle $m^\star$ (which is data-dependent). This is why we will prefer the following kind of oracle inequality, harder to prove, but more meaningful:

$$\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}}) \right] \le C\mathbb{E}\left[\inf_{m\in\mathcal{M}_n} \left\{ l(s,\widehat{s}_m) + R(m,n) \right\} \right] \ . \tag{2.5}$$

- *A "pathwise" oracle inequality*: with high probability (*e.g.* $1 - Ln^{-2}$, for some constant L),

$$l(s,\widehat{s}_{\widehat{m}}) \le C \inf_{m\in\mathcal{M}_n} \left\{ l(s,\widehat{s}_m) + R(m,n) \right\} \ . \tag{2.6}$$

  The difference with (2.5) is that we here compare $\widehat{m}$ to $m^\star$ for almost all the samples, whereas (2.5) only compares their performances in expectation. Since in practice there is one sample, (2.6) provides a stronger guarantee of performance than (2.5). Notice also that (2.6) easily implies (2.5) when the contrast is uniformly bounded by some constant $B < \infty$ (up to a small enlargement of the remainder term).

Adaptivity. Instead of an oracle inequality, an appreciated property of a model selection procedure is *adaptivity* (see for instance Birgé and Massart [**BM97**]). Basically, assume that the

distribution $P \in \mathcal{P} = \bigcup_{\beta \in \mathcal{B}} \mathcal{P}_\beta$, the parameter β representing some unknown property of P (*e.g.*, in the regression framework, one can assume that s belongs to some hölderian ball $\mathcal{H}(\alpha, R)$ without knowing the true parameter $\beta = (\alpha, R) = \beta_0$). Then, we say that $\widehat{s}_{\widehat{m}}$ is *adaptive to* β if it performs as well as any predictor $\widehat{s}_{\beta_0}$ using the knowledge of the true parameter β_0, while not using it.

A common way of measuring this performance is the *minimax risk*: given a family $\mathcal{P}$ of probability distributions, the minimax risk over $\mathcal{P}$ is defined as

$$\mathcal{R}_{\min\max}(\mathcal{P}) := \inf_{\widetilde{s}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[l(s, \widetilde{s})\right] \quad ,$$

where the infimum is taken over the set of all estimators. The minimax risk thus measures the worst case over the class $\mathcal{P}$. We can now give an accurate definition of a *minimax adaptive estimator* $\widehat{s}_{\widehat{m}}$: for every $\beta_0 \in \mathcal{B}$, for every true distribution $P \in \mathcal{P}_{\beta_0}$,

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \leq K \mathcal{R}_{\min\max}(\mathcal{P}_{\beta_0}) \tag{2.7}$$

for some constant[1] K. Of course, the smaller K, the best performances for the estimator $\widehat{s}_{\widehat{m}}$.

Margin adaptivity. In the regression setting, we have already mentioned the smoothness of s as a parameter to which adaptation is often looked for. In the classification setting, a current theoretical challenge is the construction of *margin adaptive procedures*. We briefly explain the meaning of this phrase. When $\mathcal{P}$ is the set of all distributions such that s belongs to some fixed Vapnik-Červonenkis class S of dimension V, the minimax risk over $\mathcal{P}$ is proportional to $V^{1/2} n^{-1/2}$, and it is attained by the empirical risk minimizer over S (see for instance Lugosi [**Lug02**]).

This lower bound is over-pessimistic since the class $\mathcal{P}$ is huge (the above minimax risk is thus called "global minimax risk"). It is possible to do better if P satisfies the *margin condition* introduced by Mammen and Tsybakov [**MT99**]:

$$\mathrm{var}_P\left(\gamma(t, \cdot) - \gamma(s, \cdot)\right) \leq w\left(l(s, t)\right) \tag{2.8}$$

for some nondecreasing function $w : (0, \infty) \mapsto (0, \infty)$ such that $x \mapsto w(x)/x$ is nonincreasing. For instance, if (2.8) holds with $w(\epsilon) = h^{-1}\epsilon^\theta$ for some $\theta \in (0, 1]$, then Tsybakov [**Tsy04**] showed that *fast rates of convergence* (*i.e.* $n^{-\alpha}$ for $\alpha \in (1/2, 1)$ depending on θ and the model) could be obtained by empirical risk minimization over S (with some additional complexity assumption on S in terms of entropy with bracketing). Moreover, if $|2\eta(X) - 1| \geq h > 0$ a.s., then (2.8) holds with $w(\epsilon) = h^{-1}\epsilon$. Under this assumption, Massart and Nédélec [**MN06**] showed that the empirical risk minimizer over a VC-class S of dimension V has a risk of order $V/(nh)$, which is the minimax rate over

$$\mathcal{P}(S, h) := \left\{ P \text{ s.t. } s \in S \text{ and } \mathbb{P}\left(|2\eta(X) - 1| \geq h\right) = 1 \right\} \quad .$$

A procedure is said to be margin adaptive when it adapts to the margin parameter h, or — more generally — to the unknown margin function w (in particular the exponent θ). This is a quite interesting property since it means to attain fast rates when they are available, and not only the global (pessimistic) rate $n^{-1/2}$.

What is a sharp oracle inequality*?* By "sharp oracle inequality", we mean a non-asymptotic pathwise oracle inequality like (2.6) (or (2.5) if one can not do better), with a constant $C = 1 + \epsilon_n$ ($\lim_{n \to \infty} \epsilon_n = 0$, with an explicit upper bound on ϵ_n), and a remainder term $R(m, n) \ll l(s, \widehat{s}_m)$

[1] in the best case, K is an absolute constant; in general, this can only be proved with K depending on β_0 itself, but never on P or the sample size n.

(at least for the "good" models and with large probability). In other words, this is a *non-asymptotic version* of asymptotic optimality (2.3).

There are two main reasons why we want such a result:

- predict as well as possible from the family of predictors $(\widehat{s}_m)_{m\in\mathcal{M}_n}$, even when n is not large compared to $\mathrm{Card}(\mathcal{M}_n)$.
- use a well-chosen family $(S_m)_{m\in\mathcal{M}_n}$ in order to build an adaptive estimator $\widehat{s}_{\widehat{m}}$. Since we know quite well how to choose the model S_m when P satisfies some property like the margin condition, this is almost straightforward as soon as (2.6) holds. Moreover, if the constant C in (2.6) is small and the remainder term $R(m,n)$ negligible, then (2.7) is likely to hold with a constant K close to 1.

However, a sharp oracle inequality is not a sufficient condition for non-asymptotic optimality, which is what matters in practice. When the sample size n is fixed, the optimal constant C in a sharp oracle inequality[2] is $C_n^\star > 1$ (it may also depend on other parameters like the noise-level $\sigma > 0$ or the margin w). Thus, knowing that $C \leq 1+\epsilon_n \to_{n\to\infty} 1$ is necessary, but not sufficient to show that $C \approx C_n^\star$. Moreover, *an asymptotically sub-optimal procedure (for instance an overpenalizing one) can be optimal for some sample size n* (see *e.g.* the simulations of Chap. 5 and 6).

To our knowledge, this fact is seldom taken into account, although it may be crucial for improving model selection techniques in practice. This is why we propose in this thesis *flexible procedures*, in the sense that one can use them with any overpenalization factor. Although we will only prove sharp oracle inequalities when the overpenalization factor is close to one, flexibility makes possible an optimal use in practice. Since theoretical results about non-asymptotic optimality would be quite hard to prove, it is crucial to ensure — at least — the potentiality for such an optimality. Then, an empirical optimization by practical users is likely to produce actually optimal procedures.

Nevertheless, *non-asymptotic optimal model selection* remains a theoretical open problem, which is of crucial interest for practical applications. This is probably a major reason for the gap between theory and practice, to which the next section is devoted.

2.2. A gap between theory and practice

2.2.1. Theory: hold-out, aggregation and local Rademacher complexities. In this section, we describe three strategies for selecting a model m among $\mathcal{M}_n$ (or aggregating several models), which are "good for theorists". This mainly means that some results can be proven (even adaptivity to the margin), sometimes quite easily (*e.g.* for the hold-out). We thus have a deep theoretical understanding of the reasons why they work well (at least, in theory).

Hold-out. The simplest model selection procedure is probably the hold-out. It relies on the idea that the downward bias of the resubstitution error comes from the dependence between $\widehat{s}_m$ and the data used to choose a model m. Then, splitting the sample into two separate (thus independent) parts should avoid this drawback. The first part of the data (called *training sample*, of size n_t) is used to build a family of estimators $\left(\widehat{s}_m^{(t)}\right)_{m\in\mathcal{M}_n}$. For instance, they may be empirical risk minimizers over the training sample:

$$\widehat{s}_m^{(t)} \in \arg\min_{t\in S_m} P_n^{(t)}\gamma\left(t\right) \quad ,$$

[2] the remainder term being fixed, for instance equal to zero or n^{-2}, as well as the probability $1 - Ln^{-2}$ of the favourable event.

where $P_n^{(t)}$ is the training empirical distribution. Then, we use the second part of the data (called *validation sample*, of size $n_v = n - n_t$) to estimate the prediction loss of $\widehat{s}_m^{(t)}$ for each m, and choose the model $\widehat{m}$ with the smaller estimated loss:

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n^{(v)} \gamma\left(\widehat{s}_m^{(t)}\right) \right\} \ ,$$

where $P_n^{(v)}$ denotes the validation empirical distribution.

The idea behind validation is quite simple: since $\widehat{s}_m^{(t)}$ is independent from $P_n^{(v)}$, $\widehat{m}$ minimizes an unbiased estimate of the prediction loss of $\widehat{s}_m^{(t)}$, so that it is likely to satisfy an oracle inequality of the form

$$l(s, \widehat{s}_{\widehat{m}}^{(t)}) \leq \inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m^{(t)}) + R(m, n_t, n_v) \right\} \tag{2.9}$$

with large probability. For instance, in the binary classification framework with a margin condition satisfied, Blanchard and Massart [**BM06d**] (see also Massart [**Mas07**], Sect. 8.5) gave a quite simple proof of the *margin adaptivity of the hold-out*.

However, hold-out has several drawbacks. First, looking again at (2.9), we see that $\widehat{m}$ is compared to the best predictor built with $n_t < n$ observations (and n_t can not be taken too close to n, otherwise $R(m, n_t, n_v)$ may be too large). Even if we used $\widehat{s}_{\widehat{m}}$ as a final estimator, $\widehat{m}$ is chosen according to a criterion which is close to the ideal one when the sample size is n_t, not n. As a consequence, hold-out may not satisfy sharp oracle inequalities like (2.6), and the choice of an optimal splitting ratio n_t/n_v remains an open problem.

Second, practical studies show that the hold-out procedure has poor performances, because of the variability of the criterion $P_n^{(v)} \gamma\left(\widehat{s}_m^{(t)}\right)$. A common way of reducing variability is making several data splits instead of only one (which has to be chosen arbitrarily), *i.e.* use *cross-validation* instead of hold-out. Unfortunately, this makes theory much more difficult, since we no longer have independence between the estimators and the way they are chosen. We will come back to cross-validation strategies in Sect. 2.2.2 and 2.4.

Aggregation. Instead of selecting one predictor among $(t_m)_{m \in \mathcal{M}_n}$, aggregation produces a convex combination of them (Nemirovski [**Nem00**]):

$$\widetilde{s}_{\text{agreg}} = \sum_{m \in \mathcal{M}_n} w_m t_m \qquad \text{with} \qquad \sum_{m \in \mathcal{M}_n} w_m = 1 \ .$$

The predictors t_m are generally assumed to be fixed, and the weights w_m are estimated from the data. Then, most of the theoretical results compare $\widetilde{s}_{\text{agreg}}$ to the best predictor among $(t_m)_{m \in \mathcal{M}_n}$: with large probability,

$$l(s, \widetilde{s}_{\text{agreg}}) \leq \inf_{m \in \mathcal{M}_n} \{ l(s, t_m) \} + R_n \ . \tag{2.10}$$

As for hold-out, these theoretical results rely on the independence between the weights w_m and the predictors t_m. In practice, this means that one will have to split the data into two parts: the predictors $t_m = \widehat{s}_m^{(t)}$ are built with the training sample, and the weights $w_m\left(P_n^{(v)}\right)$ are computed with the validation sample.

Several results like (2.10) about aggregation have been proved, with an "optimal" remainder term R_n. In particular, it can be used to define a margin adaptive procedure (Lecué [**Lec07a**]). Moreover, it can be better to use aggregation than model selection, once $(t_m)_{m \in \mathcal{M}_n}$ is fixed (Lecué [**Lec07b**]). However, *this does not mean that aggregation is better than model selection.*

The name "oracle inequality" often given to (2.10) can indeed be misleading. First, the predictors t_m are generally built with $n_t < n$ data, so that they have a larger excess loss than $\widehat{s}_m$.

The right-hand side in (2.10) is thus larger than the right-hand side of a "sharp oracle inequality" like (2.6). So, we do not know whether an aggregated predictor performs better than model selection among estimator built with the whole data set.

Moreover, $\widetilde{s}_{\text{agreg}}$ belongs to the set $\text{conv}\left((t_m)_{m\in\mathcal{M}_n}\right)$ of convex combinations of the predictors $(t_m)_{m\in\mathcal{M}_n}$, which is much larger than $\{t_m \text{ s.t. } m\in\mathcal{M}_n\}$. There are some results which compare $\widetilde{s}_{\text{agreg}}$ to the best predictor in $\text{conv}\left((t_m)_{m\in\mathcal{M}_n}\right)$ (Juditsky and Nemirovski [**JN00**], Yang [**Yan04**]):

$$l(s,\widetilde{s}_{\text{agreg}}) \leq \inf_{t\in\text{conv}\left((t_m)_{m\in\mathcal{M}_n}\right)} \{l(s,t)\} + R'_n \ . \tag{2.11}$$

However, it is not clear whether (2.11) substantially improves (2.10), because the remainder term R'_n has to be larger than R_n.

In addition, when $t_m = \widehat{s}_m^{(t)}$, the oracle aggregation predictor should be the best predictor in $\text{conv}\left((\widehat{s}_m)_{m\in\mathcal{M}_n}\right)$. To our knowledge, there is no general theoretical result about aggregation with this last benchmark. Moreover, in this case, one has to choose a splitting ratio n_t/n_v, on which the performances of aggregation may strongly depend. Since the optimal ratio can depend on the data (particularly in a non-asymptotic situation), it is quite unclear that one can obtain a sharp oracle inequality for practical aggregation procedures.

Finally, notice that the aggregated predictor $\widetilde{s}_{\text{agreg}}$ can be quite complex, *e.g.* if $\text{Card}(\mathcal{M}_n)$ is large and the weights w_m are all positive (which is a common situation). It may thus be hard to compute in practical applications.

Local Rademacher complexities. In classification, several model selection procedures are built upon "global penalties", *i.e.* defined by (2.2) with an estimate of

$$\text{pen}_{\text{id,g}}(m) := \sup_{t\in S_m} \{(P-P_n)\gamma(t)\}$$

as a penalty pen. For instance, one can use Rademacher complexities (introduced independently by Koltchinskii [**Kol01**] and Bartlett, Boucheron and Lugosi [**BBL02**]):

$$\widehat{R}_n(m) := \mathbb{E}\left[\sup_{t\in S_m}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i\gamma(t,(X_i,Y_i))\right\}\middle|\ (X_i,Y_i)_{1\leq i\leq n}\right] \tag{2.12}$$

(where $\epsilon_1,\ldots,\epsilon_n$ are i.i.d. Rademacher variables), or other resampling estimates[3] of $\text{pen}_{\text{id,g}}(m)$, *e.g.* Fromont's bootstrap penalties [**Fro07**]. Their main drawback is that they are much too large, compared to the ideal penalty for prediction (defined as the difference between the prediction loss and the empirical risk), *i.e.*

$$\text{pen}_{\text{id}}(m) := (P-P_n)\gamma(\widehat{s}_m) \leq \text{pen}_{\text{id,g}}(m) \ . \tag{2.13}$$

In order to attain the fast rates available under the margin condition, one has to take into account the location of $\widehat{s}_m$ in S_m.

This is why several local Rademacher complexities have been introduced in the last few years (Bartlett, Mendelson and Philips [**BMP04**]; Lugosi and Wegkamp [**LW04**]; Bartlett, Bousquet and Mendelson [**BBM05**]; Koltchinskii [**Kol06**]). Basically, they can be written in terms of the positive fixed point $r^\star$ of

$$\widehat{f}(r) = \mathbb{E}\left[\sup_{t\in S_m,\, c_1 r\leq P_n(\gamma(t)-\gamma(\widehat{s}_m))\leq c_2 r}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i\gamma(t,(X_i,Y_i))\right\}\middle|\ (X_i,Y_i)_{1\leq i\leq n}\right] \ ,$$

[3]Efron's resampling heuristics is described in Sect. 2.4.2. It produces estimators of quantities of the form $F(P,P_n)$ like $\text{pen}_{\text{id,g}}(m)$.

for some constants $c_2 > c_1 \geq 0$ (and at least a multiplicative constant in front of $r^\star$) to be tuned. Most of the results about local Rademacher complexities may be found in Koltchinskii [**Kol06**] and the subsequent discussion. In particular, they could be margin adaptive penalties if one knew how to calibrate them from the data.

The main drawback of these local penalties lies in the constants on which they depend. In a few "easy" frameworks (*e.g.* Koltchinskii [**Kol06**], Sect. 6.1, where s is assumed to belong to one of the models), we only have huge upper bounds on all of them. Hence, they are certainly not optimally calibrated, and well calibrated global penalties may perform better in most of the practical cases. In a more general framework (*e.g.* Koltchinskii [**Kol06**], Sect. 5.2), all the penalties based upon $r^\star$ for which theoretical results are proven depend on unknown quantities such as the function w in the margin condition. This is why they can not be considered as truly adaptive penalties. Moreover, even if the calibration problem was solved, their practical computation would be quite long in general. Finally, it seems quite unnatural (at least for the practical user) to consider quantities as complex as the local Rademacher complexities whereas there are several more natural procedures, which are quite efficient in practice, *e.g.* V-fold cross-validation and Efron's bootstrap penalties [**Efr83**] (see Sect. 2.2.2 and 2.4). The main argument in favour of local Rademacher complexities is theoretical: it is possible to use symmetrization tricks. This may not be a sufficient argument for their use in practice.

2.2.2. Practice: V-fold cross-validation. Despite the theoretical results detailed in the previous section, the most widely used model selection technique (in particular in classification) remains cross-validation, for which theory is much harder to derive. It is a basic improvement on the hold-out idea, where the data split is repeated several times.

The "ordinary cross-validation", also called *leave-one-out* (Allen [**All74**], Stone [**Sto74**], Geisser [**Gei75**]) uses every single observation as validation sample:

$$\widehat{m}_{\mathrm{loo}} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \gamma\left(\widehat{s}_m^{(-i)}, (X_i, Y_i)\right) \right\}$$

$$\text{where} \qquad \widehat{s}_m^{(-i)} \in \arg\min_{t\in S_m} \left\{ \frac{1}{n-1}\sum_{j\neq i} \gamma\left(t, (X_j, Y_j)\right) \right\} .$$

Then, the final leave-one-out estimator is $\widehat{s}_{\widehat{m}_{\mathrm{loo}}}$. In regression, cross-validation has been proved to be asymptotically optimal (Li [**Li87**]), but it is often criticized for its variability in classification (in particular when the algorithm producing $\widehat{s}_m$ is unstable, *e.g.* CART; *cf.* Hastie, Tibshirani, Friedman [**HTF01**] and Breiman [**Bre96**]). Then, the *leave-p-out* has been suggested as a generalization of leave-one-out, using every subset of size p of the sample as a validation sample ($n_v = p$, $n_t = n - p$).

Since an exact computation with the leave-p-out requires a huge amount of time, several other approaches have been suggested, among which *V-fold cross-validation* (VFCV, Geisser [**Gei75**]). The idea is to split the sample into V blocks of (almost) equal sizes $(B_j)_{1\leq j\leq V}$, and use each of these blocks as a validation sample:

$$\widehat{m}_{\mathrm{VFCV}} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{V}\sum_{j=1}^{V} P_n^{(B_j)} \gamma\left(\widehat{s}_m^{(-j)}\right) \right\} \qquad \text{where} \quad \widehat{s}_m^{(-j)} \in \arg\min_{t\in S_m} \left\{ P_n^{\left(B_j^c\right)} \gamma\left(t\right) \right\}$$

and $P_n^{(I)} = \mathrm{Card}(I)^{-1} \sum_{i\in I} \delta_{(X_i,Y_i)}$ for every $I \subset \{1, \ldots, n\}$. The computation time is thus reduced to V empirical risk minimizations (instead of n or much more).

Like the hold-out, VFCV is very simple to explain, requires a small computation time (at least when V is chosen smaller than, say, 20), and it can be used in almost every framework. Moreover, VFCV is much more stable than hold-out (at least if $V \geq 5$ and not too close to n), and really builds an estimator with *all the data*.

However, VFCV uses a *biased estimate of the prediction risk* (Burman [**Bur89**]). The predictor $\widehat{s}_m^{(-j)}$ being built with a sample size around $n(V-1)V^{-1}$ instead of n, VFCV slighlty overestimates the prediction loss. In practice, people usually[4] say: "With $V = 10$, the bias is small enough to be neglected". Since this may be wrong asymptotically, Burman proposed a way of correcting VFCV for its bias. Our viewpoint is between those two extreme views, because overestimating the prediction loss may be benefic in some situations (in particular non-asymptotic prediction and identification; see Sect. 2.4 and Chap. 5).

The main issue with VFCV is the *choice of* V. It depends on three factors: bias (the larger V, the smaller bias), variability ($V = 2$ is highly variable, $V = n$ can also be too variable) and computation time (which is proportional to V). Practical users often choose $V = 5$ or $V = 10$, neglecting the bias and considering that the variability is small enough. This is not true asymptotically (V should go to infinity to obtain sharp oracle inequalities), but every practical situation is non-asymptotic. We shall see in the following that large values of V can be worse than $V = 2$, even in situations where the leave-one-out is not more variable than 10-fold cross-validation. In the density estimation framework, Celisse and Robin [**CR06**] also proposed a way of choosing the optimal V from the data, according to a bias-variability trade-off. See also Politis, Romano and Wolf [**PRW99**], Chap. 9, on the choice of V (and, similarly, the choice of p in leave-p-out).

There are few theoretical results on V-fold cross-validation (van der Laan, Dudoit and Keles [**vdLDK04**], Yang [**Yan06, Yan07**], and some references therein), and none is able to distinguish it from leave-p-out (except Celisse and Robin [**CR06**] in a particular density estimation framework). As a consequence, in theory, V can only be chosen according to the bias, leading to *an asymptotically optimal choice of* V.

On the other hand, there are several simulation studies on VFCV and other resampling methods used in practice (*e.g.* the .632 bootstrap, Efron [**Efr83**], and the .632+ bootstrap, Efron and Tibshirani [**ET97**]). Apart from the references already mentioned, see for instance Efron [**Efr86**], Zhang [**Zha93**] and Molinaro, Simon and Pfeiffer [**MSP05**]. One purpose of this thesis is to use theory to enlighten some conclusions of these studies.

2.3. Accurate calibration of penalties

The main drawback of theoretical penalization procedures is often their calibration. In the worst cases, they depend on so many parameters that simulations are necessary to suggest reasonable values of these parameters. This is of course unsatisfactory since simulations can only consider a few examples of distributions P, leading to poor performances in practical problems far from these simulations. In several other cases, the shape of the penalty is known, but not the optimal multiplying factor. Think for instance of Mallows' C_p:

$$\mathrm{pen}_{\mathrm{Mallows}}(m) = \frac{2\sigma^2 D_m}{n} \ ,$$

[4] except when the goal is identification (Zhang [**Zha93**]) or testing (Dietterich [**Die98**], Alpaydin [**Alp99**]) which is much closer to identification than to prediction.

where D_m is the dimension of S_m as a vector space. In general, the (homoscedastic) noise level σ^2 is unknown. In addition, Rademacher complexities (2.12) require a factor 2 in theory which does not seem necessary in practice (Lozano [**Loz00**], Fromont [**Fro07**]).

2.3.1. A practical algorithm for calibration of penalties. Recently, Birgé and Massart [**BM06c**] proposed a practical method for making such a calibration (see their Sect. 4, or also the Sect. 2 of Blanchard and Massart [**BM06d**]). Their idea relies on the following rule of thumb:

$$\text{“optimal” penalty} \approx 2 \times \text{“minimal” penalty} \ . \tag{2.14}$$

Since penalizing less than the minimal penalty implies that the largest models are selected, it can be estimated from the data (as soon as the shape $\mathrm{pen}_0(m)$ of the optimal penalty is known). This leads to Algorithm 3.1, which provides an optimal penalty using only the shape pen_0 (which can be either known *a priori* or estimated by any other device) and the data. Let us recall it now:

(1) For every $K > 0$, compute
$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + K \,\mathrm{pen}_0(m) \right\} \ .$$
(2) Find $\widehat{K}_{\min} > 0$ such that $\widehat{m}(K)$ is very large for $K < \widehat{K}_{\min}$ and of “reasonable size” for $K > \widehat{K}_{\min}$.
(3) Choose the model $\widehat{m} = \widehat{m}\left(2\widehat{K}_{\min}\right)$.

The complexity of the first step being polynomial in $\mathrm{Card}(\mathcal{M}_n)$, this algorithm is computationally tractable. In Sect. 3.4, we give more details on the choice of $\widehat{K}_{\min}$ in practice.

Notice that in (2.14), the “optimal” penalty is the one satisfying a sharp oracle inequality like (2.6). It may not be the optimal one for a fixed sample size, in the sense of the end of Sect. 2.1.2. When some overpenalization within a factor $C_{\mathrm{ov}} > 1$ is needed, one just has to replace the factor 2 in (2.14) by $2C_{\mathrm{ov}}$. Algorithm 3.1 can thus be used to derive non-asymptotic optimal penalties (as soon as C_{ov} can be obtained from the data). When the goal of model selection is *identification*, the optimal penalty is also much larger than $\mathrm{pen}_{\mathrm{id}}$. Think for instance of BIC, which is roughly equal to the prediction criterion AIC multiplied by $\ln(n)/2$. Replacing the factor 2 by $\ln(n)$, (2.14) can then be used to calibrate identification procedures.

2.3.2. The slope heuristics. The reason why (2.14) works is the so-called “slope heuristics”. If one uses
$$p_2(m) := P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right)$$
as a penalty, then the chosen model minimizes $P_n\left(\gamma\left(s_m\right)\right)$ which is an estimate of the bias of S_m. Then, $\widehat{m}$ belongs to the largest models with high probability. And a slight enlargement of the penalty is sufficient to choose smaller models (which achieves a bias-variance trade-off, with an underestimated but positive variance term). On the other hand, the optimal penalty is
$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma\left(\widehat{s}_m\right) \approx p_1(m) + p_2(m) \qquad \text{where} \quad p_1(m) := P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right) \ ,$$
since the resulting $\widehat{m}$ minimizes the prediction loss $P\gamma\left(\widehat{s}_m\right)$. The rule of thumb (2.14) can thus be rewritten as
$$p_1(m) := P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right) \approx P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right) =: p_2(m) \ , \tag{2.15}$$
which is the “slope heuristics”[5].

[5]its name comes from the homoscedastic regression case, where both $\mathbb{E}\left[p_1(m)\right]$ and $\mathbb{E}\left[p_2(m)\right]$ are proportional to the dimension D_m of S_m in expectation. Then, the empirical risk $P_n\gamma\left(\widehat{s}_m\right)$ appears to be linear in D_m when D_m is large enough, with a slope that is equal to the opposite of the minimal constant $\widehat{K}_{\min}$.

Having in mind concentration results, the preceding argument shows that $\mathbb{E}[p_1(m)+p_2(m)]$ is the optimal penalty, whereas $\mathbb{E}[p_2(m)]$ is the minimal one. In the homoscedastic regression framework on a fixed design, Mallows' heuristics relies on the fact that $\mathbb{E}[p_1(m)] = \mathbb{E}[p_2(m)] = \sigma^2 D_m n^{-1}$, so that $2\sigma^2 D_m n^{-1}$ should be an optimal penalty. These computations also imply (2.15) (at least in expectation). Considering the gaussian case, Birgé and Massart [**BM06c**] were able to prove (2.15) on a large probability set, and then justified (2.14) in several cases (including large families $\mathcal{M}_n$, for which the optimal penalty involves an additional $\ln(n/D_m)$ factor).

2.3.3. Our contributions.

Slope heuristics with general shapes of penalty. In Chap. 3, we prove results similar to those of Birgé and Massart (*i.e.* (2.15), and its consequence (2.14)) for heteroscedastic regression on a random design, with a bounded noise. In particular, the ideal penalty is no longer assumed to be a function of the dimension D_m. However, we have to restrict to a particular form of models (that is, histograms) so that explicit computations can be done. The slope heuristics (2.15) thus remains an open problem in heteroscedastic regression in general, but our results suggest that *Algorithm 3.1 should be widely used, with any penalty shape.*

Difficulty of calibration of global penalties in classification. In Chap. 9, we consider several global resampling complexities in classification, among which the well-known Rademacher complexities. It appears that their accurate calibration can be a serious problem in some particular cases. Actually, they should be multiplied by a constant which strongly depends on the unknown distribution P (at least within a factor 2).

While our results are only partial, they suggest two possible answers to this problem:

- prove a tight control of the expectation of global complexities with some additional assumptions on P and the models (our counterexamples suggest that "no classifier has a prediction risk smaller than n^{-1}" may be sufficient). Then, if one believes that this holds true, use these theoretical constants to calibrate Rademacher complexitites.
- use a *data-dependent calibration procedure*, for instance Algorithm 3.1. We do not have theoretical evidence in favour of the slope heuristics in classification, but we conjecture that it can still be used.

Necessity of non linear shapes in heteroscedastic regression. In Chap. 4, we prove that considering general shapes for the penalty is necessary, even in regression, as soon as the noise is heteroscedastic. The main reason for this difficulty is that the ideal penalty is no longer linear in D_m when the noise level $\sigma(X)$ is not constant: if S_m is the set of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$, then

$$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(2+\delta_{n,\mathbb{P}(X\in I_\lambda)}\right)\left(\mathbb{E}\left[\sigma^2(X)\,\middle|\, X\in I_\lambda\right]+\mathbb{E}\left[(s-s_m)^2(X)\,\middle|\, X\in I_\lambda\right]\right)$$

with $\lim_{np\to+\infty}\delta_{n,p}=0$.

More precisely, we prove that linear penalties of the form $K(P_n,P)D_m$ can only select a few models in $\mathcal{M}_n$, which are all far from the oracle in some cases. Then, *even with the knowledge of the true distribution P, linear penalties are suboptimal!* This strong negative result is confirmed with a simulation study. It motivates two kinds of theoretical researches:

- account for the slope heuristics with general shapes of penalties, following the first steps we made in Chap. 3.
- propose penalization techniques that can *estimate the shape of the ideal penalty*, and prove their adaptivity to heteroscedasticity for instance.

In Chap. 5 to 8, we suggest an answer to the second point, *V-fold and Resampling penalties*, and prove their adaptivity to heteroscedasticity in the case of histogram regression. Combined with the slope heuristics, this leads to Algorithm 11.1.

2.4. Contributions on V-fold and other resampling procedures

2.4.1. Performances of V-fold cross-validation. We have already mentioned in Sect. 2.2.2 the definition and main features of V-fold cross-validation in model selection. In particular, since it overestimates the prediction loss (it considers that $n(V-1)V^{-1}$ data are used instead of n), V should not be chosen too small. In the histogram regression framework, we make this statement more accurate thanks to an explicit computation of the expectation of the V-fold criterion

$$\mathrm{crit}_{\mathrm{VFCV}}(m) := \frac{1}{V}\sum_{j=1}^{V} P_n^{(B_j)} \gamma\left(\widehat{s}_m^{(-j)}\right) \quad .$$

Looking at VFCV as if it was a penalization procedure, we compared it to the ideal penalty in expectation:

$$\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m) - P_n\gamma\left(\widehat{s}_m\right)\right] = \left(1 + \frac{1}{2(V-1)} + \epsilon(n,m)\right)\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \tag{2.16}$$

with $\lim_{\min_{\lambda\in\Lambda_m}\{n\mathbb{P}(X\in I_\lambda)\}\to\infty} \epsilon(n,m) = 0$. Thus, *$V$-fold cross-validation overpenalizes within a factor* $1 + 1/(2(V-1)) > 1$, and V has to go to infinity with n in order to obtain asymptotic optimality in general.

The non-asymptotic need for overpenalization. Equation (2.16) is consistent with several known results, in particular those of Burman [**Bur89**] (who suggests to add a correction term, so that V-fold no longer overpenalizes) and Zhang [**Zha93**]. However, we claim that it should be related to the *gain of overpenalization when the signal-to-noise ratio is large*. Although this fact is seldom mentioned, we observed it in a simulation study in regression (Sect. 5.4 and 6.5). Intuitively, taking a slightly enlarged penalty allows to make sure that the large models are sufficiently penalized to avoid their selection, which is likely to appear if the noise-level is high (because the increments of the penalized criterion $\mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$ then have a large variance). If this overpenalization factor is not too large, the selected model is only a bit smaller than the oracle, which induces a loss in performance much smaller than if a huge model had been selected.

Notice that the same phenomenon arises when $\mathrm{Card}(\mathcal{M}_n)$ grows exponentially with n. Indeed, $\mathrm{pen}_{\mathrm{id}}$ has to be multiplied by a factor roughly proportional to $\ln(n/D_m)$ (Birgé and Massart [**BM06c**]) if $\mathrm{Card}(\mathcal{M}_n) \propto e^{an}$, even asymptotically in n. Then, when n is fixed, it is tempting to write $n^k = e^{an}$ with $a = k\ln(n)/n > 0$.

The non-asymptotic need for overpenalization can also be seen in the classical way of proving oracle inequalities for penalization procedures. Using only (2.2), it is straightforward to derive

$$l(s,\widehat{s}_{\widehat{m}}) + \left(\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}\right)(\widehat{m}) \leq \inf_{m\in\mathcal{M}_n}\left\{l(s,\widehat{s}_m) + \left(\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}\right)(m)\right\} \quad . \tag{2.17}$$

This is for instance done in Sect. 1.2.3 on page 34. In order to go from (2.17) to (2.6), it is thus much more important to prove that $\mathrm{pen}(\widehat{m}) \geq \mathrm{pen}_{\mathrm{id}}(\widehat{m})$ (up to some remainder term, to be compared to $l(s,\widehat{s}_{\widehat{m}})$) than to ensure that $\mathrm{pen}(m)$ is not too large. Indeed, the constant C in (2.6) depends on the first point (and it can make C very large, if the remainder term is comparable to $l(s,\widehat{s}_{\widehat{m}})$), whereas the second one only governs the remainder term $R(m,n)$. We do not have theoretical evidence of the non-asymptotic optimality of a wise overpenalization, but we believe that this should be deeply investigated. In Sect. 6.6.1, we propose two natural ways to overpenalize from the data, both based on resampled-quantiles (6.17), but we do not know

whether it works, neither in theory nor in practice. Further remarks on overpenalization are also given in Sect. 11.3.3.

Non-asymptotic optimal choice of V. We now come back to V-fold cross-validation. In a non-asymptotic framework, choosing a small V can be better than a larger one. Even the practical rule "$V = 5$ or $V = 10$ always work well" should be moderated since we observed in several cases that $V = 2$ performs better for prediction[6]. This means that if $V = n$ is asymptotically optimal, *the non-asymptotic optimal choice of V can be $V = 2$*. There is thus a serious gap between theory and practice.

2.4.2. V-fold and resampling penalties.

V-fold penalties. It is unsatisfactory to use two-fold cross-validation, which is a poor improvement on hold-out in terms of variability, even if it sometimes appears to be the best among the VFCV procedures. Moreover, if one wants to identify the true model (*i.e.* the smallest one in $(S_m)_{m\in\mathcal{M}_n}$ to which s belongs), consistency requires to "overpenalize" within a factor of order $\ln(n)/2$ (this is the ratio between BIC and AIC). Shao [**Sha97**] thus showed that leave-p-out is consistent for identification only when $n \sim p$. Having in mind concentration inequalities, V-fold cross-validation is equivalent to the $p = n/V$ case, so that it is not consistent for identification, even with $V = 2$. Notice that VFCV is consistent in some other frameworks (Yang [**Yan07**]), but one could rightfully want to use a V-fold like procedure for identification in general.

Decreasing the variability and providing a more flexible procedure are the main reasons why we propose "V-fold penalization" in Sect. 5.3. It has the same complexity as V-fold cross-validation, and allows to tune separately the variability-complexity trade-off and the bias. It is built upon the *resampling heuristics* (Efron [**Efr79**]), which states that one can mimic the relationship between P and P_n by building a "resample" from the sample distribution P_n. Since the ideal penalty $(P-P_n)\gamma\left(\widehat{s}_m\right)$ can be written $F(P, P_n)$, we can estimate it by resampling. The V-fold resampling scheme is a sort of subsampling (Politis, Romano, Wolf [**PRW99**]): a subsample of the data is chosen uniformly[7] among $\left\{(X_i, Y_i)_{i\notin B_j},\, 1 \le j \le V\right\}$. Then, using the notations of Sect. 2.2.2, we can define the V-fold penalty (penVF) as

$$\mathrm{pen}_{\mathrm{VFCV}}(m) := \frac{C}{V}\sum_{j=1}^{V}\left(P_n - P_n^{\left(B_j^c\right)}\right)\gamma\left(\widehat{s}_m^{\left(B_j^c\right)}\right) \quad , \tag{2.18}$$

where the constant C has to be calibrated.

In order to make V-fold penalties unbiased estimates of $\mathrm{pen}_{\mathrm{id}}$, we showed in the histogram case that one must take $C = V-1$. Interestingly, penVF then turns out to coincide with Burman's corrected V-fold cross-validation (see Remark 5.2). The main novelty with penVF is that one can choose to overpenalize within a factor C_{ov} by taking $C = (V-1)C_{\mathrm{ov}}$.

In Chap. 5, we provide several evidence in favour of penVF. The first one is theoretical: in the histogram regression framework, when $C \sim V-1$, it satisfies a *sharp oracle inequality* similar to (2.6) (Thm. 5.1, Sect. 5.3.3). In particular, it is asymptotically optimal, even when the noise is highly heteroscedastic.

The second evidence comes from a simulation study. In several "difficult" frameworks (heteroscedastic noise, regression function with jumps, to name but a few), it clearly performs better

[6]This is more surprising than for identification or test, as already noticed by Zhang [**Zha93**], Dietterich [**Die98**], Alpaydin [**Alp99**]. See for instance Aerts, Claeskens and Hart [**ACH99**] about the need for overpenalization for controlling the type I error in an identification–testing framework.

[7]and, as in classical VFCV, we then compute expectations w.r.t. this random choice.

than Mallows' C_p and other linear penalties. Moreover, it outperforms VFCV as soon as the overpenalization factor is not taken too small (this is mainly due to the high signal-to-noise ratio). Considering that choosing V for VFCV also needs to know the right overpenalization factor, we can conclude that penVF performs better[8] than VFCV in general.

Finally, the main arguments in favour of penVF should come from their practical use. They have the advantages of both VFCV (simplicity, generality, small computation time, robustness, adaptivity to heteroscedasticity — and probably several other properties —, small variability if V is large enough) and penalization procedures (flexibility, since one may choose between asymptotic optimality, overpenalization and identification, through the tuning parameter C). So, considering that penVF really use *all the data* for both fit and model selection, we hope them to be *non-asymptotically optimal* (provided C is wisely chosen and V can be taken large enough).

Resampling penalties. Using the resampling heuristics in order to build penalties has already been proposed by Efron [**Efr83**] (with a bootstrap resampling scheme) and Shao [**Sha96**] (with the "m out of n" bootstrap). A close procedure, called AICb, has also been proposed by Cavanaugh and Shumway [**CS97**]. Shibata [**Shi97**] showed, in the log-likelihood framework, that Efron's bootstrap penalty, AIC and AICb are all asymptotically equivalent. Then, they are asymptotically optimal for prediction. On the contrary, when the goal is identification, Shao's result is their inconsistency, while the "m out of n" bootstrap with $n \gg m \to \infty$ is consistent.

In Chap. 6, we define a much broader class of penalties, called "Resampling Penalties" (RP), that generalizes Efron's and Shao's penalties. It relies on resampling like V-fold penalties, with an exchangeable weighted bootstrap resampling scheme (Mason and Newton [**MN92**], Præstgaard and Wellner [**PW93**]). The resampling empirical distribution is written

$$P_n^W := \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{(X_i,Y_i)} \ ,$$

where $W = (W_1, \ldots, W_n)$ is an exchangeable weight vector independent from the data. The resampling heuristics then suggests the penalty:

$$\mathrm{pen}(m) = C\mathbb{E}_W \left[\left(P_n - P_n^W \right) \gamma \left(\widehat{s}_m^W \right) \right] \qquad \text{where} \quad \widehat{s}_m^W \in \arg\min_{t \in S_m} \left\{ P_n^W \gamma(t) \right\} \ , \tag{2.19}$$

where $C \geq C_{W,\infty}$ an explicit constant which only depends on the variability of the W_i. Up to the choice of the weights, one can recover Efron's bootstrap penalties, Shao's m out of n penalties, Burman's corrected n-fold cross-validation (which are n-fold penalties with $C = n - 1$). We also propose the use of i.i.d. Rademacher and Random hold-out resampling schemes, which are completely new, up to our best knowledge.

Considering again the histogram regression case, we are able to make explicit computations that *enlighten simultaneously several known results*, from Efron [**Efr83**], Shao [**Sha96**] and Shibata [**Shi97**]. We can indeed provide explicit non-asymptotic bounds on $C_{W,\infty}$ so that (2.19) gives an almost unbiased penalty. For instance, in the m out of n bootstrap case, $C_{W,\infty} = m/n \to 0$ in Shao's framework, while Shao keeps a constant $C = 1$. We can thus conjecture that more generally, *RP with $C \gg C_{W,\infty}$ is consistent for identification.*

We then derive from several concentration inequalities two theoretical "optimality" results. The first one is a *sharp non-asymptotic pathwise oracle inequality* (Thm. 6.1) under several sets of assumptions, that allow heteroscedastic and unbounded noises with general moment conditions.

[8] at least, taking $C = V - 1/2$ leads to a criterion equal to $\mathrm{crit}_{\mathrm{VFCV}}$ in expectation, thus similar performances.

RP thus appear to be theoretically robust (at least when $C \sim C_{W,\infty}$). The second one is *adaptation*[9] *to the smoothness of* s (Thm. 6.2), even when *the noise is heteroscedastic*. Moreover, when s is Lipschitz, RP also appears to *adapt to heteroscedasticity*.

We also perform a simulation study showing that *RP performs even better than penVF*. As a consequence, RP may be quite interesting in the same "difficult cases". Moreover, our simulations allow us to make a comparison between exchangeable weights. If all the penalties are first-order equivalent, second-order terms (as well as non-asymptotic simulations) show the following order[10] between penalties:

$$\text{Rademacher} \approx \text{Random hold-out (half-sampling)} > \text{Leave-one-out} >> \text{Efron's bootstrap} \ ,$$

which is also the order of their performances (remember that overpenalization is needed non-asymptotically). The less biased penalty is the Leave-one-out one, which confirms that *Efron's bootstrap penalty is underpenalizing*, so performs worse than the other ones (but this is mainly due to the choice $C = 1$).

Comparing RP to penVF, our simulation study shows a *small gain in considering exchangeable weights* (at the price of a much longer computation time). This is confirmed by our theoretical concentration results, which show better bounds for exchangeable RP than penVF. However, if computational complexity is a real problem, 10-fold or 20-fold penalties can be chosen without loosing too much.

Actually, the main problem for a non-asymptotic optimal tuning of RP appears to be the overpenalization factor, not the choice of V. With penalties defined by (2.19), we propose in Sect. 6.6 to replace the expectation by a quantile at level α (α having to be chosen by the final user). We have neither theoretical nor simulation evidence for this new procedure, but we believe that it is an interesting research prospect.

Classification. Although the results of Chap. 5 and 6 are restricted to regression on histograms, we expect penVF and RP to be good candidates for being margin adaptive in classification. They are indeed *local penalties*, since they take into account the location of $\widehat{s}_m$ in S_m by estimating directly $\mathrm{pen}_{\mathrm{id}}$ instead of $\mathrm{pen}_{\mathrm{id,g}}$. Compared to local Rademacher complexities, they have several advantages:

- a much *smaller computation time* (in particular penVF).
- they are *easier to calibrate accurately*, since they only involve a multiplicative constant C. Even if one does not believe in $C_{W,\infty}$ (which is known asymptotically, and non-asymptotically for histogram regression), the slope heuristics can be used to calibrate C (*cf.* Sect. 2.3 and Algorithm 11.1).
- they are *more natural*: both RP (including penVF) and local Rademacher complexities use the resampling heuristics (with i.i.d. Rademacher weights for the latest, with more general weights for RP). RP use it to estimate the ideal penalty. Local Rademacher penalties use it to estimate the fixed point $r^\star$ of

$$f(r) = \sup_{t \in S_m,\, c_1 r \leq l(s,t) \leq c_2 r} \left\{ (P - P_n)\left(\gamma(t) - \gamma(s)\right) \right\} \ ,$$

  which is an upper bound on $\mathrm{pen}_{\mathrm{id}}$. One can argue that overpenalizing by estimating $r^\star$ instead of $\mathrm{pen}_{\mathrm{id}}$ can be benefic, but enlarging C with RP or penVF would do the same more naturally and more explicitly.

[9] as soon as we have a lower bound on the variation of s: $\mathrm{varia}_{\mathcal{X}}\, s := \sup_{\mathcal{X}} s - \inf_{\mathcal{X}} s \geq \epsilon > 0$.

[10] in the comparison, "$>$" means a small difference, while "$>>$" means a larger one. Though, remember than these four penalties are all equal at first order.

In order to advocate for the use of penVF and RP in classification, we consider in Chap. 7 a general framework that includes binary classification. Using a concentration result of Boucheron and Massart [**BM04**] on one half of the ideal penalty (p_2), we are able to show a similar concentration property for one half of the resampling penalty, in the case of subsampling weights (*i.e.* V-fold, leave-one-out and Random hold-out). However, there remains several open problems on the way to sharp oracle inequalities for penVF and RP in classification. From the complete proofs we made in the histogram regression case, we propose two ways towards a complete result in classification (one of them including the extension of the slope heuristics (2.15) to classification). In both cases, we would need two difficult results:

- the ideal penalty and its resampling estimates are proportional in expectation (the exact constant may be unknown, if we use the slope heuristics for tuning the penalties).
- concentration inequalities on $P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right)$, in particular *lower bounds* with high probability.

CHAPTER 3

Slope heuristics

RÉSUMÉ. Ce chapitre est consacré à l'étude d'une méthode de calibration de pénalités à l'aide des données, proposée par Birgé et Massart [**BM06c**] : l'heuristique de pente. Nous mettons en évidence l'existence de pénalités minimales, dans un cadre de régression héteroscédastique. Nous prouvons ensuite que le double de la pénalité minimale est «optimal», au sens où la procédure qui en résulte satisfait une inégalité oracle non-asymptotique avec constante presque 1. Il s'ensuit que l'heuristique de pente peut également s'appliquer lorsque la forme optimale de pénalité n'est pas linéaire en la dimension des modèles, ni même une fonction de la dimension.

3.1. Introduction

Model selection has received much interest in the last decades. A very common approach is penalization. In a nutshell, it chooses the model which minimizes the sum of the empirical risk (how does the algorithm fits the data) and some complexity measure of the model (called the penalty). This is the case of FPE (Akaike [**Aka70**]), AIC (Akaike [**Aka73**]) and Mallows' C_p or C_L (Mallows [**Mal73**]).

There is a huge amount of literature about the *efficiency* of such penalization procedures, *i.e.* that their quadratic risk is asymptotically equivalent to the risk of the oracle. This property is often called asymptotic optimality. We mention here the works of Shibata [**Shi81**] about Mallows' C_p and Akaike's FPE and AIC, followed by many other results under other assumptions. See the companion paper of Barron, Birgé and Massart [**BBM99**] and more recent results by Baraud [**Bar00, Bar02**] for more references about this question.

A related problem is how much should we penalize at least? In other words, is there a minimal penalty? In the framework of Gaussian regression on a fixed-design, this question has been addressed by Birgé and Massart [**BM01, BM06c**], and Baraud, Giraud and Huet [**BGH07**] (the latter considering the unknown variance case).

Apart from the theoretical understanding of penalization methods, this question is of much interest from the practical viewpoint. In Sect. 4 of [**BM06c**], Birgé and Massart describe their so-called "*slope heuristics*" (see also Massart [**Mas07**], Sect. 8.5.2). It relies on the fact that twice the minimal penalty is almost the optimal penalty. Then, if one knows that a good penalty has the form $\mathrm{pen}(m) = KF(D_m)$ (where D_m is the dimension of the model and $K > 0$ a tuning parameter), they propose the following strategy for choosing K from the data. Define $\widehat{m}(K)$ the selected model as a function of K. First, compute $K_{\min}$ such that $D_{\widehat{m}(K)}$ is huge for $K < K_{\min}$ and reasonable when $K \geq K_{\min}$. Secondly, define $\widehat{m} := \widehat{m}(2K_{\min})$.

Such a method has been successfully applied for multiple change points detection by Lebarbier [**Leb05**]. Applications are also being developped in several frameworks: mixture models (Maugis

and Michel [**MM07**]), clustering (Baudry [**Bau07**]), spatial statistics (Verzelen [**Ver07**]), estimation of oil reserves (Lepez [**Lep02**]) and genomics (Villers [**Vil07b**]).

However, all the results about minimal penalties concern the homoscedastic fixed-design framework, where the penalty is a function of the dimension, often linear. In this chapter, we prove that a similar phenomenon occurs in the *heteroscedatic random-design* case. Our main advance here is that penalties are neither assumed to be linear in the dimension, nor even functions of the dimension.

From the practical viewpoint, this means that the slope heuristics may be applied when the ideal penalty has a general shape. One can for instance use V-fold or general Resampling penalties defined in Chap. 5 and 6 for estimating the shape of the penalty. The interested reader should refer to Chap. 4 for further considerations about this suggest.

This chapter is organized as follows. We describe our framework and give some notations in Sect. 3.2. Our main theoretical results are stated in Sect. 3.3. We then discuss their practical consequences in Sect. 3.4. All the proofs are given in Sect. 3.5.

3.2. Framework

3.2.1. Regression. We observe some data $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, i.i.d. with common law P. Denoting by s the regression function, we have

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{3.1}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms, possibly dependent from X_i, but with variance 1 conditionally to X_i. Typically, the feature space $\mathcal{X}$ is a compact set of $\mathbb{R}^d$. Throughout this chapter, we make two main assumptions:

- The data is bounded: $\|Y\|_\infty \leq A < \infty$.
- Uniform lower-bound on the noise-level: $\sigma(X) \geq \sigma_{\min} > 0$ a.s.

Given a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$, its quality is measured by the (quadratic) prediction loss

$$\mathbb{E}_{(X,Y)\sim P}\left[\gamma(t,(X,Y))\right] =: P\gamma(t) \qquad \text{where} \quad \gamma(t,(x,y)) = (t(x)-y)^2$$

is the least-square contrast. Then, the Bayes predictor[1] is the regression function s, and we define the excess loss as

$$l(s,t) := P\gamma\left(t\right) - P\gamma\left(s\right) = \mathbb{E}_{(X,Y)\sim P}\left(t(X)-s(X)\right)^2 \quad .$$

Given a particular set of predictors S_m (called a *model*), we define the best predictor over S_m

$$s_m := \arg\min_{t\in S_m} \left\{ P\gamma(t) \right\} \quad ,$$

and its empirical counterpart

$$\widehat{s}_m := \arg\min_{t\in S_m} \left\{ P_n\gamma(t) \right\}$$

(when it exists and is unique), where $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. This estimator is the well-known *empirical risk minimizer*, also called least-square estimator since γ is the least-square contrast.

3.2.2. Model selection. We now assume that we have a family of models $(S_m)_{m\in\mathcal{M}_n}$, hence a family of estimators $(\widehat{s}_m)_{m\in\mathcal{M}_n}$. We are looking for some data-dependent $\widehat{m} \in \mathcal{M}_n$ such that $l(s,\widehat{s}_{\widehat{m}})$ is as small as possible. This is the model selection problem. For instance, we would like

[1] *i.e.* the minimizer of $P\gamma(t)$ over the set of all predictors.

to prove some oracle inequality of the form

$$l(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) \} + R_n$$

in expectation or on a set of large probability, with C close to 1 and $R_n = o(n^{-1})$.

General penalization procedures can be described as follows. Let $\text{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$ be some penalty function, possibly data-dependent. Then, define $\widehat{m} \in \mathcal{M}_n$ which minimizes

$$\text{crit}(m) = P_n \gamma(\widehat{s}_m) + \text{pen}(m) \ .$$

Throughout the chapter, we always assume that the penalty is non-negative. Otherwise, the question of minimal penalties would be ill-posed, since for every $c \in \mathbb{R}$, pen and $\text{pen} - c$ lead to the same model selection procedure.

With such a level of generality, we can show that the calibration of a penalty reduces to the estimation of an "ideal penalty". Indeed, by definition of $\widehat{m}$,

$$\forall m \in \mathcal{M}_n, \quad P_n \gamma(\widehat{s}_{\widehat{m}}) \leq P_n \gamma(\widehat{s}_m) + \text{pen}(m) - \text{pen}(\widehat{m}) \ .$$

For every $m \in \mathcal{M}_n$, we define

$$p_1(m) = P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) \qquad p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$$
$$\delta(m) = (P_n - P)\gamma(s_m) \qquad \overline{\delta}(m) = \delta(m) - (P_n - P)\gamma(s)$$

so that

$$l(s, \widehat{s}_m) = P_n \gamma(\widehat{s}_m) + p_1(m) + p_2(m) - \overline{\delta}(m) - P_n \gamma(s) \ .$$

We then have, for every $m \in \mathcal{M}_n$,

$$l(s, \widehat{s}_{\widehat{m}}) + (\text{pen} - p_1 - p_2 + \overline{\delta})(\widehat{m}) \leq l(s, \widehat{s}_m) + (\text{pen} - p_1 - p_2 + \overline{\delta})(m) \ . \tag{3.2}$$

In order to derive an oracle inequality from (3.2), we have to give lower and upper bounds on $\text{pen} - p_1 - p_2 + \overline{\delta}$ in terms of $l(s, \widehat{s}_m)$. Define the ideal penalty[2] as

$$\text{pen}_{\text{id}}(m) := P\gamma(\widehat{s}_m) - P_n \gamma(\widehat{s}_m) = p_1(m) + p_2(m) - \delta(m) \ .$$

If we could prove that pen is close to pen_{id}, or equivalently[3] that pen is close to

$$\text{pen}'_{\text{id}}(m) := p_1(m) + p_2(m) - \overline{\delta}(m) = \text{pen}_{\text{id}}(m) + (P_n - P)\gamma(s) \ ,$$

then (3.2) would lead to an oracle inequality. Notice also that taking pen too large in (3.2) only enlarges the constant C in the oracle inequality, whereas allowing $\text{pen}(m)$ to be much smaller than $\text{pen}_{\text{id}}(m)$ can make (3.2) trivial.

3.2.3. Histograms. We will often assume that models are made of histograms. This means that each model in $(S_m)_{m \in \mathcal{M}_n}$ is the set of piecewise constant functions (histograms) on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$. It is thus a vector space of dimension $D_m = \text{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda \in \Lambda_m}$. As this basis is orthogonal in $L^2(\mu)$ for any probability measure on $\mathcal{X}$, computations are quite easy. This is the only reason why we make such an assumption in Sect. 3.3. The following notations will be useful throughout this chapter.

$$p_\lambda := P(X \in I_\lambda) \qquad \widehat{p}_\lambda := P_n(X \in I_\lambda)$$
$$(\sigma_\lambda^r)^2 := \mathbb{E}\left[\sigma(X)^2 \,\middle|\, X \in I_\lambda\right] \qquad \left(\sigma_\lambda^d\right)^2 := \mathbb{E}\left[(s(X) - s_m(X))^2 \,\middle|\, X \in I_\lambda\right]$$

[2] *i.e.* the penalty such that $\widehat{m}$ minimizes the criterion $P\gamma(\widehat{s}_m)$, which is the ideal one for prediction.
[3] since pen_{id} and pen'_{id} does not depend on m.

$$s_m := \arg\min_{t\in S_m} P\gamma(t) = \sum_{\lambda\in\Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \qquad \text{with} \quad \beta_\lambda = \mathbb{E}_P\left[Y \mid X \in I_\lambda\right]$$

$$\widehat{s}_m := \arg\min_{t\in S_m} P_n\gamma(t) = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \text{with} \quad \widehat{\beta}_\lambda = \frac{1}{n\widehat{p}_\lambda} \sum_{X_i\in I_\lambda} Y_i$$

Remark that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i. Otherwise, we will consider that the model m can not be chosen. In order to make $\mathbb{E}\left[p_1(m)\right]$ well-defined and finite, we choose a convention for $p_1(m)$ when $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda = 0$ (see (3.13) in Sect. 3.5).

In order to understand the main difference between our framework and the homoscedastic fixed-design, let us compare the expectations of the ideal penalty.

In the homoscedastic fixed-design framework[4], it is quite straightforward to show that

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 D_m}{n} \quad . \tag{3.3}$$

On the other hand, in our framework, we can prove (*cf.* Sect. 5.7.2) the following. Denote by $\mathbb{E}^{\Lambda_m}[\cdot]$ the expectation conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\, \lambda\in\Lambda_m}$. If for every $\lambda\in\Lambda_m$, $\widehat{p}_\lambda > 0$, then

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m} \left(\frac{p_\lambda}{\widehat{p}_\lambda} + 1\right)\left(\left(\sigma_\lambda^d\right)^2 + (\sigma_\lambda^r)^2\right) \quad . \tag{3.4}$$

Apart from the difference between $p_\lambda/\widehat{p}_\lambda$ and 1 (which does not matter with large probability, see Sect. 3.5.4), there are two main differences between (3.3) and (3.4). Firstly, the bias term $\left(\sigma_\lambda^d\right)^2$, which is due to the randomness of the design. If s is highly non-smooth, this term can be significant. Secondly, the variance term $(\sigma_\lambda^r)^2$ depends on $\lambda\in\Lambda_m$, whereas it is constant equal to σ^2 in the homoscedastic case. When $(p_\lambda)_{\lambda\in\Lambda_m}$ are far from the uniform weights, $n^{-1}\sum_{\lambda\in\Lambda_m}(\sigma_\lambda^r)^2$ is far from $D_m n^{-1}\mathbb{E}\left[\sigma(X)^2\right]$. As shown in Chap. 4, in such cases, it may happen that any linear penalization procedure is suboptimal.

3.3. Theoretical results

In this section, we restrict ourselves to the histogram regression case. Remember that we do not consider histograms as a final goal. We only make this assumption in order to make explicit computations and obtain results from which we can derive heuristics for practical applications.

Let $(S_m)_{m\in\mathcal{M}}$ be a family of histogram models such that

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.
(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0\in\mathcal{M}_n$ s.t. $D_{m_0}\in[\sqrt{n}, c_{\mathrm{rich}}\sqrt{n}]$.

For any penalty function $\mathrm{pen}:\mathcal{M}_n\mapsto\mathbb{R}^+$, we define the following model selection procedure:

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n,\, \min_{\lambda\in\Lambda_m}\{\widehat{p}_\lambda\}>0} \left\{P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m)\right\} \quad . \tag{3.5}$$

3.3.1. Optimal penalties. Our first result is an oracle inequality. The following theorem shows that the penalization procedure (3.5) is efficient provided that the penalty is large enough.

THEOREM 3.1. *Assume that the data $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d. and satisfy the following:*

(**Ab**) *Bounded data:* $\|Y_i\|_\infty \leq A < \infty$.
(**An**) *Noise-level bounded from below:* $\sigma(X_i) \geq \sigma_{\min} > 0$ *a.s.*

[4]Notice that the true distribution P gives weights n^{-1} to each of the design points $X_1,\ldots,X_n$. The unknown distribution is only the one of $(\epsilon_i)_{1\leq i\leq n}$.

(**Ap**) *Polynomial decreasing of the bias: there exists* $\beta_1 \geq \beta_2 > 0$ *and* $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ *such that*

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \leq l(s, s_m) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for* $\mathcal{L}(X)$*: there exists* $c_{\mathrm{r},\ell}^X > 0$ *such that for every* $m \in \mathcal{M}_n$, $D_m \min_{\lambda \in \Lambda_m} p_\lambda \geq c_{\mathrm{r},\ell}^X$.

Let $c_1, c_2, C_1, C_2 \geq 0$ *such that* $c_2 > 1$ *and assume that for every* $m \in \mathcal{M}_n$,

$$\begin{aligned} \mathbb{E}\left[c_1 P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) + c_2 P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] &\leq \mathrm{pen}(m) \\ &\leq \mathbb{E}\left[C_1 P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) + C_2 P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \end{aligned} \tag{3.6}$$

with probability at least $1 - Ln^{-2}$.

Then, if $\widehat{m}$ *is defined by* (3.5)*, there exists a constant* K_1 *and a sequence* ϵ_n *converging to zero at infinity such that, with probability at least* $1 - K_1 n^{-2}$,

$$l(s, \widehat{s}_{\widehat{m}}) \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n\right] \inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\} \quad . \tag{3.7}$$

Moreover, we have the oracle inequality

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n\right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\}\right] + \frac{A^2 K_1}{n^2} \quad . \tag{3.8}$$

The constant K_1 *may depend on* c_1, c_2 *and constants in* (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) *and* ($\mathbf{Ar}_\ell^{\mathbf{X}}$)*, but not on* n*. The small term* ϵ_n *depends only on* n *(it can for instance be upperbounded by* $\ln(n)^{-1/5}$*).*

A particular case is $c_1 + c_2 = 2 - \delta_n$ and $C_1 + C_2 = 2 + \delta_n$ for some absolute sequence δ_n converging to zero at infinity. Thm. 3.1 states that if pen is uniformly close to $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$, the model selection procedure defined by (3.5) is asymptotically optimal. Notice that (3.6) can be assumed only for the models of dimension larger than $\ln(n)^\xi$ (with K_1 depending on ξ). In particular, resampling penalties defined in Chap. 6 satisfy such a condition.

The rationale behind this theorem is that if pen is close to $c_1 p_1 + c_2 p_2$, then $\mathrm{crit}(m) = l(s, s_m) + c_1 p_1(m) + (c_2 - 1) p_2(m)$. If $c_1 = c_2 = 1$, this is exactly the ideal criterion $l(s, \widehat{s}_m)$. If $c_1 + c_2 = 2$ with $c_1 \geq 0$ and $c_2 > 1$, we obtain the same result because $p_1(m)$ and $p_2(m)$ are quite close (at least when D_m is large). This closeness between p_1 and p_2 is the keystone of the slope heuristics. Notice that if $\max_{m \in \mathcal{M}_n} D_m \leq K_1'(\ln(n))^{-1} n$ (for some constant K_1' depending only on the assumptions of Thm. 3.1, like K_1), one can replace the condition $c_2 > 1$ by $c_1 + c_2 > 1$ and $c_1, c_2 \geq 0$.

We now make a few comments about the assumptions of Thm. 3.1:

- (**Ab**) and (**An**) are rather mild. In particular, they allow quite general heteroscedastic noises. For results with a noise that can vanish or be unbounded, see Sect. 6.4 and 8.3.
- ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied for "almost regular" histograms when X has a lower bounded density w.r.t. Leb. It is for instance satisfied in the example[5] of Chap. 4, where we find out that linear penalties are suboptimal.
- The upper bound in (**Ap**) holds when $(I_\lambda)_{\lambda \in \Lambda_m}$ is regular and s α-hölderian with $\alpha \in (0, 1]$. The lower bound is more surprising. Indeed, it is classical to assume that $l(s, s_m) > 0$ for every $m \in \mathcal{M}_n$ for proving the asymptotic optimality of Mallows' C_p (*cf.* for instance by Shibata [**Shi81**], Li [**Li87**] and Birgé and Massart [**BM06c**]). We need an explicit lower bound in order to obtain a non-asymptotic lower bound on the dimension of the oracle and selected models.

[5] X in uniform on $\mathcal{X} = [0, 1]$ and S_m contains histograms regular on $[0, 1/2]$ and on $[1/2, 1]$.

The reason why this assumption is not too restrictive is that non-constant α-hölderian functions satisfy (**Ap**) when $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular and X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X}\subset\mathbb{R}^k$ (*cf.* Sect. 8.10 for more details). Notice that Stone [**Sto85**] used the same assumption in the density estimation framework.

3.3.2. Minimal penalties. We now come to the problem of minimal penalties. The following result needs slighlty less assumptions than Thm. 3.1.

THEOREM 3.2. *Assume that the data $(X_i,Y_i)_{1\leq i\leq n}$ are i.i.d. and satisfy the following:*

(**Ab**) *Bounded data: $\|Y_i\|_\infty\leq A<\infty$.*

(**An**) *Noise-level bounded from below: $\sigma(X_i)\geq\sigma_{\min}>0$ a.s.*

($\mathbf{Ap_u}$) *Polynomial upper bound on the bias: there exists $\beta_2>0$ and $C_{\mathrm{b}}^+>0$ such that*

$$l(s,s_m)\leq C_{\mathrm{b}}^+ D_m^{-\beta_2}\quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for $\mathcal{L}(X)$: $D_m\min_{\lambda\in\Lambda_m}p_\lambda\geq c_{\mathrm{r},\ell}^X$.*

Let $C_2\in[0;1)$ and assume that

$$0\leq\mathrm{pen}(m)\leq\mathbb{E}\left[C_2P_n\left(\gamma(s_m)-\gamma(\widehat{s}_m)\right)\right] \tag{3.9}$$

with probability at least $1-Ln^{-2}$.

Then, if $\widehat{m}$ is defined by (3.5), there exists two constants K_2, K_3 such that, with probability at least $1-K_2n^{-2}$,

$$D_{\widehat{m}}\geq K_3n\ln(n)^{-1}\quad . \tag{3.10}$$

On the same event,

$$l(s,\widehat{s}_{\widehat{m}})\geq\ln(n)\inf_{m\in\mathcal{M}_n}\left\{l(s,\widehat{s}_m)\right\}\quad . \tag{3.11}$$

*The constant K_2 and K_3 may depend on C_1, C_2 and constants in (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) and ($\mathbf{Ar}_\ell^{\mathbf{X}}$), but not on n.*

As in the results of Birgé and Massart [**BM06c**], Thm. 3.2 points out two simultaneous phenomena when the penalty is too small. First, the dimension of the selected model explodes (3.10). Secondly, the efficiency of the model selection strongly decreases (3.11). This coupling is quite interesting. Indeed, we want to avoid underpenalization because of the second phenomenon, while the blow up of the dimension allows us to detect it more easily.

The minimal penalty pointed out by Thm. 3.2 is $p_2(m)=P_n\left(\gamma(s_m)-\gamma(\widehat{s}_m)\right)$. This is quite intuitive since $\mathrm{crit}(m)=l(s,s_m)+\mathrm{pen}(m)-p_2(m)$, so that $\mathrm{pen}\leq p_2$ make $\mathrm{crit}(m)$ decreases with D_m.

3.3.3. Comments. The comparison between Thm. 3.1 and 3.2 has two main consequences.

First, the minimal penalty is

$$\mathrm{pen}_{\min}(m):=\mathbb{E}\left[p_2(m)\right]=\mathbb{E}\left[P_n\left(\gamma(s_m)-\gamma(\widehat{s}_m)\right)\right]$$

whereas $\mathrm{pen}(m)=2\,\mathrm{pen}_{\min}(m)$ satisfies a non-asymptotic oracle inequality with constant almost one (take $c_1=C_1=0$ and $c_2=C_2=2$ in Thm. 3.1). In particular, it is asymptotically optimal.

Second, without assuming any lower bound on the bias in (**Ap**), there is a blow up phenomenon for the selected dimension $D_{\widehat{m}}$. Indeed, when $\mathrm{pen}(m)$ is too small (Thm 3.2), $D_{\widehat{m}}\geq K_3n\ln(n)^{-1}$ with probability $1-K_2n^{-2}$. On the other hand, when $\mathrm{pen}(m)>\mathrm{pen}_{\min}$, the proof of Thm. 3.1 (equations (3.16) and (3.17)) shows that for every $\alpha>(1-\beta_2)_+/2$, there is a set of probability at least $1-K_1'(\alpha)n^{-2}$ on which $D_{\widehat{m}}\leq n^{1/2+\alpha}$. In other words, this means that the selected dimension $D_{\widehat{m}}$ has quite different values when $\mathrm{pen}<\mathrm{pen}_{\min}$ and when $\mathrm{pen}>\mathrm{pen}_{\min}$. This

dimension jump is a key phenomenon, which can be used in practice for determining the minimal penalty. In simulation studies as on real data sets, we clearly observe such a dimension jump. Remark that we do not assume in this paragraph the lower bound in (**Ap**), so that the dimension jump also occurs when s belongs to one of the models.

These two points may have great applications for the practical users of penalization criteria. This is the object of the next section.

3.4. Practical use of slope heuristics: data-driven penalties

Following Sect. 4 of Birgé and Massart [**BM06c**], we can combine the asymptotic optimality of $2\,\mathrm{pen}_{\min}$ and the dimension jump around $\mathrm{pen} = \mathrm{pen}_{\min}$ in order to build data-driven penalties.

ALGORITHM 3.1 (Data-driven penalization with slope heuristics).

(1) Choose a shape of penalty $\mathrm{pen}_{\mathrm{shape}} : \mathcal{M}_n \mapsto \mathbb{R}^+$.
(2) Compute the selected model $\widehat{m}(K)$ as a function of $K > 0$

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m) + K\,\mathrm{pen}_{\mathrm{shape}}(m) \right\} \quad .$$

(3) Find $\widehat{K}_{\min} > 0$ such that $D_{\widehat{m}(K)}$ is too large for $K < \widehat{K}_{\min}$ and "reasonably small" for $K > \widehat{K}_{\min}$.
(4) Select the model $\widehat{m} = \widehat{m}(2\widehat{K}_{\min})$.

3.4.1. Computation of $\widehat{K}_{\min}$. In the above procedure, step 2 can be made with a complexity smaller than $\mathrm{Card}(\mathcal{M}_n)^2$, with the following algorithm (notice that Algorithm 3.2 can be stopped earlier if the only goal is to identify $\widehat{K}_{\min}$.

ALGORITHM 3.2 (Step 2 of Algorithm 3.1). For every $m \in \mathcal{M}_n$, define $f(m) = P_n\gamma\left(\widehat{s}_m\right)$ and $g(m) = \mathrm{pen}_{\mathrm{shape}}(m)$. Choose $\preceq$ any total ordering on $\mathcal{M}_n$ such that g is non-decreasing.

- Init: $K_0 = 0$, $m_0 = \arg\min_{m \in \mathcal{M}_n} \{ f(m) \}$ (when this minimum is attained several times, m_0 is defined as the smallest one for $\preceq$).
- Step i, $i \geq 1$: Let

$$G(m_{i-1}) := \{ m \in \mathcal{M}_n \text{ s.t. } f(m) > f(m_{i-1}) \quad \text{and} \quad g(m) < g(m_{i-1}) \} \quad .$$

If $G(m_{i-1}) = \emptyset$, then put $K_i = +\infty$, $i_{\max} = i$ and stop.
Otherwise, define

$$K_i := \inf \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \text{ s.t. } m \in G(m_{i-1}) \right\} \tag{3.12}$$

and m_i the smallest element (for $\preceq$) of

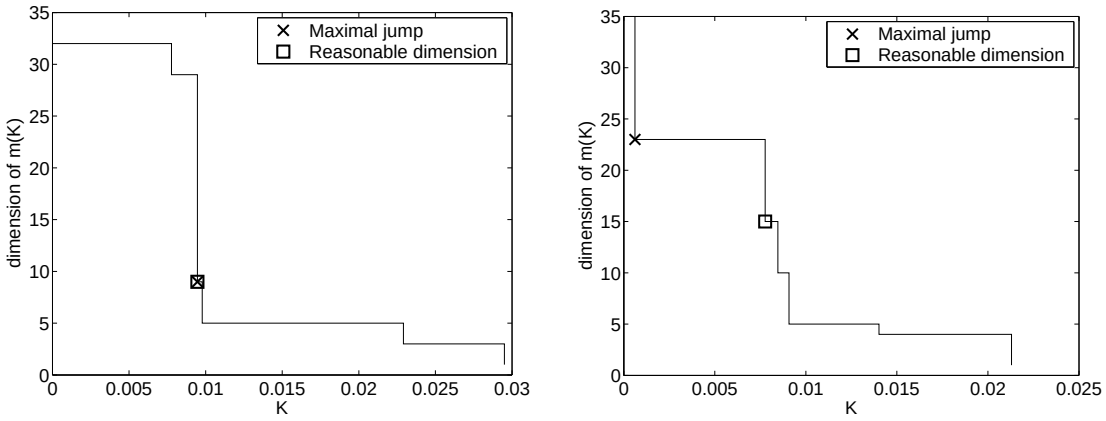
$$F_i := \arg\min_{m \in G(m_{i-1})} \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \right\} \quad .$$

PROPOSITION 3.1. *If $\mathcal{M}_n$ is finite, algorithm 3.2 terminates and $i_{\max} \leq \mathrm{Card}(\mathcal{M}_n)$. Using the notations of Algorithm 3.2, and defining $\widehat{m}(K)$ as the smallest element (for $\preceq$) of*

$$E(K) := \{ f(m) + Kg(m) \text{ s.t. } m \in \mathcal{M}_n \} \quad ,$$

$(K_i)_{0 \leq i \leq i_{\max}}$ is increasing and $\forall i \in \{0, \ldots, i_{\max} - 1\}$, $\forall K \in [K_i, K_{i+1})$, $\widehat{m}(K) = m_i$.

3.4.2. Definition of $\widehat{K}_{\min}$. In the two algorithms above, the exact definition of $\widehat{K}_{\min}$ is not completely clear. Actually, we do not know exactly which one should be the best one in practice. According to some preliminary experiments, the dimension jump is often quite clear,



(a) One clear jump. (b) Two jumps, two values for $\widehat{K}_{\min}$.

FIGURE 3.1. $D_{\widehat{m}(K)}$ as a function of K for two different samples. Data are simulated from experiment (S1). "Reasonable dimensions" are below $n/(2\ln(n)) \approx 19$.

with $D_{\widehat{m}(K)}$ quite large for $K > \widehat{K}_{\min}$ and much smaller when $K < \widehat{K}_{\min}$. Fig. 3.1a gives an example of such a situation. In that case, any definition ("large jump", or "reasonable dimension") gives the same value for $\widehat{K}_{\min}$. However, it sometimes happens that there is no clear jump, or two comparable jumps with the larger one between two models of large dimension (15% of the samples in experiment S1). See for instance Fig. 3.1b. In those cases, the choice of the definition can influence the selected model $\widehat{m}(2\widehat{K}_{\min})$ (6.5% of the samples in experiment S1).

Moreover, if we finally choose the "reasonable dimension" definition, we should precise what is a "reasonable" model. On Fig. 3.1b, we choose that it is a model of dimension lower than $n/(2\ln(n))$. Considering Thm. 3.2, in the histogram case, a model is reasonable when it has a dimension $\ll n\ln(n)^{-1}$. This only means that we have more than the minimal amount of data to compute the empirical risk minimizer $\widehat{s}_m$.

In order to make easier the choice of $\widehat{K}_{\min}$, we can suggest the following modification of $\mathcal{M}_n$, at least in the histogram case. First, remove all the huge models (*i.e.* of dimension larger than, say $n/\ln(n)$) from $\mathcal{M}_n$. This is often done in practice, because those models are obviously wrong and they only enlarge the computation time. Then, add a few models of dimension $\approx n/2$, so that at least one[6] has a well-defined $\widehat{s}_m$. For instance, in the histogram case, reorder the data $X_{(1)} < \cdots < X_{(n)}$ and consider the model of histograms adapted to the partition $(I_\lambda)_{\lambda\in\Lambda_m}$ with endpoints $X_{(2)}, X_{(4)}, \ldots, X_n$. Finally, let $\widehat{K}_{\min}$ be the minimal value of K for which none of these huge models is selected.

We do not exactly know whether this method really solves the problem of choosing $\widehat{K}_{\min}$, and if we should prefer the "large jump" or the "reasonable dimension" definition. Further experimental investigations should be done in order to suggest an answer to this problem.

3.4.3. Shape of the penalty. Choosing a shape $\mathrm{pen}_{\mathrm{shape}}$ for the penalty is not a simple answer. In the homoscedastic regression on a fixed-design framework, $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$ works provided that $\mathrm{Card}(\mathcal{M}_n)$ is polynomial in n. In our framework, $\mathrm{pen}_{\min}$ may have quite general shapes, as shown for instance in Chap. 4. While this point is a main advance compared to the results of Birgé and Massart [**BM06c**], we need to find a practical answer to this new question.

[6] several huge models may be necessary in practice, in order to decrease the variability of $\widehat{K}_{\min}$.

Following a remark made after Thm. 3.1, $\text{pen}_{\text{shape}}$ can be taken among the Resampling Penalties defined by algorithm 6.2 (in Sect. 6.3.1) and the V-fold penalties defined by algorithm 5.2 (in Sect. 5.3.1). Of course, in the histogram framework, we already know the optimal constant $2K_{\min}$, so that the use of algorithm 3.1 may not be necessary. However, in a general framework, the right non-asymptotic optimal constant may differ from the asymptotic one (or the non-asymptotic one derived from the histogram case). Although we do not have theoretical evidence, we conjecture that our results still holds in general "reasonable" frameworks.

3.4.4. Large number of models. Finally, notice that we assume the collection of models to be small, *i.e.* it size is polynomial in n. On that point, we differ from Birgé and Massart [**BM06c**] who consider several other cases. In particular, they show how the minimal penalty has to be enlarged when $\text{Card}(\mathcal{M}_n)$ is exponential in n.

Following (42) and the surrounding comments in [**BM06c**], we suggest to group the models according to some complexity index C_m (for instance their dimensions): for $C \in \left\{1, \dots, n^k\right\}$, define $\widetilde{S_C} = \bigcup_{C_m = C} S_m$. Then, the model selection with the family $(S_m)_{m \in \mathcal{M}_n}$ should be similar to the selection of a complexity C, with the family of models $\left(\widetilde{S_C}\right)_{1 \leq C \leq n^k}$. If we had a result for models of the form $\widetilde{S_C}$, the polynomial complexity case could then be applied. Notice that the S_m all are histogram models, $\widetilde{S_C}$ is not necessarily the model of histograms adapted to some partition of $\mathcal{X}$. We conjecture that such a grouping of the models allows the use of slope heuristics for tuning a penalization procedure.

3.5. Proofs

In the following, when we do not want to write explicitly some constants, we use the letter L. It means "some absolute constant, possibly different from a line to another, or even within the same line". When L is not numerical, but depends on some parameters $p_1, \dots, p_k$, it is written $L_{p_1,\dots,p_k}$ of $L(p_1, \dots, p_k)$. $L_{(\mathbf{SH1})}$ (resp. $L_{(\mathbf{SH2})}$) denotes a constant that depends only on the set of assumptions of Thm. 3.1 (resp. Thm. 3.2), including (**P1**) and (**P2**).

Moreover, since $\mathbb{E}[p_1]$ is not well-defined (because of the event $\{\min_{\lambda \in \Lambda_m} \{\widehat{p}_\lambda\} = 0\}$), we have to take the following convention:

$$p_1(m) = \widetilde{p_1}(m) := \sum_{\lambda \in \Lambda_m \text{ s.t. } \widehat{p}_\lambda > 0} p_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2) + \sum_{\lambda \in \Lambda_m \text{ s.t. } \widehat{p}_\lambda = 0} p_\lambda \left((\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right) \quad . \tag{3.13}$$

Remark that $p_1(m) = \widetilde{p_1}(m)$ when $\min_{\lambda \in \Lambda_m} \{\widehat{p}_\lambda\} > 0$), so that this convention has no consequences on the final results (Thm. 3.1 and 3.2).

3.5.1. Proof of Thm. 3.1. This proof is very similar to the one of Thm. 5.1 stated in Sect. 5.3.3. We give it for the sake of completeness.

From (3.2), we have for each $m \in \mathcal{M}_n$ such that $A_n(m) := \min_{\lambda \in \Lambda_m} \{n\widehat{p}_\lambda\} > 0$

$$l(s, \widehat{s}_{\widehat{m}}) - \left(\text{pen}'_{\text{id}}(\widehat{m}) - \text{pen}(\widehat{m})\right) \leq l(s, \widehat{s}_m) + \left(\text{pen}(m) - \text{pen}'_{\text{id}}(m)\right) \quad . \tag{3.14}$$

with $\text{pen}'_{\text{id}}(m) = p_1(m) + p_2(m) - \overline{\delta}(m) = \text{pen}(m) + (P - P_n)\gamma(s)$. It is sufficient to control $\text{pen} - \text{pen}'_{\text{id}}$ for every $m \in \mathcal{M}_n$.

We will thus use the concentration inequalities of Sect. 3.5.4 with $x = \gamma \ln(n)$ and $\gamma = 2 + \alpha_{\mathcal{M}}$. Define $B_n(m) = \min_{\lambda \in \Lambda_m} \{np_\lambda\}$. Let Ω_n be the event on which

- for every $m \in \mathcal{M}_n$, (3.6) holds

- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$:

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{(\mathbf{SH1})}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n(m)}\right]\mathbb{E}\left[p_2(m)\right] \tag{3.30}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{(\mathbf{SH1})}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n(m)}\right]\mathbb{E}\left[p_2(m)\right] \tag{3.31}$$

- for every $m \in \mathcal{M}_n$ such that $B_n(m) > 0$:

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n(m)^{-1}\ln(n)} - \frac{L_{(\mathbf{SH1})}\ln(n)^2}{\sqrt{D_m}}\right)\mathbb{E}\left[p_2(m)\right] \ . \tag{3.32}$$

$$\left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right| \leq \frac{L_{(\mathbf{SH1})}\ln(n)}{\sqrt{D_m}}\left[l(s,s_m) + \mathbb{E}\left[p_2(m)\right]\right] \tag{3.29}$$

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s,s_m)}{\sqrt{D_m}} + L_{(\mathbf{SH1})}\frac{\ln(n)}{\sqrt{D_m}}\mathbb{E}\left[p_2(m)\right] \tag{3.27}$$

From Prop. 3.5 (for $\widetilde{p_1}$), Prop. 3.4 (for p_2), Prop. 3.3 (for $\overline{\delta}(m)$), we have

$$\mathbb{P}\left(\Omega_n\right) \geq 1 - L\sum_{m\in\mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L(c_{\mathcal{M}})n^{-2} \ .$$

For every $m \in \mathcal{M}_n$ such that $D_m \leq L_{c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}$, $(\mathbf{Ar}_\ell^{\mathbf{X}})$ implies that $B_n(m) \geq L^{-1}\ln(n) \geq 1$. As a consequence, on Ω_n, if $\ln(n)^7 \leq D_m \leq L_{c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}$:

$$\begin{aligned}\max\left\{\left|\widetilde{p_1}(m) - \mathbb{E}\left[\widetilde{p_1}(m)\right]\right|, \left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right|, \left|\overline{\delta}(m)\right|\right\} \\ \leq \frac{L_{(\mathbf{SH1})}\mathbb{E}\left[l(s,s_m) + p_2(m)\right]}{\ln(n)}\end{aligned}$$

Using (3.33) (in Prop. 3.6) and the fact that $B_n(m) \geq L^{-1}\ln(n)$,

$$\frac{(c_1+c_2)\left(1-\widetilde{\delta_n}\right)}{2} \leq \mathbb{E}\left[\mathrm{pen}(m)\right] \leq \frac{(C_1+C_2)\left(1+\widetilde{\delta_n}\right)}{2}\mathbb{E}\left[\widetilde{p_1}(m) + p_2(m)\right]$$

with $0 \leq \widetilde{\delta_n} \leq L\ln(n)^{-1/4}$. We deduce: if $n \geq L_{(\mathbf{SH1})}$, for every $m \in \mathcal{M}_n$ such that $\ln(n)^7 \leq D_m \leq L_{c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}$, on Ω_n,

$$\begin{aligned}\left[\left(c_1+c_2-2\right)_- - \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right]p_1(m) \leq \left(\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}\right)(m) \\ \leq \left[\left(C_1+C_2-2\right)_+ + \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right]p_1(m) \ .\end{aligned}$$

We need to assume that n is large enough in order to upper bound $\mathbb{E}\left[p_2(m)\right]$ in terms of $p_1(m)$, since we only have

$$p_1(m) \geq \left[1 - \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right]_+ \mathbb{E}\left[p_2(m)\right]$$

in general.

Combined with (3.14), this gives: if $n \geq L_{(\mathbf{SH1})}$,

$$\begin{aligned}l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\ln(n)^5 \leq D_{\widehat{m}} \leq L_{c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}} \leq \left[\frac{1 + \left(C_1+C_2-2\right)_+}{\left(c_1+c_2-1\right)\wedge 1} + \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right] \\ \times \inf_{m\in\mathcal{M}_n \text{ s.t. } \ln(n)^7 \leq D_m \leq L_{\alpha_{\mathcal{M}}, c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}} \left\{l(s,\widehat{s}_m)\right\} \ .\end{aligned} \tag{3.15}$$

Define the oracle model $m^\star \in \arg\min\{\, l(s,\widehat{s}_m)\,\}$. We prove below that for any $c > 0$ and $\alpha > (1-\beta_2)_+/2$, if $n \geq L_{(\mathbf{SH1}),c,\alpha}$, then, on Ω_n:

$$\ln(n)^7 \leq D_{\widehat{m}} \leq n^{1/2+\alpha} \leq cn\ln(n)^{-1} \tag{3.16}$$

$$\ln(n)^7 \leq D_{m^\star} \leq n^{1/2+\alpha} \leq cn\ln(n)^{-1} \ . \tag{3.17}$$

The result follows since $L_{(\mathbf{SH1})}\ln(n)^{-1/4} \leq \epsilon_n = \ln(n)^{-1/5}$ for $n \geq L_{(\mathbf{SH1})}$. We finally remove the condition $n \geq n_0 = L_{(\mathbf{SH1})}$ by choosing $K_1 = L_{(\mathbf{SH1})}$ such that $K_1 n_0^{-2} \geq 1$.

Proof of (3.16). By definition, $\widehat{m}$ minimizes $\mathrm{crit}(m)$ over $\mathcal{M}_n$. It thus also minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = l(s,s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $\mathcal{M}_n$.

(1) Lower bound on $\mathrm{crit}'(m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^7$. We then have

$$l(s,s_m) \geq C_{\mathrm{b}}^- (\ln(n))^{-7\beta_1} \qquad \text{from } (\mathbf{Ap})$$
$$\mathrm{pen}(m) \geq 0$$
$$p_2(m) \leq L_{(\mathbf{SH1})}\sqrt{\frac{\ln(n)}{n}} + L_{(\mathbf{SH1})}\frac{D_m}{n} \leq L_{(\mathbf{SH1})}\sqrt{\frac{\ln(n)}{n}} \qquad \text{from (3.28)}$$

and from (3.27) (in Prop. 3.3),

$$\overline{\delta}(m) \geq -L_A\sqrt{\frac{l(s,s_m)\ln(n)}{n}} + L_A\frac{\ln(n)}{n} \geq -L_A\sqrt{\frac{\ln(n)}{n}} \ .$$

We then have

$$\mathrm{crit}'(m) \geq L_{(\mathbf{SH1})}(\ln(n))^{-L(\beta_1)} \ .$$

(2) Lower bound for large models: let $m \in \mathcal{M}_n$ such that $D_m \geq n^{1/2+\alpha}$. From (3.6) and (3.28) (in Prop. 3.4),

$$\begin{aligned}\mathrm{pen}(m) - p_2(m) &\geq (c_2-1)\,\mathbb{E}\,[p_2(m)] - L_A\sqrt{\frac{\ln(n)}{n}} \\ &\geq \frac{(c_2-1)\sigma_{\min}^2 D_m}{n} - L_A\sqrt{\frac{\ln(n)}{n}}\end{aligned}$$

and from (3.25),

$$\overline{\delta}(m) \geq -L_{(\mathbf{SH1})}\sqrt{\frac{\ln(n)}{n}} \ .$$

Hence, if $D_m \geq n^{1/2+\alpha}$ and $n \geq L_{(\mathbf{SH1}),\alpha}$

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) + \overline{\delta}(m) - p_2(m) \geq L_{(\mathbf{SH1}),\alpha}n^{-1/2+\alpha} \ .$$

(3) There exists a better model for $\mathrm{crit}(m)$: from $(\mathbf{P2})$, there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If moreover $n \geq L_{c_{\mathrm{rich}},\alpha}$, then

$$\ln(n)^7 \leq \sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n} \leq n^{1/2+\alpha} \ .$$

By (3.34) in Lemma 3.7, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.
Using $(\mathbf{Ap})$,

$$l(s,s_{m_0}) \leq C_{\mathrm{b}}^+ c_{\mathrm{rich}}^{\beta_2} n^{-\beta_2/2}$$

so that, when $n \geq L_{(\mathbf{SH1})}$,

$$\begin{aligned}\mathrm{crit}'(m_0) &\leq l(s, s_{m_0}) + \left|\overline{\delta}(m)\right| + \mathrm{pen}(m) \\ &\leq L_{(\mathbf{SH1})} \left(n^{-\beta_2/2} + n^{-1/2} \right) \quad .\end{aligned}$$

If $n \geq L_{(\mathbf{SH1}),\alpha}$, this upper bound is smaller than the previous lower bounds for small and large models.

Proof of (3.17). Recall that $m^\star$ minimizes $l(s, \widehat{s}_m) = l(s, s_m) + p_1(m)$ over $m \in \mathcal{M}_n$, with the convention $l(s, \widehat{s}_m) = \infty$ if $A_n(m) = 0$.

(1) Lower bound on $l(s, \widehat{s}_m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^7$. From (**Ap**), we have

$$l(s, \widehat{s}_m) \geq l(s, s_m) \geq C_{\mathrm{b}}^- \left(\ln(n)\right)^{-7\beta_1} \quad .$$

(2) Lower bound on $l(s, \widehat{s}_m)$ for large models: let $m \in \mathcal{M}_n$ such that $D_m > n^{1/2+\alpha}$. From (3.32), for $n \geq L_{(\mathbf{SH1}),\alpha}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)\left(c_{\mathrm{r},\ell}^X\right)^{-1} \ln(n)} - \frac{L_{(\mathbf{SH1}),\alpha}}{n^{1/4}} \right) \mathbb{E}\left[\widetilde{p_2}(m)\right]$$

$$\text{so that} \quad l(s, \widehat{s}_m) \geq \widetilde{p_1}(m) \geq L_{(\mathbf{SH1}),\alpha} n^{-1/2+\alpha} \quad .$$

(3) There exists a better model for $l(s, \widehat{s}_m)$: let $m_0 \in \mathcal{M}_n$ be as in the proof of (3.16) and assume that $n \geq L_{c_{\mathrm{rich}},\alpha}$. Then,

$$p_1(m_0) \leq L_{(\mathbf{SH1})} \mathbb{E}\left[p_2(m)\right] \leq L_{(\mathbf{SH1})} n^{-1/2}$$

and the arguments of the previous proof show that

$$l(s, \widehat{s}_{m_0}) \leq L_{(\mathbf{SH1})} \left(n^{-\beta_2/2} + n^{-1/2} \right)$$

which is smaller than the previous upper bounds for $n \geq L_{(\mathbf{SH1}),\alpha}$.

Classical oracle inequality. Let Ω_n be the event on which (3.7) holds true. Then,

$$\begin{aligned}\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] &= \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\Omega_n}\right] + \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\Omega_n^c}\right] \\ &\leq \left[2\eta - 1 + \epsilon_n\right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\}\right] + A^2 K_1 \mathbb{P}\left(\Omega_n^c\right)\end{aligned}$$

which proves (3.8).

3.5.2. Proof of Thm. 3.2. Similarly to the proof of Thm. 3.1, we consider the event Ω_n', of probability at least $1 - L(c_{\mathcal{M}}) n^{-2}$, on which:

- for every $m \in \mathcal{M}_n$, (3.9) (for pen), (3.32) (for $\widetilde{p_1}$), (3.28)–(3.29) (for p_2, with $x = \gamma \ln(n)$ and $\theta = \sqrt{\ln(n)/n}$) and (3.25)–(3.27) (for $\overline{\delta}$, with $x = \gamma \ln(n)$ and $\eta = \sqrt{\ln(n)/n}$) hold true.
- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$, (3.30) and (3.31) hold (for $\widetilde{p_1}$).

Lower bound on $D_{\widehat{m}}$. By definition, $\widehat{m}$ minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = l(s, s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$. As in the proof of Thm. 3.1, we define $c = L_{c_{\mathrm{r},\ell}^X} > 0$ such that for every model of dimension $D_m \leq cn \ln(n)^{-1}$, $B_n(m) \geq L^{-1} \ln(n) \geq 1$. Let $d < 1$ to be chosen later.

(1) Lower bound on $\mathrm{crit}'(m)$ for "small" models: assume that $m \in \mathcal{M}_n$ and $D_m \leq dcn\ln(n)^{-1}$. Then, $l(s,s_m)+\mathrm{pen}(m) \geq 0$ and from (3.25),

$$\overline{\delta}(m) \geq -L_A\sqrt{\frac{\ln(n)}{n}} \ .$$

If $D_m \geq \ln(n)^4$, (3.29) implies that

$$p_2(m) \leq \left(1+\frac{L_{(\mathbf{SH2})}}{\ln(n)}\right)\mathbb{E}\left[p_2(m)\right] \leq \frac{L_{(\mathbf{SH2})}D_m}{n} \leq \frac{cdL_{(\mathbf{SH2})}}{\ln(n)} \ .$$

On the other hand, if $D_m < \ln(n)^4$, (3.28) implies that

$$p_2(m) \leq L_{(\mathbf{SH2})}\sqrt{\frac{\ln(n)}{n}} \ .$$

We then have

$$\mathrm{crit}'(m) \geq -dL_{(\mathbf{SH2})}\left(\ln(n)\right)^{-1} \ .$$

(2) There exists a better model for $\mathrm{crit}(m)$: let $m_1 \in \mathcal{M}_n$ such that

$$\ln(n)^4 \leq \frac{cdn}{c_{\mathrm{rich}}\ln(n)} \leq D_{m_1} \leq \frac{cn}{\ln(n)} \leq n \ .$$

From (**P2**), this is possible as soon as $n \geq L_{c_{\mathrm{rich}},c,d}$. By (3.34) in Lemma 3.7, $A_n(m_0) \geq 1$ with probability at least $1-Ln^{-2}$.
We then have

$$\begin{aligned}
l(s,s_{m_1}) &\leq L_{(\mathbf{SH2}),c}\ln(n)^{\beta_2}n^{-\beta_2} && \text{by } (\mathbf{Ap})\\
p_2(m_1) &\leq \left(1+\frac{L_{(\mathbf{SH2})}}{\ln(n)}\right)\mathbb{E}\left[p_2(m_1)\right] && \text{by (3.29)}\\
\mathrm{pen}(m_1) &\leq C_2\mathbb{E}\left[p_2(m_1)\right] && \text{by (3.9)}\\
\left|\overline{\delta}(m_1)\right| &\leq L_A\sqrt{\frac{\ln(n)}{n}} && \text{by (3.25)}
\end{aligned}$$

so that

$$\begin{aligned}
\mathrm{crit}'(m_1) &\leq L_{(\mathbf{SH2}),c}\ln(n)^{\beta_2}n^{-\beta_2} + \left(C_2-1-\frac{L_{(\mathbf{SH2})}}{\ln(n)}\right)\mathbb{E}\left[p_2(m_1)\right] + L_A\sqrt{\frac{\ln(n)}{n}}\\
&\leq \frac{(C_2-1)\sigma_{\min}^2 c}{2\ln(n)}
\end{aligned}$$

if $n \geq L_{(\mathbf{SH2}),c}$.

We now choose d such that the constant $dL_{(\mathbf{SH2})}$ appearing in the lower bound on $\mathrm{crit}'(m)$ for "small" models is smaller than $(1-C_2)\sigma_{\min}^2 c/2$, *i.e.* $d \leq L_{(\mathbf{SH2}),c}$. Then, we assume that $n \geq n_0 = L_{(\mathbf{SH2}),c,d} = L_{(\mathbf{SH2})}$. Finally, we remove this condition as before by enlarging K_2.

Risk of $D_{\widehat{m}}$. The proof of (3.11) is quite similar to the one of (3.17). First, for every model $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$ and $D_m \geq K_3 n\ln(n)^{-1}$, we have

$$l(s,\widehat{s}_m) \geq \widetilde{p}_1(m) \geq L_{(\mathbf{SH2})}K_3\ln(n)^{-2} \qquad \text{by (3.32)} \ .$$

Then, the model $m_0 \in \mathcal{M}_n$ defined previously satisfies $A_n(m) \geq 1$, and

$$l(s,\widehat{s}_{m_0}) \leq L_{(\mathbf{SH2})}\left(n^{-\beta_2/2}+n^{-1/2}\right) \ .$$

If $n \geq L_{(\mathbf{SH2})}$, the ratio between these two bounds is larger than $\ln(n)$, so that (3.11) holds.

3.5.3. Proof of Prop. 3.1. First of all, by construction, $g(m_i)$ decreases with i, so that all the $m_i \in \mathcal{M}_n$ are distinct. Hence, algorithm (3.2) terminates and $i_{\max} \leq \mathrm{Card}(\mathcal{M}_n)$.

We now prove by induction the following property for every $i \in \{0, \dots, i_{\max} - 1\}$:

$$\mathcal{P}_i : \qquad K_i < K_{i+1} \quad \text{and} \quad \forall K \in [K_i, K_{i+1}), \quad \widehat{m}(K) = m_i \ .$$

Notice also that K_i can always be defined by (3.12) with the convention $\inf \emptyset = +\infty$.

$\mathcal{P}_0$ holds true. By definition of K_1, it is clear that $K_1 > 0$ (it may be equal to $+\infty$ if $G(m_0) = \emptyset$). When $K = K_0 = 0$, the definition of m_0 is the one of $\widehat{m}(0)$, so that $\widehat{m}(K) = m_0$. When $K \in (0, K_1)$, then Lemma 3.2 shows that either $\widehat{m}(K) = \widehat{m}(0)$ or $\widehat{m}(K) \in G(0)$. In the latter case, by definition of K_1,

$$\frac{f(\widehat{m}(K)) - f(m_0)}{g(m_0) - g(\widehat{m}(K))} \geq K_1 > K$$

so that

$$f(\widehat{m}(K)) + Kg(\widehat{m}(K)) > f(m_0) + Kg(m_0)$$

which is contradictory with the definition of $\widehat{m}(K)$.

$\mathcal{P}_i \Rightarrow \mathcal{P}_{i+1}$ for every $i \in \{0, \dots, i_{\max} - 2\}$. Assume that $\mathcal{P}_i$ holds true. First, we have to prove that $K_{i+2} > K_{i+1}$. Since $K_{i_{\max}} = +\infty$, this is clear if $i = i_{\max} - 2$. Otherwise, $K_{i+2} < +\infty$ and m_{i+2} exists. Then, by definition of m_{i+2} and K_{i+2} (resp. m_{i+1} and K_{i+1}), we have

$$f(m_{i+2}) - f(m_{i+1}) = K_{i+2}(g(m_{i+1}) - g(m_{i+2})) \tag{3.18}$$

$$f(m_{i+1}) - f(m_i) = K_{i+1}(g(m_i) - g(m_{i+1})) \ . \tag{3.19}$$

Moreover, $m_{i+2} \in G(m_{i+1}) \subset G(m_i)$, and $m_{i+2} \prec m_{i+1}$ (because g is non-decreasing). Using again the definition of K_{i+1}, we have

$$f(m_{i+2}) - f(m_i) > K_{i+1}(g(m_i) - g(m_{i+2})) \tag{3.20}$$

(otherwise, we would have $m_{i+2} \in F_{i+1}$ and $m_{i+2} \prec m_{i+1}$, which is not possible). Combining the difference of (3.20) and (3.19) with (3.18), we have

$$K_{i+2}(g(m_{i+1}) - g(m_{i+2})) > K_{i+1}(g(m_{i+1}) - g(m_{i+2})) \ ,$$

so that $K_{i+2} > K_{i+1}$ (since $g(m_{i+1}) > g(m_{i+2})$).

Second, we prove that $\widehat{m}(K_{i+1}) = m_{i+1}$. From $\mathcal{P}_i$, we know that for every $m \in \mathcal{M}_n$, for every $K \in [K_i, K_{i+1})$, $f(m_i) + Kg(m_i) \leq f(m) + Kg(m)$. Taking the limit when K goes to K_{i+1}, we obtain that $m_i \in E(K_{i+1})$. By (3.19), we then have $m_{i+1} \in E(K_{i+1})$. On the other hand, if $m \in E(K_{i+1})$, Lemma 3.2 shows that either $f(m) = f(m_i)$ and $g(m) = g(m_i)$ or $m \in G(m_i)$. In the first case, $m_{i+1} \prec m$ (because g is non-decreasing). In the second one, $m \in F_{i+1}$, so $m_{i+1} \preceq m$. Since $\widehat{m}(K_{i+1})$ is the smallest element of $E(K_{i+1})$, we have proven that $m_{i+1} = \widehat{m}(K_{i+1})$.

Last, we have to prove that $\widehat{m}(K) = m_{i+1}$ for every $K \in (K_1, K_2)$. From the last statement of Lemma 3.2, we have either $\widehat{m}(K) = \widehat{m}(K_1)$ or $\widehat{m}(K_1) \in G(\widehat{m}(K))$. In the latter case (which is only possible if $K_{i+2} < \infty$), by definition of K_{i+2},

$$\frac{f(\widehat{m}(K)) - f(m_{i+1})}{g(m_{i+1}) - g(\widehat{m}(K))} \geq K_{i+2} > K$$

so that

$$f(\widehat{m}(K)) + Kg(\widehat{m}(K)) > f(m_{i+1}) + Kg(m_{i+1})$$

which is contradictory with the definition of $\widehat{m}(K)$. □

LEMMA 3.2. *Use the notations of Prop. 3.1 and its proof. If $0 \leq K < K'$, $m \in E(K)$ and $m' \in E(K')$, then we have either*

(a) $f(m) = f(m')$ *and* $g(m) = g(m')$.
(b) $f(m) < f(m')$ *and* $g(m) > g(m')$.

In particular, we have either $\widehat{m}(K) = \widehat{m}(K')$ *or* $\widehat{m}(K') \in G(\widehat{m}(K))$.

PROOF OF LEMMA 3.2. By definition of $E(K)$ and $E(K')$, we have

$$f(m) + Kg(m) \leq f(m') + Kg(m') \tag{3.21}$$

$$f(m') + K'g(m') \leq f(m) + K'g(m) \ . \tag{3.22}$$

Summing (3.21) and (3.22) gives $(K' - K)g(m') \leq (K' - K)g(m)$ so that

$$g(m') \leq g(m) \ . \tag{3.23}$$

Since $K \geq 0$, (3.21) and (3.23) give $f(m) + Kg(m) \leq f(m') + Kg(m)$, *i.e.*

$$f(m) \leq f(m') \ . \tag{3.24}$$

Moreover, using (3.22), $g(m) = g(m')$, implies $f(m') \leq f(m)$, *i.e.* $f(m) = f(m')$ by (3.24). In the same way, (3.21) and (3.23) show that $f(m) = f(m')$ imply $g(m) = g(m')$. In both cases, (a) is satisfied. Otherwise, $f(m) < f(m')$ and $g(m) > g(m')$, *i.e.* (b) is satisfied.

The last statement follows by taking $m = \widehat{m}(K)$ and $m' = \widehat{m}(K')$, because g is non-decreasing, so that the minimum of g in $E(K)$ is attained by $\widehat{m}(K)$. □

3.5.4. Concentration inequalities used in the main proofs. We do not always assume in this section that models are made of histograms, but only that they are bounded by some finite A. First, we can control $\overline{\delta}(m)$ with general models and bounded data.

PROPOSITION 3.3. *Assume that* $\|Y\|_\infty \leq A < \infty$. *Then for all* $x \geq 0$, *on an event of probability at least* $1 - 2e^{-x}$:

$$\forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta l(s, s_m) + \left(\frac{4}{\eta} + \frac{8}{3}\right) \frac{A^2 x}{n} \ . \tag{3.25}$$

If moreover

$$Q_m^{(p)} := \frac{n\mathbb{E}\left[p_2(m)\right]}{D_m} > 0 \ , \tag{3.26}$$

on the same event,

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s, s_m)}{\sqrt{D_m}} + \frac{20}{3} \frac{A^2}{Q_m^{(p)}} \frac{\mathbb{E}[p_2(m)]}{\sqrt{D_m}} x \ . \tag{3.27}$$

REMARK 3.1. In the histogram case,

$$Q_m^{(p)} = \frac{1}{D_m} \sum_{\lambda \in \Lambda_m} \left[(\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2 \right] \geq (\sigma_{\min})^2 > 0 \ .$$

Then, we derive a concentration inequality for $p_2(m)$ in the histogram case from a general result of [**BM04**] (Thm. 2.2 in a preliminary version).

PROPOSITION 3.4. *Let* S_m *be the model of histograms associated with the partition* $(I_\lambda)_{\lambda \in \Lambda_m}$. *Assume that* $\|Y\|_\infty \leq A$ *and define* $p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$.

Then, for every $x \geq 0$, *there exists an event of probability at least* $1 - e^{1-x}$ *on which for every* $\theta \in (0; 1)$,

$$|p_2(m) - \mathbb{E}\left[p_2(m)\right]| \leq C \left[\theta l(s, s_m) + \frac{A^2 \sqrt{D_m} \sqrt{x}}{n} + \frac{A^2 x}{\theta n} \right] \tag{3.28}$$

for some absolute constant C. If moreover $\sigma(X) \geq \sigma_{\min} > 0$ a.s., we have on the same event:

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq \frac{C}{\sqrt{D_m}} \left[l(s, s_m) + \frac{A^2 \mathbb{E}[p_2(m)]}{\sigma_{\min}^2} \left(\sqrt{x} + x \right) \right] \quad . \tag{3.29}$$

Finally, we recall a concentration inequality for $p_1(m)$ that comes from Sect. 5.7.4. Its proof is particular to the histogram case.

PROPOSITION 3.5 (Prop. 5.8, Sect. 5.8). *Let $\gamma > 0$ and S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda \in \Lambda_m}$. Assume that $\|Y\|_\infty \leq A < \infty$, $\sigma(X) \geq \sigma_{\min} > 0$ a.s. and $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B_n > 0$. Then, if $B_n \geq 1$, on an event of probability at least $1 - Ln^{-\gamma}$,*

$$\widetilde{p_1}(m) \geq \mathbb{E}[\widetilde{p_1}(m)] - L(A, \sigma_{\min}, \gamma) \left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n} \right] \mathbb{E}[p_2(m)] \tag{3.30}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}[\widetilde{p_1}(m)] + L(A, \sigma_{\min}, \gamma) \left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m} e^{-LB_n} \right] \mathbb{E}[p_2(m)] \quad . \tag{3.31}$$

If we only have a lower bound $B_n > 0$, then, with probability at least $1 - Ln^{-\gamma}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma + 1) B_n^{-1} \ln(n)} - \frac{L(A, \sigma_{\min}, \gamma) \ln(n)^2}{\sqrt{D_m}} \right) \mathbb{E}[p_2(m)] \quad . \tag{3.32}$$

PROOF. We changed a little the assumptions of Prop. 5.8. The result still holds since $P_m^\ell(q) \leq 4A^2 \sigma_{\min}^{-2}$. □

3.5.5. Additional results needed. A crucial result in the proofs of Thm. 3.1 and 3.2 is that $p_1(m)$ and $p_2(m)$ are close in expectation. This comes from Sect. 5.7.2.

PROPOSITION 3.6 (Lemma 5.6, Sect. 5.7.2). *Let S_m be a model of histograms adapted to some partition $(I_\lambda)_{\lambda \in \Lambda_m}$. Assume that $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B > 0$. Then,*

$$\begin{aligned} \left(1 - e^{-B}\right)^2 \mathbb{E}[p_2(m)] &\leq \mathbb{E}[\widetilde{p_1}(m)] \\ &\leq \left[2 \wedge \left(1 + 5.1 \times B^{-1/4} \right) + (B \vee 1) e^{-(B \vee 1)} \right] \mathbb{E}[p_2(m)] \quad . \end{aligned} \tag{3.33}$$

Finally, we need the following technical lemma in the proof of the main theorems.

LEMMA 3.7. *Let $(p_\lambda)_{\lambda \in \Lambda_m}$ be non-negative real numbers of sum 1, $(n\widehat{p}_\lambda)_{\lambda \in \Lambda_m}$ a multinomial vector of parameters $(n; (p_\lambda)_{\lambda \in \Lambda_m})$. Then, for all $\gamma > 0$,*

$$\min_{\lambda \in \Lambda_m} \{n\widehat{p}_\lambda\} \geq \frac{\min_{\lambda \in \Lambda_m} \{np_\lambda\}}{2} - 2(\gamma + 1) \ln(n) \tag{3.34}$$

with probability at least $1 - 2n^{-\gamma}$.

PROOF OF LEMMA 3.7. By Bernstein inequality ([**Mas07**], Prop. 2.9), for all $\lambda \in \Lambda_m$,

$$\mathbb{P}\left(n\widehat{p}_\lambda \geq (1 - \theta) np_\lambda - \sqrt{2npx} - \frac{x}{3} \right) \geq 1 - e^{-x} \quad .$$

Take $x = (\gamma + 1) \ln(n)$ above, and remark that $\sqrt{2npx} \leq \frac{np}{2} + x$. The union bound gives the result since $\mathrm{Card}(\Lambda_m) \leq n$. □

3.5.6. Proof of Prop. 3.3. Since $\|Y\|_\infty \leq A$, we have $\|s\|_\infty \leq A$ and $\|s_m\|_\infty \leq A$. In fact, everything happens as if $S_m \cup \{s\}$ was bounded by A in L^∞.

We have

$$\overline{\delta}(m) = \frac{1}{n} \sum_{i=1}^{n} \left(\gamma(s_m, (X_i, Y_i)) - \gamma(s, (X_i, Y_i)) - \mathbb{E}[\gamma(s_m, (X_i, Y_i)) - \gamma(s, (X_i, Y_i))] \right)$$

and assumptions of Bernstein inequality ([**Mas07**], Prop. 2.9) are fulfilled with

$$c = \frac{8A^2}{3n} \quad \text{and} \quad v = \frac{8A^2 l(s, s_m)}{n}$$

since

$$\|\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i)) - \mathbb{E}\left[\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right]\|_\infty \leq 8A^2$$

and

$$\begin{aligned}\operatorname{var}\left(\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right) &\leq \mathbb{E}\left[\left(\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right)^2\right] \\ &\leq 8A^2 l(s,s_m) \end{aligned} \tag{3.35}$$

because $\|s_m - s\|_\infty \leq 2A$ and

$$(\gamma(t,\cdot) - \gamma(s,\cdot))^2 = (t(X) - s(X))^2 \left(2(Y - s(X)) - t(X) + s(X)\right)^2$$

$$\text{and} \quad \mathbb{E}\left[(Y - s(X))^2 \,\middle|\, X\right] \leq \frac{(2A)^2}{4} = A \ .$$

We obtain that, with probability at least $1 - 2e^{-x}$,

$$\left|\overline{\delta}(m)\right| \leq \sqrt{2vx} + c = \sqrt{\frac{16A^2 l(s,s_m)x}{n}} + \frac{8A^2 x}{3n}$$

and (3.25) follows since $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$. Taking $\eta = D_m^{-1/2} \leq 1$ and using $Q_m^{(p)}$ defined by (3.26), we deduce (3.27).

3.5.7. Proof of Prop. 3.4. We apply here a result from [**BM04**] (Thm. 2.2 in a preliminary version), in which it is only assumed that γ takes its values in $[0;1]$. This is satisfied when $\|Y\|_\infty \leq A = 1/2$. When $A \neq 1/2$, we apply this result to $(2A)^{-1}Y$ and recover the general result by homogeneity.

First, we recall this result in the bounded least-square regression framework. For every $t : \mathcal{X} \mapsto \mathbb{R}$ and $\epsilon > 0$, we define

$$d^2(s,t) = 2l(s,t) \qquad \text{and} \qquad w(\epsilon) = \sqrt{2}\epsilon \ .$$

Let ϕ_m belong to the class of nondecreasing and continuous functions $f : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $x \mapsto f(x)/x$ is nonincreasing on $(0;+\infty)$ and $f(1) \geq 1$. Assume that for every $u \in S_m$ and $\sigma > 0$ such that $\phi_m(\sigma) \leq \sqrt{n}\sigma^2$,

$$\sqrt{n}\mathbb{E}\left[\sup_{t \in S_m,\, d(u,t) \leq \sigma} |\overline{\gamma}_n(u) - \overline{\gamma}_n(t)|\right] \leq \phi_m(\sigma) \ . \tag{3.36}$$

Let $\varepsilon_{\star,m}$ be the unique positive solution of the equation

$$\sqrt{n}\varepsilon_{\star,m}^2 = \phi_m(w(\varepsilon_{\star,m})) \ .$$

Then, there exists some absolute constant C such that for every real number $q \geq 2$ one has

$$\|p_2(m) - \mathbb{E}[p_2(m)]\|_q \leq \frac{C}{\sqrt{n}}\left[\sqrt{2q}\left(\sqrt{l(s,s_m)} \vee \varepsilon_{\star,m}\right) + q\frac{2}{\sqrt{n}}\right] \ . \tag{3.37}$$

For every model S_m of histograms, of dimension D_m as a vector space, we can take

$$\phi_m(\sigma) = 3\sqrt{2}\sqrt{D_m} \times \sigma \qquad \text{in (3.36).} \tag{3.38}$$

The proof of this statement is made below. Then, $\varepsilon_{\star,m} = 6\sqrt{D_m}n^{-1/2}$.

Combining (3.37) with the classical link between moments and concentration (Lemma 8.10 in Sect. 8.6.2), the first result follows. The second result is obtained by taking $\theta = D_m^{-1/2}$, as in Prop. 3.3.

PROOF OF (3.38). Let $u \in S_m$ and $d(u,t) = \sqrt{2}\,\|u(X) - t(X)\|_2$ for every $t : \mathcal{X} \mapsto \mathbb{R}$. Define $\psi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ by

$$\psi(\sigma) = \mathbb{E}\left[\sup_{d(u,t)\leq\sigma,\, t\in S_m} |(P_n - P)(\gamma(u,\cdot) - \gamma(t,\cdot))|\right] .$$

We are looking for some nondecreasing and continuous function $\phi_m : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\phi_m(x)/x$ is nonincreasing, $\phi_m(1) \geq 1$ and for every $u \in S_m$,

$$\forall \sigma > 0 \quad \text{such that} \quad \phi_m(\sigma) \leq \sqrt{n}\sigma^2 \ , \qquad \phi_m(\sigma) \geq \sqrt{n}\psi(\sigma) \ .$$

We first look at a general upperbound on ψ.

Assume that $u = s_m$. If this is not the case, the triangular inequality shows that $\psi_{\text{general } u} \leq 2\psi_{u=s_m}$. Let us write

$$t = \sum_{\lambda\in\Lambda_m} t_\lambda \mathbb{1}_{I_\lambda} \qquad u = s_m = \sum_{\lambda\in\Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \ .$$

Computation of $P(\gamma(t,\cdot) - \gamma(u,\cdot))$. for some general $t \in S_m$:

$$\begin{aligned} P(\gamma(t,\cdot) - \gamma(u,\cdot)) &= \mathbb{E}\left[(t(X) - Y)^2 - (u(X) - Y)^2\right] \\ &= \mathbb{E}\left[(t(X) - u(X))^2\right] + 2\mathbb{E}\left[(t(X) - u(X))(u(X) - s(X))\right] \\ &= \mathbb{E}\left[(t(X) - u(X))^2\right] = \sum_{\lambda\in\Lambda_m} p_\lambda (t_\lambda - \beta_\lambda)^2 \end{aligned}$$

since for every $\lambda \in \Lambda_m$, $\mathbb{E}\left[s(X) \,|\, X \in I_\lambda\right] = \beta_\lambda$.

Computation of $P_n(\gamma(t,\cdot) - \gamma(u,\cdot))$. for some general $t \in S_m$: with $\eta_i = Y_i - u(X_i)$, we have

$$\begin{aligned} P_n(\gamma(t,\cdot) - \gamma(u,\cdot)) &= \frac{1}{n}\sum_{i=1}^n (t(X_i) - u(X_i))^2 - \frac{2}{n}\sum_{i=1}^n \left[(t(X_i) - u(X_i))\eta_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda - u_\lambda)^2 \mathbb{1}_{X_i\in I_\lambda} - \frac{2}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda - u_\lambda)\mathbb{1}_{X_i\in I_\lambda}\eta_i \ . \end{aligned}$$

Back to $(P_n - P)$. We sum the two inequalities above and use the triangular inequality

$$\begin{aligned} |(P_n - P)(\gamma(t,\cdot) - \gamma(u,\cdot))| &\leq \left|\frac{1}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda - u_\lambda)^2 (\mathbb{1}_{X_i\in I_\lambda} - p_\lambda)\right| \\ &\quad + \left|\frac{2}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda - u_\lambda)\mathbb{1}_{X_i\in I_\lambda}\eta_i\right| \\ &\leq \frac{2A}{n}\sum_{\lambda\in\Lambda_m}\left[(\sqrt{p_\lambda}\,|t_\lambda - u_\lambda|)\,\frac{|\sum_{i=1}^n (\mathbb{1}_{X_i\in I_\lambda} - p_\lambda)|}{\sqrt{p_\lambda}}\right] \\ &\quad + \frac{2}{n}\sum_{\lambda\in\Lambda_m}\left[(\sqrt{p_\lambda}\,|t_\lambda - u_\lambda|)\,\frac{|\sum_{i=1}^n \mathbb{1}_{X_i\in I_\lambda}\eta_i|}{\sqrt{p_\lambda}}\right] \end{aligned}$$

since $|t_\lambda - u_\lambda| \leq 2A$ for every $t \in S_m$.

We now assume that $d(u,t) \leq \sigma$ for some $\sigma > 0$, *i.e.*

$$d(u,t)^2 = 2 \sum_{\lambda \in \Lambda_m} p_\lambda (t_\lambda - u_\lambda)^2 \leq \sigma^2 \ .$$

From Cauchy-Schwarz inequality, we obtain for every $t \in S_m$ such that $d(u,t) \leq \sigma$

$$|(P_n - P)(\gamma(t,\cdot) - \gamma(u,\cdot))| \leq \frac{2A\sigma}{\sqrt{2}n} \sqrt{\sum_{\lambda \in \Lambda_m} \frac{\left(\sum_{i=1}^n (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2}{p_\lambda}} + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \frac{\left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2}{p_\lambda}}$$

Back to ψ. The upper bound above does not depend on t, so that the left-hand side of the inequality can be replaced by a supremum over $\{t \in S_m \text{ s.t. } d(u,t) \leq \sigma\}$. Taking expectations and using Jensen's inequality ($\sqrt{\cdot}$ being concave), we obtain an upper bound on ψ:

$$\begin{aligned} \psi(\sigma) \leq \frac{2A\sigma}{\sqrt{2}n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E}\left[\frac{\left(\sum_{i=1}^n (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2}{p_\lambda}\right]} \\ + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2}{p_\lambda}\right]} \end{aligned} \tag{3.39}$$

For every $\lambda \in \Lambda_m$, we have

$$\mathbb{E}\left(\sum_{i=1}^n (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2 = \sum_{i=1}^n \mathbb{E}\left(\mathbb{1}_{X_i \in I_\lambda} - p_\lambda\right)^2 = n p_\lambda \left(1 - p_\lambda\right) \tag{3.40}$$

which simplifies the first term. For the second term, notice that

$$\begin{aligned} \forall i \neq j, \qquad & \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \mathbb{1}_{X_j \in I_\lambda} \eta_i \eta_j\right] = \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i\right] \mathbb{E}\left[\mathbb{1}_{X_j \in I_\lambda} \eta_j\right] \\ \text{and} \quad \forall i, \qquad & \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i\right] = \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \mathbb{E}\left[\eta_i \,|\, \mathbb{1}_{X_i \in I_\lambda}\right]\right] = 0 \end{aligned}$$

since η_i is centered conditionally to $\mathbb{1}_{X_i \in I_\lambda}$. Then,

$$\mathbb{E}\left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2 = \sum_{i=1}^n \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i^2\right] \leq n p_\lambda \left\|\eta\right\|_\infty^2 \leq n p_\lambda (2A)^2 \ . \tag{3.41}$$

Combining (3.39) with (3.40) and (3.41), we deduce that

$$\begin{aligned} \psi(\sigma) &\leq \frac{2A\sigma}{\sqrt{2}\sqrt{n}} \sqrt{D_m - 1} + \frac{2\sqrt{2}A\sigma}{\sqrt{n}} \sqrt{D_m} \\ &\leq 3A\sqrt{2} \frac{\sqrt{D_m}}{\sqrt{n}} \times \sigma \ . \end{aligned}$$

As already noticed, we have to multiply this bound by 2 so that it is valid for every $u \in S_m$ and not only $u = s_m$.

The resulting upper bound (multiplied by $\sqrt{n}$) has all the desired properties for ϕ_m since $6A\sqrt{2}\sqrt{D_m} = 3\sqrt{2D_m} \geq 1$. The result follows. □

CHAPTER 4

Limitations of linear penalties

RÉSUMÉ. Dans ce chapitre, nous mettons en défaut les méthodes de sélection de modèles par pénalisation qui utilisent une pénalité linéaire en la dimension des modèles. C'est le cas de nombreux critères classiques comme C_p de Mallows, AIC ou BIC. Dans un cadre particulier de régression hétéroscédastique, il s'avère que tout modèle choisi par pénalisation linéaire (éventuellement dépendant des données) a un risque bien supérieur à celui de l'oracle. Une telle procédure est donc sous-optimale (au moins asymptotiquement), ce qui montre la nécessité de proposer des méthodes d'estimation de la forme de la pénalité à l'aide des données.

4.1. Introduction

A penalization procedure chooses a model that minimizes the sum of the empirical risk (how does the model fit the data) and some complexity measure of the model (called the penalty). The simplest ones are linear penalties, where the penalty is a linear function of the dimension of the models. By "dimension", we mean any intrinsic complexity measure, *e.g.* the dimension of a vector space, a number of parameters, the Vapnik-Červonenkis dimension (for sets of classifiers), to name but a few. This is for instance the case of AIC (Akaike [**Aka73**]) and Mallows' C_p or C_L (Mallows [**Mal73**]). The latter penalty can be written

$$\mathrm{pen}_{\mathrm{Mallows}}(m) = \frac{2\sigma^2 D_m}{n}$$

or $\dfrac{2\widehat{\sigma^2} D_m}{n}$ where $\widehat{\sigma^2}$ is an estimator of the homoscedastic noise level σ^2 .

The main practical advantage of such procedures is that they do not need much more computations than minimizing the empirical risk over each model. Their complexity is thus reduced to its minimum.

Linear penalties are efficient in several frameworks, *i.e.* in homoscedastic regression. See for instance the works of Shibata [**Shi81**], Li [**Li87**], Baraud [**Bar00, Bar02**] and references therein. In a recent work, Birgé and Massart [**BM06c**] suggest some data-driven calibration of penalties of the form $K \times D_m$. Using the "slope heuristics", they propose an algorithm that computes a constant $\widehat{K}_{\mathrm{slope}}$ from the data. Following a remark that we made in Sect. 3.4, this algorithm still has a rather small complexity, compared to other classical model selection procedures (*e.g.* cross-validation).

All the above penalties are linear, in the sense that we can write

$$\mathrm{pen}(m) = \widehat{K} D_m$$

for some data-dependent constant $\widehat{K}$. The aim of this chapter is to show that linear penalties may fail in some heteroscedastic regression framework. This fact can be conjectured from the computation of the "ideal penalty" made in Chap. 5. Then, we prove that this induces the suboptimality of any linear penalization procedure in some particular framework. This result motivates the introduction of a "slope heuristics" algorithm with non-linear shapes for the penalty, based upon the results of Chap. 3. In view of Chap. 5 and 6, these non-linear shapes can be obtained with V-fold or Resampling penalties.

This chapter is organized as follows. The non-linearity of the ideal penalty is recalled in Sect. 4.2. Then, the suboptimality of linear penalties is tackled in Sect. 4.3. Finally, a preliminary simulation study is reported in Sect. 4.4. The proofs are made in Sect. 4.5.

4.2. Non-linearity of the ideal penalty in the histogram framework

In this chapter, we consider the least-square regression framework, and we assume that each model S_m is the set of histograms adapted to some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of the feature space $\mathcal{X}$. See *e.g.* Sect. 3.2 for the precise framework and some notations. We only recall that the n observations $(X_i, Y_i) \in \mathcal{X}\times\mathcal{Y}$ are i.i.d. and satisfy

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i$$

with $\mathbb{E}\left[\epsilon_i^2 \,\middle|\, X_i\right] = 1$. The noise-level $\sigma : \mathcal{X} \mapsto \mathbb{R}^+$ is unknown and may be heteroscedastic.

A penalization procedure chooses the model

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \right\} \ ,$$

so that the ideal penalty is

$$\mathrm{pen}_{\mathrm{id}}(m) := \left(P - P_n\right)\gamma(\widehat{s}_m) \ .$$

Of course, this quantity is unknown from the user, so it can not be used as a practical penalty. However, according to Thm. 3.1 in Sect. 3.3.1, $\widehat{m}$ satisfies an oracle inequality when $\mathrm{pen}(m)$ is close to $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ for every $m \in \mathcal{M}_n$ (provided $\mathcal{M}_n$ is not too large). Then, the shape of $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ is likely to be the "optimal" shape for a penalty.

In the histogram case, we can compute explicitly this quantity: from (6.6),

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(1 + e^0_{\mathcal{B}(n,\mathbb{P}(X\in I_\lambda))}\right)\left((\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right) \tag{4.1}$$

where

$$\begin{aligned}
e^0_{\mathcal{B}(n,p)} &:= np\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z>0}\right] \qquad \text{with} \quad Z \sim \mathcal{B}(n,p) \\
(\sigma_\lambda^r)^2 &:= \mathbb{E}\left[\left(Y - s(X)\right)^2 \,\middle|\, X \in I_\lambda\right] = \mathbb{E}\left[\left(\sigma(X)\right)^2 \,\middle|\, X \in I_\lambda\right] \\
\left(\sigma_\lambda^d\right)^2 &:= \mathbb{E}\left[\left(s(X) - s_m(X)\right)^2 \,\middle|\, X \in I_\lambda\right] \ .
\end{aligned}$$

There are three differences with the homoscedastic fixed-design case, where $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = 2\sigma^2 D_m n^{-1}$:

(1) $\sum_{\lambda\in\Lambda_m}(\sigma_\lambda^r)^2 \neq D_m\mathbb{E}\left[\sigma(X)^2\right]$. Hence, when the noise is highly heteroscedatic and the weights $(p_\lambda)_{\lambda\in\Lambda_m}$ are far from uniform weights, this part of the penalty is far from being linear in D_m. This holds for instance in experiments S2 and HSd2 (described in Sect. 4.4.2).

(2) $\left(\sigma_\lambda^d\right)^2 \neq 0$. This term may be large when s is far from s_m, *i.e.* when s is highly non-smooth or when m has a small dimension. This may happen for instance when s is the

HeaviSine or the Doppler functions (defined by Donoho and Johnstone [**DJ95**]; see also Sect. 4.4.2).

(3) $e^0_{\mathcal{B}(n,\mathbb{P}(X\in I_\lambda))} \neq 1$. However, according to Lemma 5.3 in Sect. 5.6.1, this term is uniformly close to 1 when $\min_{\lambda\in\Lambda_m}\{n\mathbb{P}(X\in I_\lambda)\}$ is large enough.

We have identified two main reasons why penalties should not be taken linear in the dimension in some particular framework. The next section consider a case where the first situation holds. It indeed turns out that linear penalization procedures are suboptimal in this particular framework.

4.3. Suboptimality of linear penalization

We consider in this section the following particular framework:

$$Y = X + \mathbb{1}_{X\geq\frac{1}{2}}\epsilon$$

with ϵ such that $\mathbb{E}\left[\epsilon^2 \,\middle|\, X\right] = 1$ a.s. and $X \sim \mathcal{U}([0;1])$. Then, $s(X) = X$. For every $k_1, k_2 \in \mathbb{N}\backslash\{0\}$, $S_{(k_1,k_2)}$ denotes the model of histograms adapted to a regular partition of $[0,1/2]$ with k_1 pieces, and a regular partition of $[1/2,1]$ with k_2 pieces. Then, consider the "regular with two bin sizes" family $(S_m)_{m\in\mathcal{M}_n} := \left(S_{(k_1,k_2)}\right)_{1\leq k_1,k_2\leq D_{\max}}$ with $D_{\max} = \lfloor n/(2\ln(n))\rfloor$. In the following, for every $m = (k_1,k_2) \in \mathcal{M}_n$, $D_{m,1} := k_1$ and $D_{m,2} := k_2$ denote the two components of the dimension $D_m = D_{m,1} + D_{m,2}$ of the model m.

The following proposition shows that any linear penalty is suboptimal with such data, if we only consider expectations.

PROPOSITION 4.1. *Let $(X_i,Y_i)_{1\leq i\leq n}$ and $\mathcal{M}_n$ be defined as above. For any $m \in \mathcal{M}_n$ and $K > 0$, we define*

$$\mathrm{crit}_{\mathrm{lin}}(m,K) = \mathbb{E}\left[P_n\gamma(\widehat{s}_m)\right] + KD_m - P\gamma(s)$$

the expected linear penalized criterion. For any $\epsilon \geq 0$, we define

$$\mathcal{M}_{\mathrm{lin},\epsilon} = \left\{ \widetilde{m} \in \mathcal{M}_n \text{ s.t. } \exists K > 0, \quad \mathrm{crit}_{\mathrm{lin}}(\widetilde{m},K) \leq \inf_{m\in\mathcal{M}_n}\left\{\mathrm{crit}_{\mathrm{lin}}(m,K) + \epsilon\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\} \right\} .$$

Then, there are absolute constants n_0 and $\kappa > 0$ such that for every $n \geq n_0$ and $\epsilon > 0$,

$$\inf_{m\in\mathcal{M}_{\mathrm{lin},\epsilon}}\left\{\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\} \geq \left(1 + \kappa\left(\epsilon\vee 1\right)^{-2}\right)\inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\} . \tag{4.2}$$

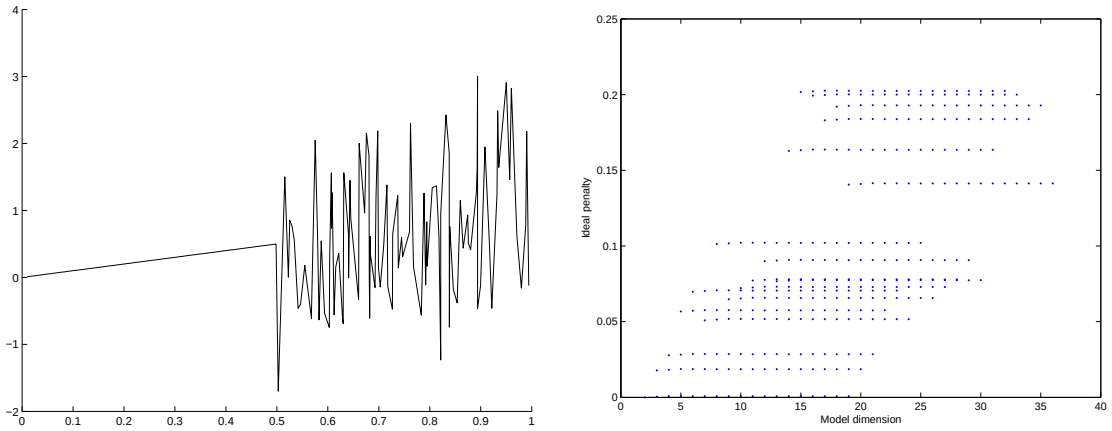
Having in mind concentration results (similar to those of Sect. 5.7.4), the linear criterion $P_n\gamma\left(\widehat{s}_m\right) + KD_m - P\gamma\left(s\right)$ is close to $\mathrm{crit}_{\mathrm{lin}}$ with high probability, up to a remainder smaller than $\epsilon\mathbb{E}\left[l(s,\widehat{s}_m)\right]$, for some $\epsilon < \infty$. Then, the following *optimal linear penalization algorithm* has to belong to $\mathcal{M}_{\mathrm{lin},\epsilon}$:

$$m^\star_{\mathrm{lin}} = \widehat{m}(K^\star) \quad \text{where} \quad K^\star = \arg\min_{K>0}\left\{P\gamma(\widehat{s}_{\widehat{m}(K)})\right\} .$$

So, Prop. 4.1 shows that selecting $m^\star_{\mathrm{lin}}$ is suboptimal.

This is a quite strong result: no linear penalization method can be asymptotically optimal in this case-example, even a method using the data and the knowledge of s and σ ! This rely on the fact that linear penalties can only select a small number of models, which was already noticed by Breiman [**Bre92**] (in its Sect. 5, Breiman caracterizes what those models, called "RSS-extreme submodels"). Whereas Breiman states that this limitation can be benefic in some cases, our Prop. 4.1 shows that it also induces suboptimality in some heteroscedastic frameworks.

4.4. Simulation study



(a) data sample (b) Ideal penalty *vs.* D_m.

FIGURE 4.1. $s(x) = x$, $\sigma(x) = \mathbb{1}_{x \geq \frac{1}{2}}$, $n = 200$.

4.4.1. Framework of Prop. 4.1. We consider in this section data generated as in Prop. 4.1, with a sample size $n = 200$. An instance of data sample is reported on Fig. 4.1a. The corresponding ideal penalty $\mathrm{pen}_{\mathrm{id}}(m)$ is reported as a function of D_m on Fig. 4.1b. It is obviously non-linear.

For each sample, we compared the oracle model

$$m^\star \in \arg\min_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) \}$$

with the models $(\widehat{m}(K))_{K \geq 0}$ that can be chosen by a linear penalty, defined by

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \{ P_n \gamma(\widehat{s}_m) + K D_m \} \quad .$$

The results are reported on Fig. 4.2: the broken lines represent $(D_{\widehat{m}(K),1}, D_{\widehat{m}(K),2})_{K>0}$ for 250 samples, and the stars (on the bottom right) are the 250 corresponding oracle models. Notice that 250 broken lines and 250 stars are superposed, so that we can not distinguish individual paths $(D_{\widehat{m}(K),1}, D_{\widehat{m}(K),2})_{K>0}$. However, we can conclude that the "RSS-extreme submodels" (following Breiman's terminology) are far from the oracle model for each sample.

We observe that the oracle model can never be selected, *i.e.* $m^\star \neq m^\star_{\mathrm{lin}}$. Moreover, on the 250 experiments of Fig. 4.2, we evaluate the benchmarks

$$C_{\mathrm{or}} := \frac{\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m) \right]} \qquad C_{\mathrm{path-or}} := \mathbb{E}\left[\frac{l(s, \widehat{s}_{\widehat{m}})}{\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)} \right] \quad . \tag{4.3}$$

Basically, C_{or} is the constant that should appear in an oracle inequality like (5.12), and $C_{\mathrm{path-or}}$ corresponds to a pathwise oracle inequality like (5.11). We estimate $C_{\text{path-or}}(m^\star_{\mathrm{lin}}) \approx 2.87 \pm 0.12$ and $C_{\mathrm{or}}(m^\star_{\mathrm{lin}}) \approx 2.14 \pm 0.06$. So, the distance between $m^\star_{\mathrm{lin}}$ and $m^\star$ has serious consequences on the risk.

Intuitively, this means that penalizing low-noise and high-noise regions with the same constant K leads to choosing the bin size as if the noise had a high-level everywhere. In the example above, any linear penalization method has thus an additive bias term in its risk, making the model selection suboptimal.

4.4.2. Twelve more experiments. We now consider twelve experiments, under which we also study V-fold and Resampling penalties in Sect. 5.4 and 6.5.

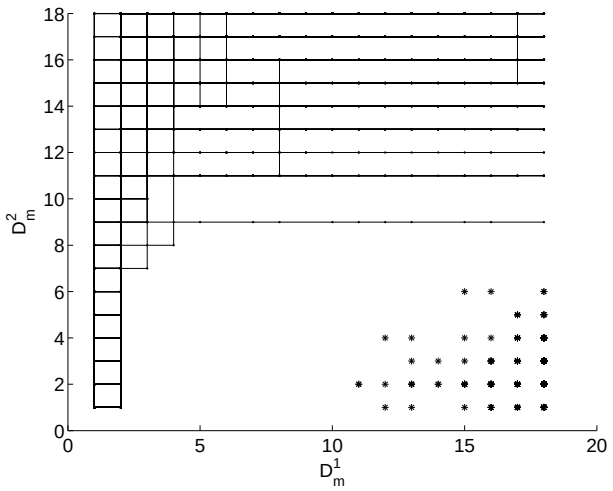


FIGURE 4.2. Simulation results ($n = 200$): no linear penalty can select the oracle on 250 independent samples.

Data are generated according to

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i$$

with X_i i.i.d. uniform on $\mathcal{X} = [0;1]$ and $\epsilon_i \sim \mathcal{N}(0,1)$ independent from X_i. The experiments differ from the regression function s (smooth for S, see Fig. 4.3; smooth with jumps for HS, see Fig. 4.4; Sqrt, His6 and Doppler are represented on Fig. 4.13, 4.15 and 4.17), the noise type (homoscedastic for S1 and HSd1, heteroscedastic for S2 and HSd2) and the number n of data. Instances of data sets are given in Fig. 4.5 to 4.18. Their last difference lies in the families of models. Define

$$\mathcal{M}_n \subset (\mathbb{N}\backslash\{0\}) \cup (\mathbb{N}\backslash\{0\})^2$$

$$\text{where } \forall k, k_1, k_2 \in \mathbb{N}\backslash\{0\}, \quad (I_\lambda)_{\lambda\in\Lambda_k} = \left(\left[\frac{j}{k}; \frac{j+1}{k}\right)\right)_{0\le j\le k-1} \text{ and}$$

$$(I_\lambda)_{\lambda\in\Lambda_{(k_1,k_2)}} = \left(\left[\frac{j}{2k_1}; \frac{j+1}{2k_1}\right)\right)_{0\le j\le k_1-1} \cup \left(\left[\frac{1}{2} + \frac{j}{2k_2}; \frac{1}{2} + \frac{j+1}{2k_2}\right)\right)_{0\le j\le k_2-1}$$

and the following four families of models:

- "regular": regular histograms with $1 \le D \le \frac{n}{\ln(n)}$ pieces, *i.e.*

$$\mathcal{M}_n = \left\{1, \dots, \left\lfloor \frac{n}{\ln(n)} \right\rfloor\right\}$$

- "2 bin sizes": histograms regular on $\left[0; \frac{1}{2}\right]$ and on $\left[\frac{1}{2}; 1\right]$, with D_1 (resp. D_2) pieces, $1 \le D_1, D_2 \le \frac{n}{2\ln(n)}$. The model of constant functions is added to $\mathcal{M}_n$, *i.e.*

$$\mathcal{M}_n = \{1\} \cup \left\{1, \dots, \left\lfloor \frac{n}{2\ln(n)} \right\rfloor\right\}^2$$

- "dyadic": dyadic regular histograms with 2^k pieces, $0 \le k \le \ln_2(n) - 1$, *i.e.*

$$\mathcal{M}_n = \left\{2^k \text{ s.t. } 0 \le k \le \ln_2(n) - 1\right\}$$

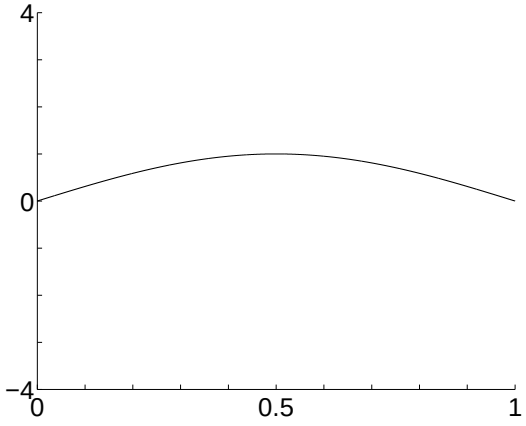


FIGURE 4.3. $s(x) = \sin(\pi x)$

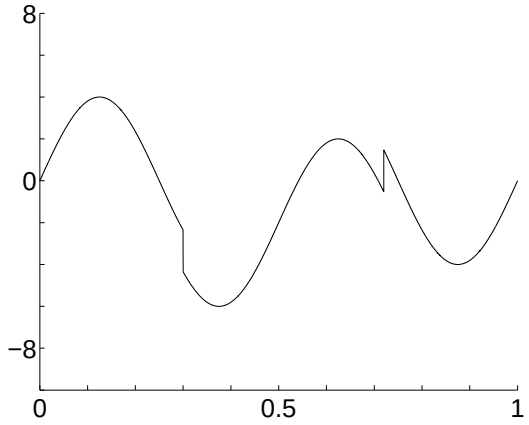


FIGURE 4.4. $s(x) =$ HeaviSine(x) (see [**DJ95**])

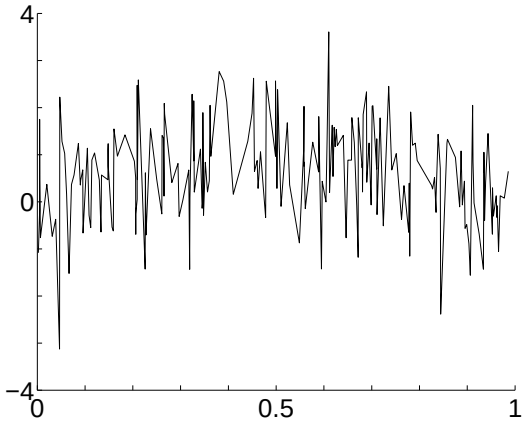


FIGURE 4.5. S1: $s(x) = \sin(\pi x)$, $\sigma \equiv 1$, $n = 200$

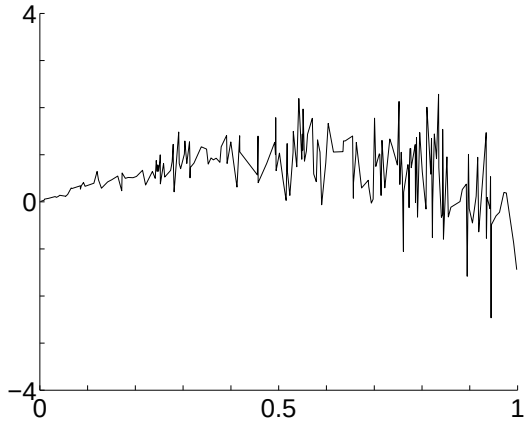


FIGURE 4.6. S2: $s(x) = \sin(\pi x)$, $\sigma(x) = x$, $n = 200$

- "dyadic, 2 bin sizes": dyadic regular histograms with bin sizes 2^{-k_1} and 2^{-k_2}, $0 \leq k_1, k_2 \leq \ln_2(n) - 2$ (dyadic version of S2). The model of constant functions is added to $\mathcal{M}_n$, *i.e.*

$$\mathcal{M}_n = \{1\} \cup \left\{ 2^k \text{ s.t. } 0 \leq k \leq \ln_2(n) - 2 \right\}^2 .$$

For each experiment, s, $\sigma(x)$, n and $\mathcal{M}_n$ are given in Tab. 4.1.

We then compare the following procedures:

- Mal: Mallows' penalty, $\text{pen}(m) = 2\widehat{\sigma}^2 D_m n^{-1}$ where $\widehat{\sigma}^2$ is the classical variance estimator

$$\widehat{\sigma}^2 = \frac{d^2\left(Y_{1\ldots n}, S_{\lfloor n/2 \rfloor}\right)}{n - \lfloor n/2 \rfloor}$$

  (where $S_{\lfloor n/2 \rfloor}$ is any model of dimension $\lfloor n/2 \rfloor$, d the Euclidean distance on $\mathbb{R}^n$ and $Y_{1\ldots n} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$) used by Baraud [**Bar00**], Sect. 6.
- Mal+: Mallows' penalty multiplied by 5/4
- Opt. lin.: Optimal linear penalty $m^\star_{\text{lin}}$
- L.S.H. jump: The linear slope heuristics algorithm of Birgé and Massart [**BM06c**], *i.e.* algorithm 3.1 with $\text{pen}_{\text{shape}}(m) = D_m$ and $\widehat{K}_{\min}$ corresponding to the largest dimension jump.

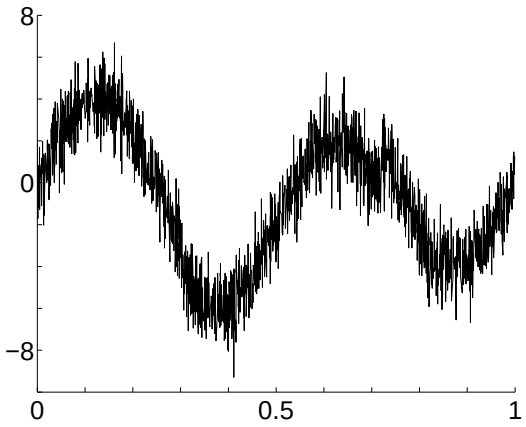


FIGURE 4.7. HSd1: HeaviSine, $\sigma \equiv 1$, $n = 2048$

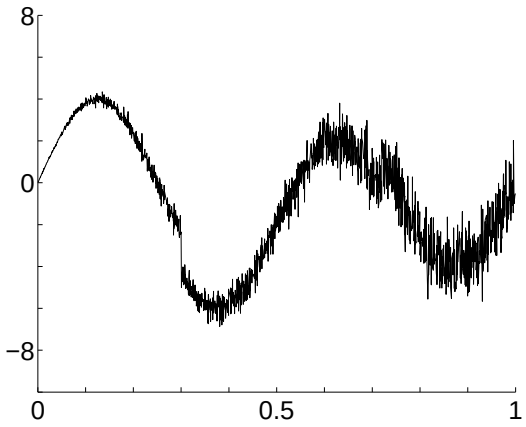


FIGURE 4.8. HSd2: HeaviSine, $\sigma(x) = x$, $n = 2048$

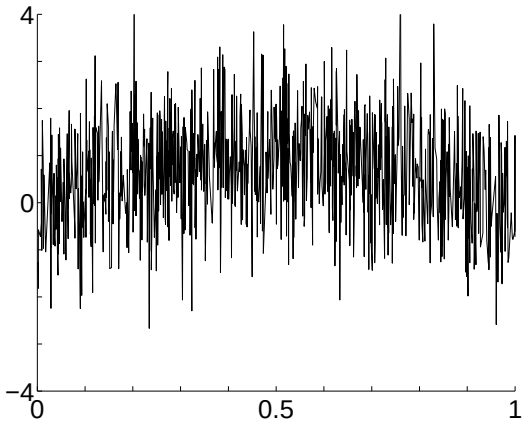


FIGURE 4.9. S1000: sin, $\sigma \equiv 1$, $n = 1000$

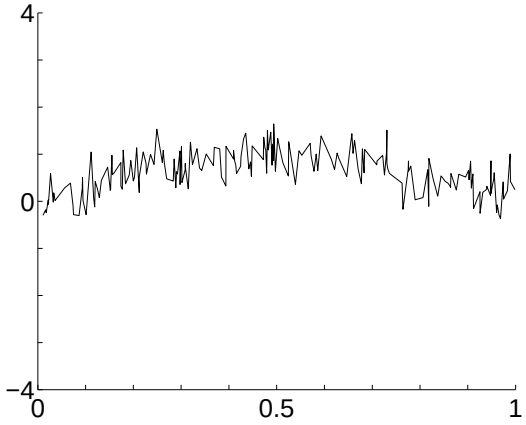


FIGURE 4.10. S$\sqrt{0.1}$: sin, $\sigma \equiv \sqrt{0.1}$, $n = 200$

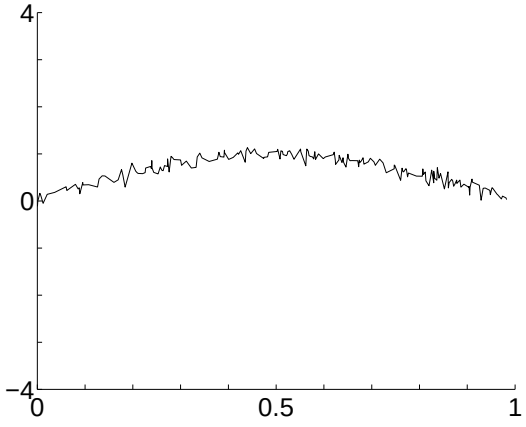


FIGURE 4.11. S0.1: sin, $\sigma \equiv 0.1$, $n = 200$

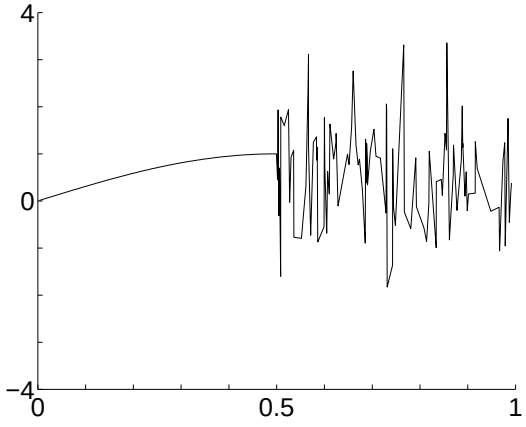


FIGURE 4.12. Svar2: sin, $\sigma(x) = \mathbb{1}_{x \geq \frac{1}{2}}$, $n = 200$

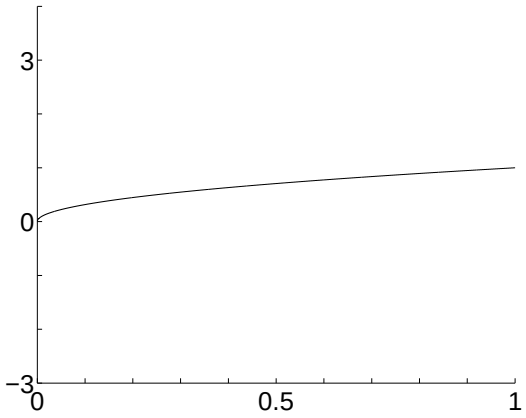


FIGURE 4.13. $s(x) = \sqrt{x}$

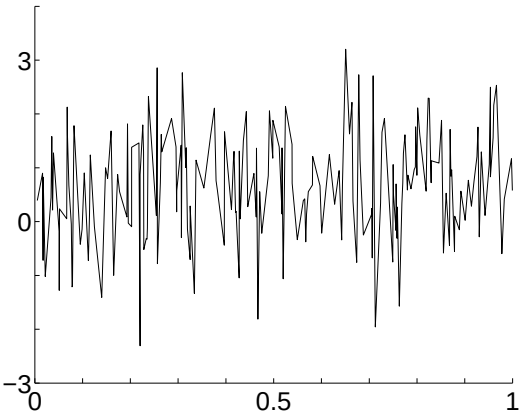


FIGURE 4.14. Sqrt, $\sigma \equiv 1$, $n = 200$

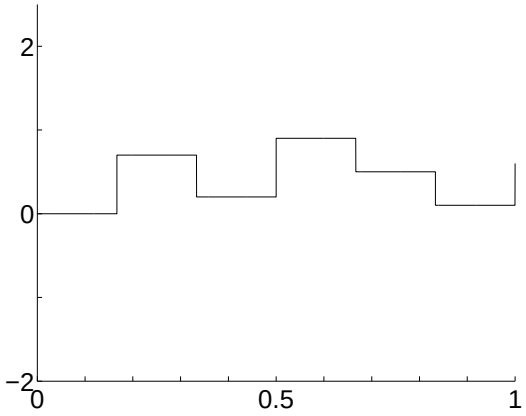


FIGURE 4.15. $s(x) = \mathrm{His}_6(x)$

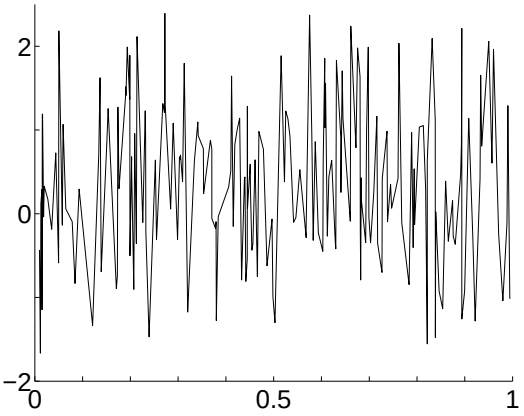


FIGURE 4.16. His6, $\sigma \equiv 1$, $n = 200$

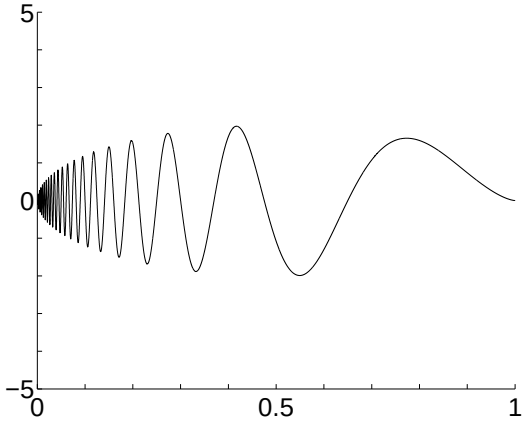


FIGURE 4.17. $s(x) = \mathrm{Doppler}(x)$ (see [**DJ95**])

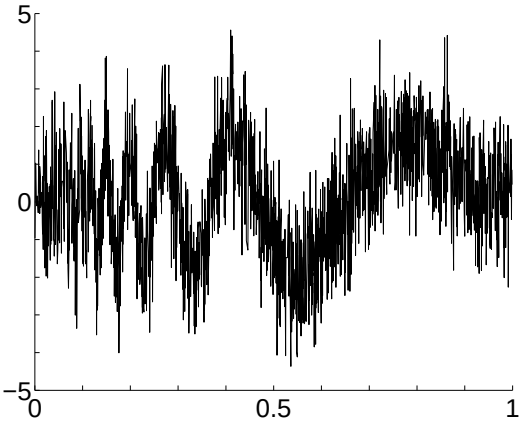
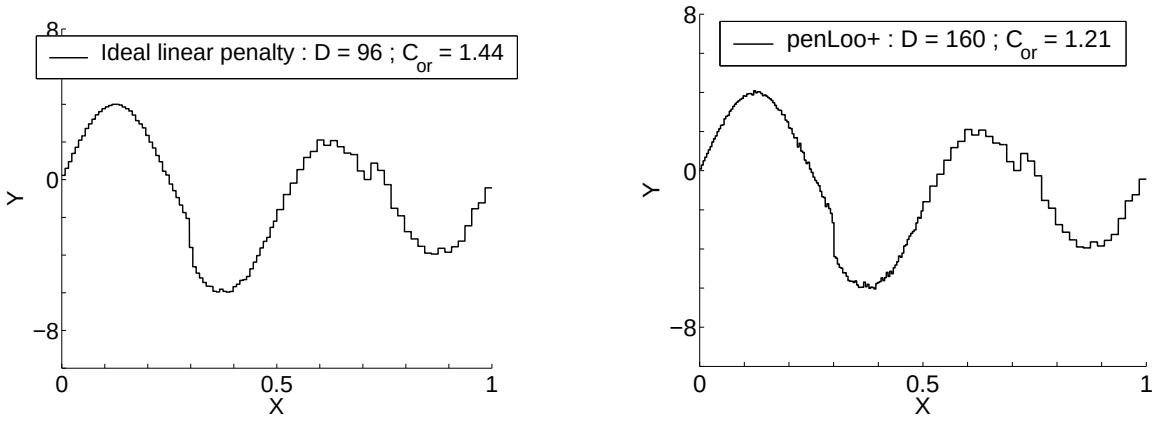


FIGURE 4.18. DopReg=Dop2bin, $\sigma \equiv 1$, $n = 2048$

- L.S.H. reas.: The linear slope heuristics algorithm with $\widehat{K}_{\min}$ corresponding to the "reasonable dimension" criterion, *i.e.* the minimal K such that $D_{\widehat{m}(K)} \leq D_{\max}/2$.
- penLoo: Leave-one-out Resampling penalty (defined as in Sect. 6.5).



(a) Ideal linear penalty (b) penLoo+

FIGURE 4.19. Selected estimators for one sample of experiment HSd2 (for which the dimension of the oracle is 192). The constant C_{or} reported here is the ratio $l(s, \widehat{s}_{\widehat{m}})/l(s, \widehat{s}_{m^\star})$ for this particular sample.

- penLoo+: penLoo multiplied by 5/4.

In each experiment, for each simulated data set, we first remove the models with less than 2 data points in one piece of their associated partition. Then, we compute the least-square estimators $\widehat{s}_m$ for each $m \in \widehat{\mathcal{M}}_n$. Finally, we select $\widehat{m} \in \widehat{\mathcal{M}}_n$ using each algorithm and compute its true excess risk $l(s, \widehat{s}_{\widehat{m}})$ (and the excess risk of each model $m \in \mathcal{M}_n$). Since we simulate N data sets ($N = 1000$ in the four first experiments, $N = 250$ for the eight other ones), we can then estimate C_{or} and $C_{\text{path}-\text{or}}$. As C_{or} and $C_{\text{path}-\text{or}}$ approximatively give the same rankings between algorithms, we only report C_{or} in Tab. 4.1.

It appears that $m^\star_{\text{lin}}$ has a better performance for 11 experiments. This is not surprising in "easy" situations, where Mallows' is almost optimal, since $m^\star_{\text{lin}}$ is always better than Mallows' C_p. It is less intuitive for Svar2 and Dop2bin, which are more difficult, because of heteroscedasticity or bias. Considering that $m^\star_{\text{lin}}$ uses the knowledge of the true distribution P, one can understand that it is sufficient to keep a good performance for "intermediate" problems.

However, in experiment HSd2, the ideal linear penalization has a constant $C_{\text{or}} = 1.18 \pm 0.01$. This is worse than resampling penalization, for which $C_{\text{or}} \leq 1.11$. Thus, the most difficult problem of Sect. 6.5 (with a complex family of models, heteroscedasticity and bias) gives another example where linear penalties are definitely not adapted. On Fig. 4.19, we compared on one sample the estimators selected by (a) the ideal linear penalty and (b) the leave-one-out penalty multiplied by 1.25. Their main difference relies in the estimation of the jump at $X = 0.3$. Whereas the ideal linear penalty splits this jump into two parts, the penLoo+ estimator has only one large jump near $X = 0.3$ (the dyadic partitioning does not allow to find the exact position of the jump). This makes penLoo+ better for prediction than any linear penalization procedure (as shown by the comparison of the prediction risks, see Fig. 4.19.

The results concerning the linear slope heuristics are more difficult to understand. It appears that Mallows' C_p is often better than L.S.H. in frameworks where linear penalties are reasonable. On the contrary, when Mallows' strongly fails, L.S.H. is slighlty better, but worse than non-linear penalties like Leave-one-out penalties. The point here is that we know the asymptotically optimal constant in front of Mallows' penalty in the homoscedastic case. Then, L.S.H. can not do better

TABLE 4.1. Accuracy indexes C_{or} for each algorithm in twelve experiments, $\pm$ a rough estimate of uncertainty of the value reported (*i.e.* the empirical standard deviation divided by $\sqrt{N}$). In each column, the more accurate algorithms (taking the uncertainty into account) are bolded.

Experiment	S1	S2	HSd1	HSd2
s	sin	sin	HeaviSine	HeaviSine
$\sigma(x)$	1	x	1	x
n (data)	200	200	2048	2048
$\mathcal{M}_n$	regular	2 bin sizes	dyadic, regular	dyadic, 2 bin sizes
Mal	1.928 ± 0.04	3.864 ± 0.02	1.606 ± 0.015	1.487 ± 0.011
Mal+	1.800 ± 0.03	4.047 ± 0.02	1.606 ± 0.015	1.487 ± 0.011
Opt. lin.	$\mathbf{1.472 \pm 0.02}$	$\mathbf{1.747 \pm 0.02}$	$\mathbf{1.000 \pm 0.002}$	1.180 ± 0.005
L.S.H. jump	2.008 ± 0.04	3.289 ± 0.04	1.606 ± 0.015	1.487 ± 0.011
L.S.H. reas.	1.882 ± 0.03	3.565 ± 0.04	1.606 ± 0.015	1.487 ± 0.011
penLoo	2.080 ± 0.05	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penLoo+	1.844 ± 0.03	2.215 ± 0.05	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.096 \pm 0.004}$

Experiment	S1000	$S\sqrt{0.1}$	S0.1	Svar2
s	sin	sin	sin	sin
$\sigma(x)$	1	$\sqrt{0.1}$	0.1	$\mathbb{1}_{x \geq 1/2}$
n (data)	1000	200	200	200
$\mathcal{M}_n$	regular	regular	regular	2 bin sizes
Mal	1.667 ± 0.04	1.611 ± 0.03	1.400 ± 0.02	3.520 ± 0.03
Mal+	1.619 ± 0.03	1.593 ± 0.03	1.426 ± 0.02	3.672 ± 0.03
Opt. lin.	$\mathbf{1.414 \pm 0.03}$	$\mathbf{1.373 \pm 0.02}$	$\mathbf{1.253 \pm 0.01}$	$\mathbf{1.980 \pm 0.05}$
L.S.H. jump	1.764 ± 0.05	1.644 ± 0.03	1.602 ± 0.03	2.286 ± 0.10
L.S.H. reas.	1.710 ± 0.04	1.598 ± 0.03	2.238 ± 0.05	2.227 ± 0.11
penLoo	1.776 ± 0.05	1.641 ± 0.04	1.379 ± 0.02	2.656 ± 0.15
penLoo+	1.626 ± 0.03	1.587 ± 0.03	1.401 ± 0.02	2.349 ± 0.13

Experiment	Sqrt	His6	DopReg	Dop2bin
s	$\sqrt{\cdot}$	His_6	Doppler	Doppler
$\sigma(x)$	1	1	1	1
n (data)	200	200	2048	2048
$\mathcal{M}_n$	regular	regular	dyadic, regular	dyadic, 2 bin sizes
Mal	2.295 ± 0.11	1.969 ± 0.11	1.130 ± 0.011	1.469 ± 0.013
Mal+	1.989 ± 0.08	1.799 ± 0.09	1.130 ± 0.011	1.459 ± 0.014
Opt. lin.	$\mathbf{1.405 \pm 0.04}$	$\mathbf{1.266 \pm 0.06}$	$\mathbf{1.000 \pm 0.005}$	$\mathbf{1.009 \pm 0.004}$
L.S.H. jump	2.517 ± 0.11	2.034 ± 0.10	1.130 ± 0.011	1.403 ± 0.016
L.S.H. reas.	2.368 ± 0.13	1.899 ± 0.09	1.130 ± 0.011	1.375 ± 0.016
penLoo	2.695 ± 0.14	2.063 ± 0.12	1.026 ± 0.005	1.058 ± 0.006
penLoo+	2.152 ± 0.10	1.858 ± 0.10	1.082 ± 0.005	1.048 ± 0.006

when $\mathrm{pen}_{\mathrm{id}}(m)$ is indeed linear, and L.S.H. has the faults of linear penalties in more complex frameworks.

Moreover, when the situation of Fig. 3.1b (see Sect. 3.4) occurs, L.S.H. behaves very bad. Since this holds approximatively on 6% of the samples in experiment S1, this problem is sufficient to significantly enlarge the risk of L.S.H. As a consequence, we do not suggest to use slope heuristics in "easy" situations (where the optimal constant is almost known), but only in *a priori* hard frameworks. Notice also that we did not tried to overpenalize L.S.H. within a factor 5/4. In most of the experiments, this may improve a little the performances of the method, as it does with Mallows' and penLoo.

4.5. Proofs

PROOF OF PROP. 4.1. We first introduce the following relevant dimension indexes:

$$D_{m,1} = \mathrm{Card}\left\{\lambda \in \Lambda_m \text{ s.t. } I_\lambda \cap \left[0;\frac{1}{2}\right] \neq \emptyset\right\} \qquad D_{m,2} = D_m - D_{m,1} \ .$$

For every $m \in \mathcal{M}_n$ and every $\lambda \in \Lambda_m$, we have

$$\left(\sigma_\lambda^d\right)^2 = \frac{\mathrm{Leb}(I_\lambda)^2}{12} = \frac{\mathbb{1}_{I_\lambda \subset \left[0;\frac{1}{2}\right]}}{12D_{m,1}^2} + \frac{\mathbb{1}_{I_\lambda \subset \left[\frac{1}{2};1\right]}}{12D_{m,2}^2} \qquad \text{and} \qquad (\sigma_\lambda^r)^2 = \mathbb{1}_{I_\lambda \subset \left[\frac{1}{2};1\right]} \ .$$

Combining this with (4.1) (and its proof, see Sect. 5.7.2), we have

$$\begin{aligned}
\mathbb{E}\left[p_2(m)\right] &= n^{-1} \sum_{\lambda \in \Lambda_m} \left[\left(\sigma_\lambda^d\right)^2 + (\sigma_\lambda^r)^2\right] = \frac{1}{n}\left(\frac{1}{12D_{m,1}} + D_{m,2} + \frac{1}{12D_{m,2}}\right) \\
\mathbb{E}\left[\widetilde{p_1}(m)\right] &= n^{-1} \sum_{\lambda \in \Lambda_m} e^+_{\mathcal{B}(n,p_\lambda)} \left[\left(\sigma_\lambda^d\right)^2 + (\sigma_\lambda^r)^2\right] \\
&= n^{-1}\left[\frac{e^+_{\mathcal{B}(n,D_{m,1}^{-1})}}{12D_{m,1}} + e^+_{\mathcal{B}(n,D_{m,2}^{-1})}\left(\frac{1}{12D_{m,2}} + D_{m,2}\right)\right] \\
&= \frac{1}{n}\left(\frac{1}{12D_{m,1}} + D_{m,2} + \frac{1}{12D_{m,2}}\right) + \frac{\delta_n(m)}{n}
\end{aligned}$$

with $\lim_{n\to\infty} \sup_{m\in\mathcal{M}_n} \delta_n(m) = 0$. This last fact uses Lemma 5.3 (Sect. 5.6.1) and $nD_{m,i}^{-1} \geq nD_{\max} \geq \ln(n)$. Moreover,

$$l(s,s_m) = \sum_{\lambda\in\Lambda_m} \left(\sigma_\lambda^d\right)^2 = \frac{1}{24}\left(\frac{1}{D_{m,1}^2} + \frac{1}{D_{m,2}^2}\right) \ .$$

Then, taking the convention $p_1(m) = \widetilde{p_1}(m)$ in the definition of $P\gamma(\widehat{s}_m)$, we have

$$\begin{aligned}
\mathbb{E}\left[l(s,\widehat{s}_m)\right] &= l(s,s_m) + \mathbb{E}\left[\widetilde{p_1}(m)\right] \\
&= \frac{1}{24}\left(\frac{1}{D_{m,1}^2} + \frac{1}{D_{m,2}^2}\right) + \frac{1}{n}\left(\frac{1}{12D_{m,1}} + D_{m,2} + \frac{1}{12D_{m,2}}\right) + \frac{\delta_n(m)}{n} \ . \qquad (4.4)
\end{aligned}$$

Taking

$$D_{m,1} = D_{\max} \qquad \text{and} \qquad D_{m,2} = \left(\frac{n}{12}\right)^{1/3} \ ,$$

we deduce that for $n \geq L$,

$$\inf_{m\in\mathcal{M}_n} \mathbb{E}\left[l(s,\widehat{s}_m)\right] \leq \kappa_1 n^{-2/3} + \left(n^{-1}\right) \qquad (4.5)$$

with $\kappa_1 = 2^{-4/3}3^{-2/3} + 2^{-2/3}3^{-1/3}$.

On the other hand, for any $K > 0$ and $m \in \mathcal{M}_n$,

$$\begin{aligned}\mathrm{crit}_{\mathrm{lin}}(m,K) &= l(s,s_m) - \mathbb{E}\left[p_2(m)\right] + KD_m \\ &= KD_{m,1} - \frac{1}{12nD_{m,1}} + \frac{1}{24D_{m,1}^2} + \left(K - \frac{1}{n}\right) D_{m,2} + \frac{1}{24D_{m,2}^2} - \frac{1}{12nD_{m,2}} \ . \quad (4.6)\end{aligned}$$

Since $\max(D_{m,1}, D_{m,2}) \leq n\ln(n)^{-1}$, $\mathrm{crit}_{\mathrm{lin}}(m,K)$ is minimal over $m \in \mathcal{M}_n$ when

$$\begin{aligned}D_{m,1} \approx \widehat{D}_1(K) &:= 1 \vee \left[\left(\frac{1}{12K}\right)^{1/3} \wedge D_{\max}\right] \\ D_{m,2} \approx \widehat{D}_2(K) &:= 1 \vee \left[\left(\frac{1}{12(K-n^{-1})_+}\right)^{1/3} \wedge D_{\max}\right] \ .\end{aligned}$$

By convention, when $K \leq n^{-1}$, $\widehat{D}_2(K) = D_{\max}$.

Let $m(K)$ be the model of dimensions $(\widehat{D}_1(K), \widehat{D}_2(K))$. Then, for every $m \in \mathcal{M}_{\mathrm{lin},\epsilon}$, there exists some $K = K(m) > 0$ such that

$$\begin{aligned}\mathrm{crit}_{\mathrm{lin}}(m,K) &\leq \mathrm{crit}_{\mathrm{lin}}(m(K),K) + \epsilon\mathbb{E}\left[l(s,\widehat{s}_{m(K)})\right] \\ &\leq \mathrm{crit}_{\mathrm{lin}}(m(K),K) + L\epsilon\left(n^{-2/3} + K^{2/3}\right) \ . \quad (4.7)\end{aligned}$$

We first assume that $K \leq 1/(2n)$. Then, $D_{m(K)} \geq \widehat{D}_2(K) = D_{\max}$ and

$$\mathrm{crit}_{\mathrm{lin}}(m(K),K) \leq 3^{2/3}2^{-5/3}K^{2/3} + \left(K - \frac{1}{n}\right) D_{\max} + \frac{1}{24D_{\max}^2} \leq -\frac{1}{3\ln(n)}$$

when $n \geq L$. As a consequence, when $n \geq L$,

$$-\frac{D_{m,2}}{n} - \frac{1}{12nD_{m,2}} \leq \mathrm{crit}_{\mathrm{lin}}(m,K) \leq -\frac{1}{3\ln(n)} + Ln^{-2/3} \leq -\frac{1}{4\ln(n)} \ .$$

When $n \geq L$, this implies $D_{m,2} \geq n/(5\ln(n)) \geq 2\kappa_1 n^{1/3}$. According to (4.5), $D_{m,2} \geq 2\kappa_1 n^{1/3}$ implies that for $n \geq L$,

$$\mathbb{E}\left[l(s,\widehat{s}_m)\right] \geq \frac{3}{2} \inf_{m\in\mathcal{M}_n} \left\{\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\} \ .$$

We now only have to consider $m \in \mathcal{M}_{\mathrm{lin},\epsilon}$ such that $D_{m,2} < 2\kappa_1 n^{1/3}$ and for which $K(m) > 1/(2n)$. Starting from (4.6) and (4.7), we have (if $n \geq L$)

$$\begin{aligned}KD_{m,1} - \kappa_1 n^{-2/3} &\leq \mathrm{crit}_{\mathrm{lin}}(m,K) \\ &\leq \mathrm{crit}_{\mathrm{lin}}(m(K),K) + L \times \epsilon\left(n^{-2/3} + K^{2/3}\right) \\ &\leq 3^{2/3}2^{-2/3}K^{2/3} + L \times \epsilon\left(n^{-2/3} + K^{2/3}\right)\end{aligned}$$

so that

$$D_{m,1} \leq L\left(1+\epsilon\right)K^{-1}\left(n^{-2/3} + K^{2/3}\right) \leq \kappa_2(\epsilon)n^{-1/3}$$

where $\kappa_2(\epsilon) = L\left(1+\epsilon\right)$ depends only on ϵ. Then, (4.4) implies

$$\begin{aligned}\mathbb{E}\left[l(s,\widehat{s}_m)\right] &\geq \inf_{1\leq D_{m,2}\leq D_{\max}} \left\{\frac{1}{12D_{m,2}^2} + \frac{1}{12nD_{m,2}} + \frac{D_{m,2}}{n}\right\} + \left(\frac{n^{-2/3}}{24\kappa_2^2} + \frac{n^{-4/3}}{12\kappa_2}\right) + \frac{\delta_n(m)}{n} \\ &\geq \left(1 + \left(25\kappa_1\kappa_2^2\right)^{-1}\right) \inf_{m\in\mathcal{M}_n} \left\{\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\}\end{aligned}$$

for $n \geq L$, and the result follows. □

CHAPTER 5

V-fold cross-validation

RÉSUMÉ. Ce chapitre est consacré à la validation-croisée «V-fold» — qui est très utilisée en pratique — et à une nouvelle méthode de sélection de modèles par pénalisation, que nous appelons la *pénalisation V-fold*.

Nous montrons tout d'abord que la validation-croisée «V-fold» (VFCV) surpénalise d'autant plus que V est petit. Du point de vue asymptotique, il faut donc faire tendre V vers l'infini avec n pour obtenir une procédure optimale. En revanche, lorsque le rapport signal sur bruit est fixé, il peut être bénéfique de surpénaliser. On observe alors que le choix optimal de V n'est pas toujours la plus grande valeur possible compte-tenu du temps de calcul. Des simulations montrent ainsi un exemple où $V = 2$ est meilleur que les choix classiques $V = 5$ ou $V = 10$.

Afin de simplifier le choix de V et d'améliorer les performances de la VFCV, nous définissons la pénalisation V-fold. Cette procédure se fonde sur l'heuristique de rééchantillonnage (Efron [**Efr79**]), et présente une complexité similaire à la VFCV. En revanche, elle permet de choisir le niveau de surpénalisation séparément de V. Dans le cadre de la régression sur des modèles d'histogrammes, nous prouvons une inégalité-oracle non-asymptotique trajectorielle, avec une constante presque 1. On peut en particulier en déduire un résultat d'adaptation à la régularite hölder de la fonction de régression, en présence d'un bruit hétéroscedastique assez général.

Une étude de simulation confirme les études théoriques de la VFCV et de la pénalisation V-fold. En particulier, cette dernière permet effectivement d'améliorer les performances de la VFCV.

5.1. Introduction

There are typically two kinds of model selection criteria. On the one-hand, penalized criteria are the sum of an empirical loss and some penalty term, often measuring the complexity of the models. This is the case of AIC (Akaike [**Aka73**]), Mallows' C_p or C_L (Mallows [**Mal73**]) and BIC (Schwarz [**Sch78**]), to name but a few. On the other hand, cross-validation (Allen [**All74**], Stone [**Sto74**], Geisser [**Gei75**]) and related criteria are based on the idea of data splitting. Part of the data (the training set) is used for fitting each model, and the rest of the data (the validation set) is used to measure the performance of the models. There are several versions of cross-validation (CV), *e.g.* leave-one-out (LOO, also called ordinary CV), leave-p-out (LPO, also called delete-p CV) and generalized CV (Craven and Wahba [**CW79**]).

In practical applications, cross-validation is often computationally very expensive. This is why less greedy CV algorithms have been proposed, among which V-fold cross-validation (VFCV, Geisser [**Gei75**]) and repeated learning testing methods (Breiman *et al.* [**BFOS84**]). In this

chapter, we focus on VFCV, which seems to be the most widely used nowadays. A major problem for the practical user is the choice of V. Basically, V has to be chosen small for complexity reasons, but not too small for decreasing the variability of the algorithm. However, there is a third issue, which is bias: VFCV with V bounded gives an asymptotically biased estimate of the risk, so that it has to be corrected (Burman [**Bur89, Bur90**]). Moreover, when the aim of model selection is model identification rather than prediction, a good criterion should not be an unbiased estimate of the risk. For instance, we basically obtain the BIC criterion by multiplying the exact variance term by $\ln(n)/2$.

The properties of CV (in particular leave-p-out) for prediction and model identification have been widely studied from the asymptotical viewpoint, in particular in regression. It basically depends on the splitting ratio, *i.e.* the ratio between the sizes of the validation and training sets ($p/(n-p)$ in the leave-p-out case; *cf.* [**vdLDK04, Yan06**] and the references therein). For instance, when $p = 1$, the leave-one-out is equivalent to AIC and Mallows' C_p (Li [**Li87**]), *i.e.* it is asymptotically optimal for prediction and inconsistent for model identification. In addition, Shao [**Sha93**] showed that $p \sim n$ and $n - p \to \infty$ are necessary for leave-p-out to be consistent for identification. Several other asymptotic results about CV in regression can be found in [**Zha93, Sha97, GKKW02**].

The choice of p (in leave-p-out) or V (in V-fold CV) is thus crucial and depends on the goal of model selection: prediction or identification. If Shao's result can be extended to VFCV with $p = n/V$, it follows that VFCV can not be used for a consistent identification[1]. On the other hand, optimal prediction for VFCV without correction term needs V to go to infinity (or $p \ll n$ for leave-p-out). In a recent work in the density estimation framework, Celisse and Robin [**CR06**] propose a way of choosing p that realizes the optimal trade-off between bias and variance. When computational issues arise, taking V small is often contradictory with the above requirements. An alternative would be to use the corrected V-fold cross-validation (Burman [**Bur89**]), but its proof is only asymptotical for each model. The practical user of VFCV really needs some help at this point, either for choosing V or for correcting V-fold.

Our main goal in this chapter is to provide such an help. This can be split into two kinds of results. First, we investigate the properties of the uncorrected VFCV from the non-asymptotic viewpoint. We obtain in Sect. 5.2 results that confirms the asymptotic suboptimality of VFCV for a prediction purpose. Although we assume a particular structure for the models (where explicit computations are possible), we allow the noise to be highly *heteroscedastic*. We then expect that the behavior of VFCV is similar in much more general regression settings. For instance, our results explain a strange non-asymptotic behavior of VFCV pointed out in a simulation study (Sect. 5.4). Then, the heuristics we derive about VFCV should be helpful for practical users *in any framework*.

Secondly, we propose in Sect. 5.3 a penalized criterion (called "V-fold penalization", penVF) based upon Efron's resampling heuristics [**Efr79**] (following Fromont's bootstrap penalties [**Fro04**]) and V-fold ideas. To our knowledge, it has never been proposed in the literature, although it finally turns out to be a generalization of Burman's corrected VFCV [**Bur89**]. Thanks to its penalization form, our criteria is much more flexible than VFCV and corrected VFCV. This advantage is confirmed by a simulation study (Sect. 5.4).

This chapter is organized as follows: Sect. 5.2 gives non-asymptotic results about the "classical" VFCV in a non-asymptotic framework. Then, we define V-fold penalties and prove their efficiency

[1] in the same way, Dietterich [**Die98**] and Alpaydin [**Alp99**] proposed the choice $V = 2$ in order to minimize the probability of a type I error (see also Zhang [**Zha93**]).

in Sect. 5.3. It is compared with VFCV in a simulation study in Sect. 5.4. Finally, our results are discussed in Sect. 5.5. The two remaining sections are devoted to some probabilistic tools (Sect. 5.6) and proofs (Sect. 5.7).

5.2. Performance of V-fold cross-validation

In this section, we provide a non-asymptotic study of V-fold cross-validation (VFCV) in the least-square regression framework. In order to make explicit computations possible, we focus on the case where all the models are histograms (although we do not assume that the regression function itself is an histogram). This is only a first theoretical step. We use it to derive heuristics, that should help the practical user of VFCV in any framework.

5.2.1. General framework. First consider the general prediction setting: $\mathcal{X} \times \mathcal{Y}$ is a measurable space, P an unknown probability measure on it and we observe some data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ of common law P. Let $\mathcal{S}$ be the set of predictors (measurable functions $\mathcal{X} \mapsto \mathcal{Y}$) and $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$ a contrast function. Given a family $(\widehat{s}_m)_{m \in \mathcal{M}_n}$ of data-dependent predictors, our goal is to find the one minimizing the prediction loss $P\gamma(t)$. We will extensively use this functional notation $Q\gamma(t) := \mathbb{E}_{(X,Y)\sim Q}[\gamma(t,(X,Y))]$, for any probability measure Q on $\mathcal{X} \times \mathcal{Y}$. Notice that the expectation here is only taken w.r.t. (X, Y), so that $Q\gamma(t)$ is random when $t = \widehat{s}_m$ is random. Assuming that there exists a minimizer $s \in \mathcal{S}$ of the loss (the Bayes predictor), we will often consider the excess loss $l(s,t) = P\gamma(t) - P\gamma(s) \geq 0$ instead of the loss.

We assume that each predictor $\widehat{s}_m$ may be written as a function $\widehat{s}_m(P_n)$ of the empirical distribution of the data $P_n = n^{-1} \sum_{i=1}^n \delta_{(X_i,Y_i)}$. The case-example of such a predictor is the empirical risk minimizer $\widehat{s}_m \in \arg\min_{t \in S_m} \{ P_n\gamma(t) \}$.

In the classical version of VFCV, we first define some partition $(B_j)_{1 \leq j \leq V}$ of the indexes $\{1, \ldots, n\}$. Then, we define

$$P_n^{(j)} = \frac{1}{\mathrm{Card}(B_j)} \sum_{i \in B_j} \delta_{(X_i,Y_i)} \qquad \widehat{s}_m^{(j)} = \widehat{s}_m\left(P_n^{(j)}\right)$$

$$P_n^{(-j)} = \frac{1}{n - \mathrm{Card}(B_j)} \sum_{i \notin B_j} \delta_{(X_i,Y_i)} \qquad \widehat{s}_m^{(-j)} = \widehat{s}_m\left(P_n^{(-j)}\right) \quad .$$

The final VFCV estimator is $\widehat{s}_{\widehat{m}_{\mathrm{VFCV}}}(P_n)$ with

$$\widehat{m}_{\mathrm{VFCV}} \in \arg\min_{m \in \mathcal{M}_n} \{\mathrm{crit}_{\mathrm{VFCV}}(m)\} \quad \text{and} \quad \mathrm{crit}_{\mathrm{VFCV}}(m) := \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} \gamma\left(\widehat{s}_m^{(-j)}\right) \quad . \tag{5.1}$$

We often assume that the partition $(B_j)_{1 \leq j \leq V}$ is regular, *i.e.* that

$$\mathrm{Card}(B_j) \in \{ \lfloor n/V \rfloor, \lfloor n/V \rfloor + 1 \} \quad .$$

When V divides n, this can be done exactly, *i.e.* we can have $\mathrm{Card}(B_j) = n/V$ for every j.

5.2.2. The histogram regression case. In this chapter, our theoretical results tackle the regression framework. The data $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ are i.i.d. of common law P. Denoting by s the regression function, we have

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{5.2}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms, possibly dependent from X_i, but with variance 1 conditionally to X_i. Throughout this chapter, we always

assume that there is some noise:

$$\|\sigma\|_2^2 = \|\sigma(X)\|_2^2 = \mathbb{E}\left[\sigma(X)^2\right] = \mathbb{E}\left[\epsilon^2\right] > 0 \ .$$

The feature space $\mathcal{X}$ is typically a compact subset of $\mathbb{R}^d$. We use the least-square contrast $\gamma : (t,(x,y)) \mapsto (t(x)-y)^2$ to measure the quality of a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$. As a consequence, the Bayes predictor is the regression function s, and the excess loss is $l(s,t) = \mathbb{E}_{(X,Y)\sim P}\left(t(X)-s(X)\right)^2$. To each model S_m, we associate the *empirical risk minimizer*

$$\widehat{s}_m := \widehat{s}_m(P_n) = \arg\min_{t\in S_m} \{ P_n\gamma(t) \}$$

(when it exists and is unique).

Each model in $(S_m)_{m\in\mathcal{M}_n}$ is the set of piecewise constant functions (histograms) on some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$. It is thus a vector space of dimension $D_m = \mathrm{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda\in\Lambda_m}$. As this basis is orthogonal in $L^2(\mu)$ for any probability measure on $\mathcal{X}$, we can make explicit computations. The following notations will be useful throughout this chapter.

$$p_\lambda := P(X\in I_\lambda) \qquad \widehat{p}_\lambda := P_n(X\in I_\lambda) \qquad s_m := \arg\min_{t\in S_m} P\gamma(t)$$

$$(\sigma_\lambda^r)^2 := \mathbb{E}\left[\sigma(X)^2 \,\middle|\, X\in I_\lambda\right] \qquad \left(\sigma_\lambda^d\right)^2 := \mathbb{E}\left[(s-s_m)^2(X) \,\middle|\, X\in I_\lambda\right]$$

$$(\sigma_\lambda)^2 := \mathbb{E}\left[(Y-s(X))^2 \,\middle|\, X\in I_\lambda\right] = (\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2 \ .$$

Remark that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i.

Prop. 5.1 below compares the V-fold criterion and the ideal criterion $P\gamma(\widehat{s}_m)$ in expectation.

PROPOSITION 5.1. *Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Then, the expectation of the ideal criterion is equal to*

$$\mathbb{E}\left[P\gamma(\widehat{s}_m)\right] = P\gamma(s_m) + \frac{1}{n}\sum_{\lambda\in\Lambda_m}(1+\delta_{n,p_\lambda})\left(\left(\sigma_\lambda^d\right)^2 + (\sigma_\lambda^r)^2\right) \tag{5.3}$$

where $\delta_{n,p}$ only depends on (n,p) and is small when the product np is large: $\lim_{np\to\infty}\delta_{n,p}=0$.

Assume that the blocks $(B_j)_{1\le j\le V}$ have approximately the same size:

$$n^{-1}\max_j \mathrm{Card}(B_j) \le c_B < 1 \qquad \text{and} \qquad \sup_j\left\{\left|\frac{\mathrm{Card}(B_j)}{n}-\frac{1}{V}\right|\right\} \le \epsilon_n^{reg} \xrightarrow[n\to\infty]{} 0 \ . \quad (\mathbf{A_{reg,VF}})$$

Then, the expectation of the V-fold criterion is equal to

$$\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m)\right] = P\gamma(s_m) + \frac{V}{V-1}\times\frac{1}{n}\sum_{\lambda\in\Lambda_m}(1+\delta_{n,p_\lambda})^{(VF)}\left(\left(\sigma_\lambda^d\right)^2 + (\sigma_\lambda^r)^2\right) \tag{5.4}$$

where $\delta_{n,p}^{(VF)}$ only depends on $(n,p,c_B,\epsilon_n^{reg})$ and satisfies $\lim_{np\to\infty}\delta_{n,p}^{(VF)}=0$.

In the proof of Prop. 5.1, we give explicit non-asymptotic upper bounds on $\delta_{n,p}$ and $\delta_{n,p}^{(VF)}$.

REMARK 5.1. Since we deal with histograms, $\widehat{s}_m$ is not defined when $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda = 0$, which occurs with positive probability. We then have to take a convention for computing the expectation of the ideal criterion, and the same kind of problem occur with $\mathrm{crit}_{\mathrm{VFCV}}$. See the proof of Prop. 5.1 for more details.

Prop. 5.1 is consistent with Burman's estimate of the bias of VFCV [**Bur89**]. The major advance here is that it is non-asymptotic, and we have explicit upper bounds on the remainder

terms. The classical V-fold cross-validation is thus "overpenalizing" within a factor[2] $1+1/(2(V-1))$ because it estimates the generalization ability of $\widehat{s}_m^{(-j)}$, which is built upon less data than $\widehat{s}_m$. This interpretation is consistent with the results of van der Laan, Dudoit and Keles [**vdLDK04**] in the density estimation framework. According to the concentration inequalities proven in Sect. 5.7, the asymptotic behaviour of VFCV is given by the expectation of its criterion $\mathrm{crit}_{\mathrm{VFCV}}$ (under the assumptions of Thm. 5.1). Assuming moreover that the bias term $P\gamma(s_m)$ is slowly varying (*e.g.* if s is piecewise C^1 and X is uniform on $\mathcal{X}$), we can derive that VFCV is suboptimal in the following sense:

$$\text{If } V = \mathcal{O}(1) \ , \qquad \liminf_{n\to\infty} \frac{l(s, \widehat{s}_{\widehat{m}_{\mathrm{VFCV}}})}{\inf_{m\in\mathcal{M}_n} \{ l(s, \widehat{s}_m) \}} > 1 \qquad \text{a.s.} \tag{5.5}$$

This enlightens some clues for the choice of V:

- *computational issues*: the smaller V, the faster will be the algorithm.
- *variability* of the algorithm: $V = 2$ is known to be quite variable because of the single split, see Burman [**Bur89**] for an asymptotic study of the variance. If moreover $\widehat{s}_m$ is unstable, then $V = n$ (*i.e.* the leave-one-out) is known to be quite variable, *e.g.* in classification with CART (Hastie, Tibshirani and Friedman [**HTF01**]; see also Breiman [**Bre96**]). It seems to disappear when $\widehat{s}_m$ is more stable (Molinaro, Simon and Pfeiffer [**MSP05**]). Thus, the optimum V for variability is large (and goes to infinity with n), but not necessarily equal to n.
- *overpenalization*: $V/(V-1)$ should not be too far from 1.

Notice that the variability does not matter asymptotically (it is only a second order term), whereas it is crucial to take it into account in practical (thus non-asymptotic) applications. It is also a quite difficult problem, since "there is no universal (valid under all distributions) unbiased estimator of the variance of V-fold cross-validation" (Bengio and Grandvalet [**BG04**]).

Moreover, from the non-asymptotic viewpoint (n small and σ large, or s irregular), it is known that overpenalization (*i.e.* positively biased penalties) gives better results (*cf.* Sect. 2.4.1, 6.6.1 and 11.3.3). This means that the better V may not always be large for classical V-fold, independently from computational issues. For instance, in the simulation experiment HSd1 (see Sect. 5.4, Tab. 5.1), $V = 2$ is better than $V \in \{5, 10, 20, n\}$.

The conclusion of this section is that choosing V for V-fold is a complex issue in practice. Even independently for computational issues, the non-asymptotic trade-off between bias and variance is not that clear. We refer to Celisse and Robin [**CR06**] for an interesting approach to this question. When complexity comes on balance, nothing's certain except that our choice of V will be suboptimal. Moreover, it is impossible to choose V so that $\mathrm{crit}_{\mathrm{VFCV}}$ overpenalizes as much as the BIC criterion. This implies that VFCV is not well suited for the general identification setting (although it can sometimes be used for identification, see Yang [**Yan06, Yan07**]). The next section suggests an answer to these issues.

5.3. An alternative V-fold algorithm: V-fold penalties

According to the previous section, the main drawback of VFCV is that overpenalization (*i.e.* bias), variance and complexity depend on the same parameter V. We propose below a way of decoupling the overpenalization issue from the two other ones. It relies on the penalization idea.

5.3.1. Definition of V-fold penalties.

[2] the ideal penalty being of order $2n^{-1}\sum_{\lambda\in\Lambda_m}\left((\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right)$, the overpenalization factor is the ratio between $1 + V/(V-1)$ and 2.

General case. We come back to the general setting of Sect. 5.2.1. Recall that each predictor $\widehat{s}_m$ can be written as a function $\widehat{s}_m(P_n)$ of the empirical distribution of the data $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$.

Since our goal is prediction, the ideal choice for $\widehat{m}$ is the one which minimizes over $\mathcal{M}_n$ the true prediction risk $P\gamma(\widehat{s}_m(P_n)) = P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}_{\mathrm{id}}(m)$ where the ideal penalty is equal to

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m(P_n)) \ . \tag{5.6}$$

Of course, the ideal penalty, which depends both on the unknown distribution P and the data (through P_n), is unknown from the statistician.

The *resampling heuristics* (introduced by Efron [**Efr79**]) gives a way of estimating such quantities. Basically, it says that one can mimic the relationship between P and P_n by building a n-sample of common distribution P_n (the "resample"). P_n^W denoting the empirical distribution of the resample, then the pair (P, P_n) should be close (in distribution) to the pair (P_n, P_n^W). Then, any functional $F(P, P_n)$ is estimated by $\mathbb{E}_W\left[F(P_n, P_n^W)\right]$, where $\mathbb{E}_W[\cdot]$ denotes expectation w.r.t. the resampling randomness. This heuristics has then been generalized to other resampling schemes, with the exchangeable[3] weighted bootstrap (Mason and Newton [**MN92**], Præstgaard and Wellner [**PW93**]), where

$$P_n^W := \frac{1}{n}\sum_{i=1}^n W_i\delta_{(X_i,Y_i)} \qquad \text{with} \quad W \in \mathbb{R}^n \quad \text{an exchangeable weight vector,}$$

independent from the data. Fromont [**Fro04**] used it successfully to build bootstrap penalties in the classification framework. Considering VFCV as a subsampling scheme (which is a particular case of resampling), we followed the same principles to build the V-fold penalties below.

Algorithm 5.1 (V-fold penalization).

(1) Choose a partition $(B_j)_{1\le j\le V}$ of $\{1,\dots,n\}$ and define $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ with $J \sim \mathcal{U}(\{1,\dots,V\})$ independent from the data ($\mathcal{U}(E)$ denotes the uniform distribution over the set E).

(2) Choose a constant $C \ge C_{W,\infty} = V - 1$.

(3) Compute the following resampling penalty for each $m \in \mathcal{M}_n$:

$$\mathrm{pen}(m) = C\mathbb{E}_W\left[P_n\gamma\left(\widehat{s}_m\left(P_n^W\right)\right) - P_n^W\gamma\left(\widehat{s}_m\left(P_n^W\right)\right)\right] \ .$$

(4) Minimize the penalized empirical criterion to choose $\widehat{m}$ and thus $\widehat{s}_{\widehat{m}}$:

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n}\left\{P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}(m)\right\} \ .$$

At step 1, the partition $(B_j)_{1\le j\le V}$ should be taken as regular as possible.

Remark 5.2. There is a constant $C \neq 1$ in front of the penalty, although there isn't any in Efron's heuristics, because we did not normalize W. Indeed, each W_i has a mean 1 but a variance $(V-1)^{-1} \neq 1$. If the weights W were exchangeable, the asymptotical value of the right normalizing constant $C_{W,\infty}$ could be derived from Theorem 3.6.13 in [**vdVW96**]:

$$C_{W,\infty} \sim_{n\to\infty} \left(n^{-1}\sum_{i=1}^n \mathbb{E}\left(W_i - 1\right)^2\right)^{-1} \sim_{n\to\infty} V - 1 \ .$$

In the case example of histograms, we show that $V - 1$ works in a non-asymptotic framework.

[3] A random vector $W \in \mathbb{R}^n$ is called exchangeable when for every permutation τ, $(W_{\tau(1)},\dots,W_{\tau(n)})$ has the same distribution as W.

Although we built algorithm 5.1 upon Efron's heuristics and Fromont's idea of bootstrap penalties, it turns out to be a generalization of Burman's corrected V-fold cross-validation [**Bur89**]. Indeed, with our notations, Burman's criterion (formula (2.3) in [**Bur89**]) is

$$\begin{aligned}\mathrm{crit}_{\mathrm{corr.VF}}(m) &:= \mathrm{crit}_{\mathrm{VFCV}}(m) + P_n\gamma\left(\widehat{s}_m\right) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right) \\ &= P_n\gamma\left(\widehat{s}_m\right) + \frac{1}{V}\sum_{j=1}^{V}\left[\left(P_n^{(j)} - P_n\right)\gamma\left(\widehat{s}_m^{(-j)}\right)\right] \quad .\end{aligned}$$

On the other hand, our V-fold penalized criteria with $C = V - 1$ is equal to

$$\mathrm{crit}_{\mathrm{penVF}}(m) := P_n\gamma\left(\widehat{s}_m\right) + \frac{V-1}{V}\sum_{j=1}^{V}\left[\left(P_n - P_n^{(-j)}\right)\gamma\left(\widehat{s}_m^{(-j)}\right)\right]$$

since $P_n^W = P_n^{(-J)}$ with $J \sim \mathcal{U}\{1,\dots,V\}$. Assuming for the sake of simplicity that all the blocks of the partition have the same size n/V, we have, for every $j \in \{1,\dots,V\}$, $P_n = V^{-1}P_n^{(j)} + (V-1)V^{-1}P_n^{(-j)}$. We deduce

$$P_n^{(j)} - P_n = \frac{V-1}{V}\left(P_n^{(j)} - P_n^{(-j)}\right) = (V-1)\left(P_n - P_n^{(-j)}\right) \quad ,$$

i.e. $\mathrm{crit}_{\mathrm{corr.VF}}(m) = \mathrm{crit}_{\mathrm{penVF}}(m)$ when $C = V - 1$.

The main advance with V-fold penalization (penVF) is that one can choose a constant $C > V - 1$, *e.g.* for prediction in a non-asymptotic framework, or for model identification ($C \propto (V-1)\ln(n)$ should be convenient). This solves one drawback of V-fold cross-validation, which can not overpenalize more than within a factor 3/2 (by taking $V = 2$, as suggested by Zhang [**Zha93**], Dietterich [**Die98**] and Alpaydin [**Alp99**]).

Moreover, the choice of C is decoupled from the one of V, which has now to be chosen according to the complexity-variance trade-off. Further comments about the choice of C and V are made in Sect. 5.5.

The histogram case. Assuming that S_m is the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$, we can make algorithm 5.1 more explicit. Then, we will be able to study its performance, via Prop. 5.2 and Thm. 5.1. We recall that histograms are not our final goal, but only a convenient setting for which we can derive heuristics for practical use of penVF in any framework. We first introduce some more notations:

$$\begin{aligned}
&\beta_\lambda = \mathbb{E}_{(X,Y)\sim P}\left[Y \mid X \in I_\lambda\right] && \text{so that} \quad s_m = \sum_{\lambda\in\Lambda_m}\beta_\lambda \mathbb{1}_{I_\lambda} \\
&\widehat{\beta}_\lambda = \frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda} Y_i && \text{so that} \quad \widehat{s}_m = \sum_{\lambda\in\Lambda_m}\widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \\
&\widehat{p}_\lambda^W := P_n^W(X\in I_\lambda) = \widehat{p}_\lambda W_\lambda && \text{with} \quad W_\lambda := \frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda} W_i \\
&\widehat{s}_m^W := \arg\min_{t\in S_m} P_n^W\gamma(t) = \sum_{\lambda\in\Lambda_m}\widehat{\beta}_\lambda^W \mathbb{1}_{I_\lambda} && \text{with} \quad \widehat{\beta}_\lambda^W := \frac{1}{n\widehat{p}_\lambda^W}\sum_{X_i\in I_\lambda} W_i Y_i \quad .
\end{aligned}$$

Assuming that $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$ (otherwise, the model m should clearly not be chosen), we can compute the ideal penalty (see (5.19) and (5.26) in Sect. 5.7.2):

$$\mathrm{pen}_{\mathrm{id}}(m) = (P-P_n)\gamma(\widehat{s}_m) = \sum_{\lambda\in\Lambda_m} (p_\lambda + \widehat{p}_\lambda)\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2 + (P-P_n)\gamma(s_m) \ .$$

According to the resampling heuristics, $\mathrm{pen}_{\mathrm{id}}(m)$ is estimated (up to some normalizing constant) by

$$\begin{aligned}\mathrm{pen}(m) &= \mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m^W)\right] \\ &= \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\left(\widehat{p}_\lambda + \widehat{p}_\lambda^W\right)\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] + \mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m)\right] \\ &= \sum_{\lambda\in\Lambda_m}\left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right]\right) \qquad (5.7)\end{aligned}$$

since $\sum_i \mathbb{E}[W_i] = 1$. Indeed, $\mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m)\right]$ estimates the expectation of $(P-P_n)\gamma(s_m)$ which is centered. The penalty (5.7) is well-defined if and only if $\widehat{s}_m^W$ is a.s. uniquely defined, *i.e.* $W_\lambda > 0$ for every $\lambda \in \Lambda_m$ a.s. This is why we modified the definition of the weights in algorithm 5.1, so that this problem does not occur.

ALGORITHM 5.2 (V-fold penalization for histograms).

(1) Replace $\mathcal{M}_n$ by

$$\widehat{\mathcal{M}}_n = \left\{ m \in \mathcal{M}_n \text{ s.t. } \min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq 3 \right\} \ .$$

(2) Choose a constant $C \geq C_{W,\infty} = V - 1$.

(3) For every $m \in \widehat{\mathcal{M}}_n$, choose a partition $(B_j)_{1\leq j\leq V}$ of $\{1,\dots,n\}$ such that

$$\forall \lambda\in\Lambda_m, \forall 1\leq j\leq V, \quad \mathrm{Card}\left(B_j \cap \{i \text{ s.t. } X_i \in I_\lambda\}\right) \in \left\{ \left\lfloor \frac{n\widehat{p}_\lambda}{V} \right\rfloor, \left\lfloor \frac{n\widehat{p}_\lambda}{V} \right\rfloor + 1 \right\} \ .$$

Then, define the weights $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ with $J \sim \mathcal{U}(\{1,\dots,V\})$ independent from the data.

(4) Compute the following resampling penalty for each $m \in \mathcal{M}_n$:

$$\mathrm{pen}(m) = C\mathbb{E}_W\left[P_n\gamma\left(\widehat{s}_m\left(P_n^W\right)\right) - P_n^W\gamma\left(\widehat{s}_m\left(P_n^W\right)\right)\right] \ . \qquad (5.8)$$

(5) Minimize the penalized empirical criterion to choose $\widehat{m}$ and thus $\widehat{s}_{\widehat{m}}$:

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n}\left\{P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}(m)\right\} \ .$$

Notice that we choose here a different partition for each model. This is consistent with the proposal of Breiman *et al.* (see also Burman [**Bur90**], Sect. 2) to stratify the data and choose a partition which respects the stratas. Other modifications of algorithm 5.1 are possible. For instance, replace the definition of $\mathrm{pen}(m)$ by the following:

$$\mathrm{pen}(m) = C\left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0\right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right]\right) + \infty\mathbb{1}_{n\widehat{p}_\lambda\leq 1} \ . \qquad (5.9)$$

This is what we did in the simulations of Sect. 5.4. A short theoretical study of these V-fold weigths is done in Sect. 8.4.1. See also algorithm 6.2 in Sect. 6.3. In practical applications, both algorithms should give the same results.

5.3.2. Expectations. We are now in position to compute the expectation of the penVF criterion.

PROPOSITION 5.2. *Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ and* pen *be defined as in algorithm 5.2. Then, if* $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq 3$,

$$\mathbb{E}^{\Lambda_m}\left[P_n\gamma(\widehat{s}_m)+\mathrm{pen}(m)\right]=P\gamma(s_m)+\frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(\frac{2C}{V-1}-1+\frac{C}{V-1}\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})}\right)\left(\left(\sigma_\lambda^d\right)^2+(\sigma_\lambda^r)^2\right) \tag{5.10}$$

with $\mathbb{E}^{\Lambda_m}[\cdot]=\mathbb{E}^{\Lambda_m}\left[\cdot\,\middle|\,(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}\right]$ *and*

$$\frac{2}{n\widehat{p}_\lambda-2}\geq\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})}\geq 0\qquad\text{so that}\qquad\lim_{n\widehat{p}_\lambda\to\infty}\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})}=0\ .$$

Thus, taking $C=V-1$ in algorithm 5.2 leads to an almost unbiased procedure. Indeed, when $\min_{\lambda\in\Lambda_m}\{np_\lambda\}$ goes to infinity faster than some constant times $\ln(n)$, so does $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}$ with a large probability. Moreover, following the proof of Lemma 5.3, we can show that

$$\mathbb{E}\left[\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})}\mathbb{1}_{n\widehat{p}_\lambda\geq 3}\right]\leq\kappa\left(1\wedge(np_\lambda)^{-1/4}\right)\xrightarrow[np_\lambda\to\infty]{}0$$

for some absolute constant $\kappa>0$. This is consistent with the asymptotic computations of Burman [**Bur89**].

Moreover, the flexibility of choice of the constant $C>0$ allows to overpenalize within any factor. This may be crucial in the non-asymptotic framework. Moreover, for an identification purpose, $C=(V-1)\ln(n)/2$ gives a criterion analogous to BIC in the least-square regression framework.

5.3.3. Theoretical results for histograms. In this section, we prove that penVF (algorithm 5.2) is asymptotically optimal for prediction. We assume the existence of some non-negative constants $\alpha_{\mathcal{M}}$, $c_{\mathcal{M}}$, c_{rich}, η such that:

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n)\leq c_{\mathcal{M}}n^{\alpha_{\mathcal{M}}}$.
(**P2**) Richness of $\mathcal{M}_n$: $\exists m\in\mathcal{M}_n$ s.t. $D_m\in[\sqrt{n};c_{\mathrm{rich}}\sqrt{n}]$.
(**P3**) The constant C is well chosen: $\eta(V-1)\geq C\geq V-1$.

THEOREM 5.1. *Assume that the (X_i,Y_i)'s satisfy the following:*

(**Ab**) *Bounded data:* $\|Y_i\|_\infty\leq A<\infty$.
(**An**) *Noise-level bounded from below:* $\sigma(X_i)\geq\sigma_{\min}>0$ *a.s.*
(**Ap**) *Polynomial decreasing of the bias: there exists* $\beta_1\geq\beta_2>0$ *and* $C_{\mathrm{b}}^+,C_{\mathrm{b}}^-<0$ *such that*

$$C_{\mathrm{b}}^-D_m^{-\beta_1}\leq l(s,s_m)\leq C_{\mathrm{b}}^+D_m^{-\beta_2}\ .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for $\mathcal{L}(X)$:* $D_m\min_{\lambda\in\Lambda_m}p_\lambda\geq c_{\mathrm{r},\ell}^X$.

Let $\widehat{m}$ be the model chosen by algorithm 5.2 (under restrictions (**P1** $-$ **3**)*). Then, there exists a constant K_1 and a sequence ϵ_n converging to zero at infinity such that, with probability at least* $1-K_1n^{-2}$,

$$l(s,\widehat{s}_{\widehat{m}})\leq[2\eta-1+\epsilon_n]\inf_{m\in\mathcal{M}_n}\{l(s,\widehat{s}_m)\}\ . \tag{5.11}$$

Moreover, we have the oracle inequality

$$\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right]\leq[2\eta-1+\epsilon_n]\,\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\{l(s,\widehat{s}_m)\}\right]+\frac{A^2K_1}{n^2}\ . \tag{5.12}$$

The constant K_1 may depend on V and constants in (**Ab**), (**An**), (**Ap**), ($\mathbf{Ar}_\ell^{\mathbf{X}}$) *and* (**P1** $-$ **4**)*, but not on n. The small term ϵ_n depends only on n (it can for instance be upperbounded by* $\ln(n)^{-1/5}$*).*

Thm. 5.1 implies the *asymptotic optimality* of the V-fold penalization (algorithm 5.2) as soon as $\lim_{n\to\infty} C = V - 1$ and V is fixed. According to the computations of Sect. 5.2.2, this is not always the case for the classical V-fold cross-validation. Actually, we even have a much stronger result with (5.11) when $C = V - 1$: an oracle inequality with constant almost one on a set of large probability.

Moreover, our result can handle several kinds of *heteroscedastic noises*, while algorithm 5.2 does not use any knowledge about σ, $\|Y\|_\infty$ or the smoothness of s. As a consequence, as long as $\mathcal{M}_n$ allows adaptation, V-fold penalization is a *naturally adaptive algorithm*.

REMARK 5.3.
- In Thm. 5.1, we assume that V is fixed when n grows. A careful look at the proof shows that we only need $V \leq \ln(n)$ for n large enough. With some more work, we could go up to V of order n^ϵ for some $\epsilon > 0$ depending on the assumptions of Thm. 5.1, but we can not handle the leave-one-out case ($V = n$). In Chap. 6, we prove a result similar to Thm. 5.1 for several exchangeable weights, including leave-one-out.
- Assumptions (**Ab**) and (**An**) can be relaxed. Several alternative assumptions are given in Chap. 6.
- Assumption ($\mathbf{Ar}_\ell^\mathbf{X}$) is satisfied if X has a lower bounded density w.r.t. Leb and the histograms are "almost regular". For instance, all the simulation experiments of Sect. 5.4 satisfy this assumption, even S2 or HSd2 in which the histograms are quite irregular.
- The lower bound in (**Ap**) seems a bit surprising. It is not too restrictive because non-constant hölderian functions satisfy (**Ap**) with

$$\beta_1 = k^{-1} + \alpha^{-1} - (k-1)k^{-1}\alpha^{-1} \qquad \text{and} \qquad \beta_2 = 2\alpha k^{-1}$$

when X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$ (*cf.* Sect. 8.10 for more details).

It is quite classical to assume that $l(s, s_m) > 0$ for every model m in the proof of the asymptotic optimality (for prediction) of Mallows' C_p and cross-validation (*e.g.* by Shibata [**Shi81**], Li [**Li87**], Birgé and Massart [**BM06c**]). In our non-asymptotic framework, we need an explicit lower bound on $D_{\widehat{m}}$ (which has to go to infinity at least at the speed $\ln(n)^7$). The polynomial decreasing (**Ap**) is a convenient sufficient condition for such a lower bound. For the same kind of reasons, Stone [**Sto85**] used this assumption in the density estimation framework (see also Burman [**Bur02**], Lemma 3, for the multidimensional case).

5.4. Simulations

As an illustration of the results of the two previous sections, we compare the performances of penVF (for several values of V), Mallows' C_p and VFCV on some simulated data.

5.4.1. Experimental setup. In the following simulation study, we consider the same data sets as in Sect. 4.4.2. We briefly describe them again below. First, we focus on four main experiments, called S1, S2, HSd1 and HSd2. Data are generated according to (5.2) with X_i i.i.d. uniform on $\mathcal{X} = [0;1]$ and $\epsilon_i \sim \mathcal{N}(0,1)$ independent from X_i. The experiments differ from the regression function s (smooth for S, see Fig. 4.3; smooth with jumps for HS, see Fig. 4.4), the noise type (homoscedastic for S1 and HSd1, heteroscedastic for S2 and HSd2), the number n of data and the families of models (see the top of Tab. 6.2; "regular", "with two bin sizes", "dyadic" and "dyadic, 2 bin sizes" are defined on page 111). Instances of data sets are given in Fig. 4.5 to 4.8 (in Sect. 4.4.2).

We compare the following algorithms:

Mal Mallows' C_p penalty: $\mathrm{pen}(m) = 2\widehat{\sigma}^2 D_m n^{-1}$ where $\widehat{\sigma}^2$ is the classical variance estimator

$$\widehat{\sigma}^2 = \frac{d^2\left(Y_{1\dots n}, S_{\lfloor n/2 \rfloor}\right)}{n - \lfloor n/2 \rfloor}$$

(where $S_{\lfloor n/2 \rfloor}$ is any model of dimension $\lfloor n/2 \rfloor$, d the Euclidean distance on $\mathbb{R}^n$ and $Y_{1\dots n} = (Y_1, \dots, Y_n) \in \mathbb{R}^n$) used by Baraud [**Bar00**], Sect. 6.

VFCV Classical V-fold cross-validation, defined by (5.1), with $V \in \{2, 5, 10, 20\}$.

LOO Classical Leave-one-out (*i.e.* VFCV with $V = n$).

penVF V-fold penalty, with $V \in \{2, 5, 10, 20\}$. $C = C_{W,\infty} = V - 1$. The partition (B_j) is chosen once, as in algorithm 5.1. Then, we do not take into account realizations of W such that $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda^W = 0$, as in (5.9). In practice, this is almost the same as algorithm 5.2.

penLoo V-fold penalty, with $V = n$. $C = C_{W,\infty} = n - 1$.

For Mal, penLoo and penVF, we also consider the same penalties multiplied by 5/4 (denoted by a + symbol added after its shortened name). This intends to test for overpenalization.

In each experiment, for each simulated data set, we first remove the models with less than 2 data points in one piece of their associated partition. Then, we compute the least-square estimators $\widehat{s}_m$ for each $m \in \widehat{\mathcal{M}}_n$. Finally, we select $\widehat{m} \in \widehat{\mathcal{M}}_n$ using each algorithm and compute its true excess risk $l(s, \widehat{s}_{\widehat{m}})$ (and the excess risk of each model $m \in \mathcal{M}_n$). Since we simulate N data sets ($N = 1000$ in the four main experiments), we can then estimate the two following benchmarks:

$$C_{\mathrm{or}} = \frac{\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)\right]} \qquad C_{\mathrm{path-or}} = \mathbb{E}\left[\frac{l(s, \widehat{s}_{\widehat{m}})}{\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)}\right]$$

Basically, C_{or} is the constant that should appear in an oracle inequality like (5.12), and $C_{\mathrm{path-or}}$ corresponds to a pathwise oracle inequality like (5.11). As C_{or} and $C_{\mathrm{path-or}}$ approximatively give the same rankings between algorithms, we only report C_{or} in Tab. 5.1.

5.4.2. Results and comments. First of all, the experiments of Tab. 5.1 show the interest of VFCV (or penVF) in a non-asymptotic framework. Although it can not compete with simple procedures such as Mallows' C_p from the computational viewpoint, it is much more efficient in difficult situations. This is for instance the case when the noise is heteroscedastic (S2 and HSd2) or when the regression function is non-smooth[4] (HSd1 and HSd2). On the other hand, it is slighlty worse than Mallows' for easy problems (S1), but this is only the price for robustness.

Secondly, in the four experiments, the best procedures are always the overpenalizing ones. This is mainly due to the small sample size compared to the high noise-level, since this is no longer the case when σ is smaller (see S0.1, Tab. 5.2), and less obvious when n is larger (see S1000, Tab. 5.2). We would like to enlighten this phenomenon, since it vanishes in the asymptotic framework, and it is quite hard to find in theoretical results. For instance, in Thm. 5.1, the constant K_1 is too large, so that one can not use (5.11) or (5.12) to find the "optimal" non-asymptotic value of the constant C.

Moreover, the need for overpenalization is quite important for understanding VFCV, because it can influence the choice of V. For instance, in HSd1, $V = 2$ is significantly better than $V \in \{5, 10, 20, n\}$, which is highly non intuitive. In S1 and S2, the better V is not quite obvious, but it is certainly not $V = 20$ or n. On the contrary, $V = 20$ and $V = n$ are the best choices in

[4] In the particular case of HSd1, we must add that Mallows' C_p performs quite better when σ^2 is known ($C_{\mathrm{or}} \approx 1.044 \pm 0.004$ for Mal, which is still worse than VFCV and penVF for any V; and $C_{\mathrm{or}} \approx 1.606 \pm 0.015$ for Mal+). This is mainly due to the difficulty of estimating σ^2 accurately when even large models can have a large bias. However, this is no longer the case for HSd2, in which the knowledge of σ^2 does not improve Mal and Mal+.

TABLE 5.1. Accuracy indexes C_{or} for each algorithm in four experiments, $\pm$ a rough estimate of uncertainty of the value reported (*i.e.* the empirical standard deviation divided by $\sqrt{N}$). In each column, the more accurate algorithms (taking the uncertainty into account) are bolded.

Experiment	S1	S2	HSd1	HSd2
s	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	HeaviSine	HeaviSine
$\sigma(x)$	1	x	1	x
n (data)	200	200	2048	2048
$\mathcal{M}_n$	regular	2 bin sizes	dyadic, regular	dyadic, 2 bin sizes
Mal	1.928 ± 0.04	3.864 ± 0.02	1.606 ± 0.015	1.487 ± 0.011
Mal+	$\mathbf{1.800 \pm 0.03}$	4.047 ± 0.02	1.606 ± 0.015	1.487 ± 0.011
2−FCV	2.078 ± 0.04	2.542 ± 0.05	$\mathbf{1.002 \pm 0.003}$	1.184 ± 0.004
5−FCV	2.137 ± 0.04	2.582 ± 0.06	1.014 ± 0.003	1.115 ± 0.005
10−FCV	2.097 ± 0.05	2.603 ± 0.06	1.021 ± 0.003	1.109 ± 0.004
20−FCV	2.088 ± 0.04	2.578 ± 0.06	1.029 ± 0.004	1.105 ± 0.004
LOO	2.077 ± 0.04	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
pen2−F	2.578 ± 0.06	3.061 ± 0.07	1.038 ± 0.004	$\mathbf{1.103 \pm 0.005}$
pen5−F	2.219 ± 0.05	2.750 ± 0.06	1.037 ± 0.004	1.104 ± 0.004
pen10−F	2.121 ± 0.05	2.653 ± 0.06	1.034 ± 0.004	1.104 ± 0.004
pen20−F	2.085 ± 0.04	2.639 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penLoo	2.080 ± 0.05	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
pen2−F+	2.175 ± 0.05	2.748 ± 0.06	1.011 ± 0.003	1.106 ± 0.004
pen5−F+	1.913 ± 0.03	2.378 ± 0.05	$\mathbf{1.006 \pm 0.003}$	$\mathbf{1.102 \pm 0.004}$
pen10−F+	1.872 ± 0.03	$\mathbf{2.285 \pm 0.05}$	$\mathbf{1.005 \pm 0.003}$	$\mathbf{1.098 \pm 0.004}$
pen20−F+	1.898 ± 0.04	$\mathbf{2.254 \pm 0.05}$	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.098 \pm 0.004}$
penLoo+	$\mathbf{1.844 \pm 0.03}$	$\mathbf{2.215 \pm 0.05}$	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.096 \pm 0.004}$

experiment HSd2. The main conclusion here should be that one has to take into account both bias and variance for choosing an optimal V. Then, the optimal procedure is the one which (almost) never underpenalize, while overpenalizing as few as possible. It is not always the larger V, so that a larger computation time does not always improve the accuracy!

Notice also that the variability of the procedures can be compared thanks to the uncertainty estimates reported in Tab. 5.1 to 5.3 (*i.e.* the empirical standard deviations of the ratio $l(s, \widehat{s}_{\widehat{m}})/\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)$). It appears that the leave-one-out is not clearly more variable than V-fold cross-validation with $V = 10$. The bias (which can be either a need or a drawback) is thus the main point for choosing V in this least-square regression framework.

Finally, these results confirm the strength of penVF, both against simple procedures as Mallows' C_p (in hard situations) and against the classical VFCV. In three over four experiments, penVF+ with any $V \in \{5, 10, 20, n\}$ does better than the best choice of V in VFCV; and it is almost the case in the fourth one. This comes from the overpenalizing ability of V-fold penalization. Indeed, in the same experiments, penVF (which coincides with Burman's corrected cross-validation) performs worse than penVF+, and sometimes even worse than VFCV. Remark that this phenomenon is not universal, and strongly non-asymptotic. In particular, the right overpenalization constant depends on the sample size n and the noise-level σ. When n becomes large,

TABLE 5.2. Accuracy indexes C_{or} for more experiments ($N = 250$).

Experiment	S1000	S$\sqrt{0.1}$	S0.1	Svar2
s	sin	sin	sin	sin
$\sigma(x)$	1	$\sqrt{0.1}$	0.1	$\mathbb{1}_{x\geq 1/2}$
n (data)	1000	200	200	200
$\mathcal{M}_n$	regular	regular	regular	2 bin sizes
Mal	$\mathbf{1.667 \pm 0.04}$	$\mathbf{1.611 \pm 0.03}$	$\mathbf{1.400 \pm 0.02}$	3.520 ± 0.03
Mal+	$\mathbf{1.619 \pm 0.03}$	$\mathbf{1.593 \pm 0.03}$	1.426 ± 0.02	3.672 ± 0.03
2−FCV	$\mathbf{1.668 \pm 0.04}$	1.663 ± 0.04	$\mathbf{1.394 \pm 0.02}$	2.960 ± 0.15
5−FCV	1.756 ± 0.07	1.693 ± 0.04	$\mathbf{1.393 \pm 0.02}$	2.950 ± 0.16
10−FCV	1.746 ± 0.04	1.664 ± 0.04	$\mathbf{1.385 \pm 0.02}$	2.681 ± 0.14
20−FCV	1.774 ± 0.05	1.645 ± 0.03	$\mathbf{1.382 \pm 0.02}$	2.742 ± 0.16
LOO	1.768 ± 0.05	1.639 ± 0.04	$\mathbf{1.379 \pm 0.02}$	2.641 ± 0.15
pen2−FCV	2.066 ± 0.08	1.809 ± 0.05	$\mathbf{1.390 \pm 0.02}$	3.209 ± 0.18
pen5−FCV	1.816 ± 0.05	1.638 ± 0.04	$\mathbf{1.400 \pm 0.02}$	2.749 ± 0.15
pen10−FCV	1.783 ± 0.05	1.706 ± 0.04	$\mathbf{1.374 \pm 0.02}$	2.598 ± 0.15
pen20−FCV	1.801 ± 0.05	1.657 ± 0.03	$\mathbf{1.385 \pm 0.02}$	2.684 ± 0.15
penLoo	1.776 ± 0.05	1.641 ± 0.04	$\mathbf{1.379 \pm 0.02}$	2.656 ± 0.15
pen2−FCV+	1.809 ± 0.05	1.714 ± 0.04	1.416 ± 0.02	2.808 ± 0.16
pen5−FCV+	$\mathbf{1.683 \pm 0.04}$	$\mathbf{1.616 \pm 0.03}$	$\mathbf{1.399 \pm 0.02}$	$\mathbf{2.460 \pm 0.14}$
pen10−FCV+	$\mathbf{1.627 \pm 0.04}$	$\mathbf{1.613 \pm 0.03}$	$\mathbf{1.385 \pm 0.02}$	$\mathbf{2.398 \pm 0.14}$
pen20−FCV+	$\mathbf{1.644 \pm 0.04}$	$\mathbf{1.583 \pm 0.03}$	$\mathbf{1.390 \pm 0.02}$	$\mathbf{2.316 \pm 0.13}$
penLoo+	$\mathbf{1.626 \pm 0.03}$	$\mathbf{1.587 \pm 0.03}$	$\mathbf{1.401 \pm 0.02}$	$\mathbf{2.349 \pm 0.13}$

corrected cross-validation (*i.e.* penVF) is of course optimal for prediction. The main point here is that overpenalization may be needed, and V-fold penalization allows to choose to overpenalize.

Then, contrary to VFCV, choosing the optimal V for penVF or penVF+ is much more simple. In all the experiments, the more accurate V is the larger one. For the practical user, the choice of V thus reduces to a trade-off between complexity and variability.

5.4.3. Additional experiments. We report the results of eight more experiments in Tab. 5.2 and 5.3. They are quite similar to the first four ones, since we only changed a few parameters among n, σ and s (s and instances of data sets are reported in Fig. 4.9 to 4.18, in Sect. 4.4.2). Remark that we simulated only $N = 250$ data sets.

The choice of V is still difficult for VFCV: $V = 2$ is optimal in S1000 and Sqrt and $V = 20$ in the six other ones. On the contrary, $V = n$ is (almost) always better for penVF and penVF+, and overpenalization often improves the quality of the algorithm (but not always: see DopReg and S0.1). These eight experiments mainly show that the assumptions of Thm. 5.1 are not necessary for penVF to be efficient.

5.5. Discussion

Time has come for us to give an accurate answer to this practical (but quite hard) question: how to use V-fold?

First of all, the classical V-fold cross-validation is biased and asymptotically suboptimal for prediction. It thus has to be corrected, and we suggest a V-fold penalization algorithm that

TABLE 5.3. Accuracy indexes C_{or} for more experiments ($N = 250$).

Experiment	Sqrt	His6	DopReg	Dop2bin
s	$\sqrt{\cdot}$	His_6	Doppler	Doppler
$\sigma(x)$	1	1	1	1
n (data)	200	200	2048	2048
$\mathcal{M}_n$	regular	regular	dyadic, regular	dyadic, 2 bin sizes
Mal	2.295 ± 0.11	$\mathbf{1.969 \pm 0.11}$	1.130 ± 0.011	1.469 ± 0.013
Mal+	$\mathbf{1.989 \pm 0.08}$	$\mathbf{1.799 \pm 0.09}$	1.130 ± 0.011	1.459 ± 0.014
2−FCV	2.489 ± 0.12	2.788 ± 0.13	1.097 ± 0.005	1.165 ± 0.009
5−FCV	2.777 ± 0.16	2.316 ± 0.12	1.064 ± 0.005	$\mathbf{1.049 \pm 0.006}$
10−FCV	2.571 ± 0.13	2.074 ± 0.11	1.043 ± 0.005	$\mathbf{1.051 \pm 0.006}$
20−FCV	2.561 ± 0.12	2.071 ± 0.11	$\mathbf{1.034 \pm 0.005}$	$\mathbf{1.053 \pm 0.006}$
LOO	2.695 ± 0.14	2.059 ± 0.12	$\mathbf{1.026 \pm 0.005}$	$\mathbf{1.058 \pm 0.006}$
pen2−FCV	4.088 ± 0.23	3.210 ± 0.14	1.048 ± 0.006	1.062 ± 0.006
pen5−FCV	3.024 ± 0.18	2.485 ± 0.13	$\mathbf{1.033 \pm 0.005}$	$\mathbf{1.055 \pm 0.006}$
pen10−FCV	3.009 ± 0.18	2.192 ± 0.12	$\mathbf{1.029 \pm 0.005}$	$\mathbf{1.056 \pm 0.006}$
pen20−FCV	2.723 ± 0.14	2.150 ± 0.12	$\mathbf{1.031 \pm 0.006}$	$\mathbf{1.056 \pm 0.006}$
penLoo	2.695 ± 0.14	2.063 ± 0.12	$\mathbf{1.026 \pm 0.005}$	$\mathbf{1.058 \pm 0.006}$
pen2−FCV+	3.015 ± 0.17	2.728 ± 0.12	1.084 ± 0.004	1.084 ± 0.008
pen5−FCV+	2.409 ± 0.13	2.080 ± 0.09	1.080 ± 0.005	1.063 ± 0.007
pen10−FCV+	2.305 ± 0.11	$\mathbf{1.869 \pm 0.09}$	1.082 ± 0.005	$\mathbf{1.050 \pm 0.006}$
pen20−FCV+	2.180 ± 0.10	$\mathbf{1.832 \pm 0.09}$	1.079 ± 0.005	$\mathbf{1.052 \pm 0.006}$
penLoo+	$\mathbf{2.152 \pm 0.10}$	$\mathbf{1.858 \pm 0.10}$	1.082 ± 0.005	$\mathbf{1.048 \pm 0.006}$

provides such a correction. This algorithm is asymptotically optimal in theory, quite efficient on some simulated data, and has the same computational cost as VFCV.

Secondly, a non-asymptotic phenomenon is likely to arise, that make the problem harder: when the sample size is small and the noise-level large, overpenalizing procedures are more efficient than unbiased ones. Then, our V-fold penalization method allows to choose an overpenalizing factor, whereas VFCV imposes it (through V) and a corrected VFCV forbids it. This flexibility is the main reason why we suggest to use penVF instead of VFCV or Burman's corrected VFCV. Otherwise, V has to be chosen very carefully, taking into account the bias and the possible need for some bias.

We shall now explain how to use V-fold penalties. It depends on two tuning parameters, the number V of folds and the overpenalization factor $C/(V-1)$. The choice of V depends on the trade-off between variability of the algorithm and computational complexity. If the latter one does not matter, the optimal choice is close to $V = n$ (*e.g.* in least-squares regression; the optimal V is a little smaller for "unstable" algorithms such as CART in classification). Otherwise, the choice has to be done by the final user. We refer to asymptotic computations of Burman [**Bur89, Bur90**] and the recent work of Celisse and Robin [**CR06**] for quantitative measures of variability according to V. Further research in that direction would be very useful for practical use of V-fold model selection criteria.

The question of choosing the overpenalization factor is probably harder to solve. According to our simulation study, the optimal one depends at least on the sample size, the noise level and the smoothness of the regression function. Since the first criterion is that the penalty almost

never underestimates the ideal one, a wise choice of C depends on the fluctuations of both the V-fold penalty and the ideal penalty. We thus need a better understanding of the variability of penVF. Another idea would be to replace the conditional expectation in (5.6) by a quantile. If the computation time does not matter, one could also think of using V-fold cross-validation again for choosing the overpenalization factor. We refer to Sect. 6.6 and 11.3.3 for further discussions about overpenalization.

We focused in this chapter on prediction, but one often uses model selection for identification. Overpenalization is then needed, even from the asymptotic viewpoint (think of the BIC penalty; see also Aerts, Claeskens and Hart [**ACH99**]). As a consequence, V-fold cross-validation may be inconsistent in general for any V ($V = 2$ being the better choice, as remarked by Zhang [**Zha93**], Dietterich [**Die98**] and Alpaydin [**Alp99**]), whereas V-fold penalties could work. Indeed, following the idea of the BIC criterion, we conjecture that an overpenalization factor proportional to $\ln(n)$ implies the consistency of V-fold penalization (see Sect. 11.3.1).

5.6. Probabilistic tools

In this section, we give some results that may be interesting independently from the resampling penalization method.

5.6.1. Expectations of inverses of binomials. When we compare

$$\mathbb{E}\left[P\left(\gamma\left(\widehat{s}_m\right)-\gamma\left(s_m\right)\right)\right] \quad \text{and} \quad \mathbb{E}\left[P_n\left(\gamma\left(s_m\right)-\gamma\left(\widehat{s}_m\right)\right)\right] \ ,$$

the quantity

$$e_Z^+ = e_{\mathcal{L}(Z)}^+ := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\,\middle|\, Z>0\right]$$

appears, for some random variables Z with Binomial laws. Such quantities have been considered several times (for instance Lew [**Lew76**] consider general Z, and Znidaric [**Žni05**] investigates the case of Binomial random variables). However, these results are either asymptotic or too general to be accurate. In this section, we give some non-asymptotic bounds on e_Z^+, from which we can recover some of the well-known asymptotic results.

It is useful to notice the following general lower bound, which is a straightforward consequence of Jensen's inequality: if $\mathbb{P}(Z>0)>0$, then

$$e_Z^+ \geq \mathbb{P}\left(Z>0\right) \ . \tag{5.13}$$

We used non-asymptotic concentration inequalities to derive the following upper bound.

LEMMA 5.3. *For any $n \in \mathbb{N}\backslash\{0\}$ and $p \in (0;1]$, $\mathcal{B}(n,p)$ denotes the binomial law with parameters (n,p). Denote $\kappa_3 = 5.1$ and $\kappa_4 = 3.2$. Then, if $np \geq 1$,*

$$\kappa_4 \wedge \left(1+\kappa_3(np)^{-1/4}\right) \geq e_{\mathcal{B}(n,p)}^+ \geq 1-e^{-np} \ . \tag{5.14}$$

As a consequence,

$$\lim_{np\to+\infty} e_{\mathcal{B}(n,p)}^+ = 1 \ . \tag{5.15}$$

5.6.2. Concentration of inverses of multinomials. The concentration inequalities of Lemma 5.4 below are useful to show that $\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}(m)\right]$ is close to its expectation when S_m is an histogram model (we refer to (5.22) for a rigorous definition of $\widetilde{p_1}(m)$, which is equal to $p_1(m) := P\left(\gamma\left(\widehat{s}_m\right)-\gamma\left(s_m\right)\right)$ as soon as $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$; this allows to consider the expectation of $\widetilde{p_1}(m)$). To our knowledge, such results do not exist in the literature. In this section, we use the following notations for every $x, T \geq 0$:

$$\varphi(x) = xe^{-x} \qquad \varphi_1(x) = \varphi\left(x\vee 1\right) \qquad f_T(x) = \frac{1}{x}\wedge T$$

with the convention $f(0)=1$ and $f_T(0)=T$.

LEMMA 5.4. *Let $(X_\lambda)_{\lambda\in\Lambda_m} \sim \mathcal{M}(n;(p_\lambda)_{\lambda\in\Lambda_m})$ be a multinomial random vector such that $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n \geq 1$. Let $(a_\lambda)_{\lambda\in\Lambda_m}$ be a family of non-negative real numbers. We define*

$$Z_{m,T} = \sum_{\lambda\in\Lambda_m} a_\lambda f_T(X_\lambda) \quad .$$

(1) *Lower deviations: let $c_1 = 0.184$. For all $x \geq 0$, with probability at least $1-e^{-x}$,*

$$\begin{aligned}\mathbb{E}[Z_{m,1}] - Z_{m,1} \leq{}& \frac{\varphi_1(c_1 B_n)}{c_1} \sum_{\lambda\in\Lambda_m} \frac{a_\lambda}{np_\lambda} \\ &+ 3\sqrt{2}\sqrt{\sum_{\lambda\in\Lambda_m} \frac{a_\lambda^2}{(np_\lambda)^2}}\sqrt{4D_m\exp(-c_1B_n)+x}\end{aligned} \tag{5.16}$$

(2) *Upper deviations: for all $T\in(0;1]$ and $x\geq 0$, with probability at least $1-e^{-x}$,*

$$\begin{aligned}Z_{m,T} - \mathbb{E}[Z_{m,T}] \leq{}& \frac{\varphi_1(c_2B_n)}{c_2} \sum_{\lambda\in\Lambda_m} \left(\frac{a_\lambda}{np_\lambda}\right) \\ &+ \sqrt{\sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda}\right)^2 (D_m\exp(-c_4B_n)+x)} \\ &\times c_3 \vee \left[\frac{c_5 T\sqrt{x+\exp(-c_4B_n)}}{n\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{a_\lambda}\right\}\sqrt{\sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda}\right)^2}}\right]\end{aligned} \tag{5.17}$$

where $\quad c_2 = 0.28 \quad c_3 = 9.6 \quad c_4 = 0.09 \quad c_5 = 10.5$.

This lemma is proven in Sect. 8.8.

5.6.3. Moment inequalities for some U-statistics. Our explicit computations of both ideal and resampling penalties show that conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda\in\Lambda_m)}$, they are U-statistics of order 2. Moreover, they are all of the form (5.18) below. This is why we prove in this section moment inequalities for such U-statistics. This result may be derived from [**GLZ00**] (possibly with smaller powers of q if more information is available on the variables $\xi_{\lambda,i}$). In Sect. 8.9, we give a simple proof of it, based on moment inequalities of [**BBLM05**].

PROPOSITION 5.5. *Let $(a_\lambda)_{\lambda\in\Lambda_m}$ and $(b_\lambda)_{\lambda\in\Lambda_m}$ be two families of real numbers, $(r_\lambda)_{\lambda\in\Lambda_m}$ a family of integers. For all $\lambda\in\Lambda_m$, let $(\xi_{\lambda,i})_{1\leq i\leq r_\lambda}$ be independent centered random variables admitting $2q$-th moments $m_{2q,\lambda,i}$ for some $q\geq 2$. We define $S_{\lambda,1}$, $S_{\lambda,2}$ and Z as follows:*

$$Z = \sum_{\lambda\in\Lambda_m}\left(a_\lambda S_{\lambda,2} + b_\lambda S_{\lambda,1}^2\right) \quad \textit{with} \quad S_{\lambda,1} = \sum_{i=1}^{r_\lambda}\xi_{\lambda,i} \textit{ and } S_{\lambda,2} = \sum_{i=1}^{r_\lambda}\xi_{\lambda,i}^2 \ . \tag{5.18}$$

Then, for every $q\geq 2$,

$$\begin{aligned}\|Z-\mathbb{E}[Z]\|_q \leq{}& 4\sqrt{\kappa}\sqrt{q}\sqrt{\sum_{\lambda\in\Lambda_m}\left((a_\lambda+b_\lambda)^2\sum_{i=1}^{r_\lambda} m_{2q,\lambda,i}^4\right)} \\ &+ 8\sqrt{2}\kappa q\sqrt{\sum_{\lambda\in\Lambda_m}\left(b_\lambda^2\sum_{1\leq i\neq j\leq r_\lambda} m_{2q,\lambda,i}^2 m_{2q,\lambda,j}^2\right)}.\end{aligned}$$

5.7. Proofs

5.7.1. Notations. Before starting the proofs, we introduce some notations. In the following, when we do not want to write explicitly some constants, we use the letter L. It means "some positive absolute constant, possibly different from a line to another, or even within the same line". When L is not numerical, but depends on some parameters $p_1, \dots, p_k$, it is written $L_{p_1,\dots,p_k}$ of $L(p_1, \dots, p_k)$. When L depends on the constants that appear in a set $(\mathbf{A})$ of assumptions, it is written $L_{(\mathbf{A})}$.

For every model $m \in \mathcal{M}_n$, $\lambda \in \Lambda_m$ and $q > 0$,

$$p_1(m) := P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) \qquad p_2(m) := P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$$
$$\delta(m) := (P_n - P)\gamma(s_m) \qquad \overline{\delta}(m) := (P_n - P)\left(\gamma(s_m) - \gamma(s)\right) \ .$$

For any non-negative random variable Z and $q > 0$, we define

$$e^+_{\mathcal{L}(Z)} := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1} \,\middle|\, Z > 0\right] \qquad e^0_{\mathcal{L}(Z)} := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z>0}\right]$$
$$\text{and} \quad \forall x \geq 0, \quad \varphi(x) = xe^{-x} \qquad \varphi_1(x) = \varphi(x \vee 1) \ .$$

In the histogram case, we also define, for any random variable Z, $q > 0$, $m \in \mathcal{M}_n$ and $\lambda \in \Lambda_m$:

$$\mathbb{E}^{\Lambda_m}\left[Z\right] := \mathbb{E}\left[Z \,\middle|\, \left(\mathbb{1}_{X_i \in I_\lambda}\right)_{1 \leq i \leq n,\, \lambda \in \Lambda_m}\right]$$
$$m_{q,\lambda} := \|Y - s_m(X)\|_{q,\lambda} := \left(\mathbb{E}\left[|Y - s_m(X)|^q \,\middle|\, X \in I_\lambda\right]\right)^{1/q}$$
$$\|Z\|_q^{(\Lambda_m)} := \mathbb{E}^{\Lambda_m}\left[|Z|^q\right]^{1/q} = \mathbb{E}\left[|Z|^q \,\middle|\, \left(\mathbb{1}_{X_i \in I_\lambda}\right)_{1 \leq i \leq n,\, \lambda \in \Lambda_m}\right]^{1/q}$$
$$S_{\lambda,1} := \sum_{X_i \in I_\lambda} (Y_i - \beta_\lambda) \qquad \text{and} \qquad S_{\lambda,2} := \sum_{X_i \in I_\lambda} (Y_i - \beta_\lambda)^2 \ .$$

It is convenient to replace $p_2(m)$ by

$$\widetilde{p_2}(m) := p_2(m) + \frac{1}{n}\sum_{\lambda \in \Lambda_m}\left(\left(\sigma^d_\lambda\right)^2 + \left(\sigma^r_\lambda\right)^2\right)\mathbb{1}_{n\widehat{p}_\lambda = 0}$$

so that, if $\min_{\lambda \in \Lambda_m}\left\{n\widehat{p}_\lambda\right\} \geq 1$ (which will always be assumed in practice),

$$p_2(m) = \widetilde{p_2}(m) \qquad \text{and} \qquad \mathbb{E}^{\Lambda_m}\left[p_2(m)\right] = \mathbb{E}^{\Lambda_m}\left[\widetilde{p_2}(m)\right] = \mathbb{E}\left[\widetilde{p_2}(m)\right] \ .$$

Inside expectations, we will often write p_2 instead of $\widetilde{p_2}(m)$ by convention. When $\min_{\lambda \in \Lambda_m}\left\{np_\lambda\right\}$ is large, this does not make much difference (see Lemma 5.6).

5.7.2. Expectations.

Ideal penalties and classical V-fold.

PROOF OF PROP. 5.1.

Ideal criterion. We only have to compute $\mathbb{E}\left[p_1(m)\right]$. Since s_m minimizes $P\gamma(t)$ over $t \in S_m$, we have (provided that $\widehat{s}_m$ is well-defined)

$$p_1(m) = \sum_{\lambda \in \Lambda_m} p_\lambda\left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2 = \sum_{\lambda \in \Lambda_m} \frac{1}{n^2\widehat{p}_\lambda}\frac{p_\lambda}{\widehat{p}_\lambda}S^2_{\lambda,1} \ . \tag{5.19}$$

Thus,

$$\mathbb{E}^{\Lambda_m}\left[p_1(m)\right] = \frac{1}{n}\sum_{\lambda \in \Lambda_m}\frac{p_\lambda}{\widehat{p}_\lambda}\left(\left(\sigma^d_\lambda\right)^2 + \left(\sigma^r_\lambda\right)^2\right) \ . \tag{5.20}$$

Before computing a complete expectation, we must precise what we do when $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda = 0$. This is why we introduce the following alternative definitions for $p_1(m)$:

$$\widetilde{p_1}^{(0)}(m) = \sum_{\lambda\in\Lambda_m \text{ s.t. } \widehat{p}_\lambda>0} \frac{1}{n^2\widehat{p}_\lambda}\frac{p_\lambda}{\widehat{p}_\lambda}S^2_{\lambda,1} \tag{5.21}$$

$$\widetilde{p_1}(m) = \widetilde{p_1}^{(0)}(m) + \sum_{\lambda\in\Lambda_m \text{ s.t. } \widehat{p}_\lambda=0} p_\lambda\left(\left(\sigma^d_\lambda\right)^2 + (\sigma^r_\lambda)^2\right) \tag{5.22}$$

$$\widetilde{p_1}^{(T)}(m) = \sum_{\substack{\lambda\in\Lambda_m\\ n\widehat{p}_\lambda\geq T^{-1}}} \frac{1}{n^2\widehat{p}_\lambda}\frac{p_\lambda}{\widehat{p}_\lambda}S^2_{\lambda,1} + \sum_{\substack{\lambda\in\Lambda_m\\ n\widehat{p}_\lambda< T^{-1}}} Tp_\lambda\left[(\sigma^d_\lambda)^2 + (\sigma^r_\lambda)^2\right] \tag{5.23}$$

for every $T\in(0;\infty)$. Notice that if $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq T^{-1}\geq 1$, all the definitions of p_1 coincide. Then,

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}^{(0)}(m)\right] = \sum_{\lambda\in\Lambda_m \text{ s.t. } \widehat{p}_\lambda>0} \frac{1}{n}\frac{p_\lambda}{\widehat{p}_\lambda}\left[(\sigma^d_\lambda)^2 + (\sigma^r_\lambda)^2\right] \tag{5.24}$$

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}^{(T)}(m)\right] = \sum_{\lambda\in\Lambda_m}\left[\left(\frac{1}{n\widehat{p}_\lambda}\right)\wedge T\right]p_\lambda\left[(\sigma^d_\lambda)^2 + (\sigma^r_\lambda)^2\right] \quad . \tag{5.25}$$

If we choose $p_1(m) = \widetilde{p_1}(m)$, (5.3) holds with $\delta_{n,p_\lambda} = e^0_{\mathcal{B}(n,p_\lambda)} - 1 + np_\lambda(1-p_\lambda)^n \leq e^0_{\mathcal{B}(n,p_\lambda)} - 1 + np_\lambda\exp(-np_\lambda)$. Taking $p_1(m) = \widetilde{p_1}^{(0)}(m)$, (5.3) follows with $\delta^0_{n,p_\lambda} = e^0_{\mathcal{B}(n,p_\lambda)} - 1$ instead of δ_{n,p_λ}. The control of those two small terms when np_λ is large comes from Lemma 5.3.

V-fold criterion. By definition (5.1), if $\widehat{s}_m^{(-j)}$ is a.s. well-defined for every j,

$$\begin{aligned}\mathrm{crit}_{\mathrm{VFCV}}(m) &= \frac{1}{V}\sum_{j=1}^V P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\\ &= P\gamma\left(s_m\right) + \frac{1}{V}\sum_{j=1}^V\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) - P\gamma\left(s_m\right)\right]\end{aligned}$$

so that

$$\begin{aligned}\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m)\right] &= P\gamma\left(s_m\right) + \frac{1}{V}\sum_{j=1}^V\mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) - P\gamma\left(s_m\right)\right]\\ &= P\gamma\left(s_m\right) + \frac{1}{V}\sum_{j=1}^V\mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right) - P\gamma\left(s_m\right)\right]\\ &= P\gamma\left(s_m\right) + \frac{1}{V}\sum_{j=1}^V\sum_{\lambda\in\Lambda_m} e^0_{\mathcal{B}(n-\mathrm{Card}(B_j),p_\lambda)}\frac{(\sigma^r_\lambda)^2 + \left(\sigma^d_\lambda\right)^2}{n-\mathrm{Card}(B_j)}\\ &= P\gamma\left(s_m\right) + \frac{V}{(V-1)n}\sum_{\lambda\in\Lambda_m}\left(1+\delta^{(VF)}_{n,p_\lambda}\right)\left((\sigma^r_\lambda)^2 + \left(\sigma^d_\lambda\right)^2\right)\end{aligned}$$

with

$$\delta^{(VF)}_{n,p_\lambda} = \frac{1}{V}\sum_{j=1}^V\frac{n-n/V}{n-\mathrm{Card}(B_j)}\left(e^0_{\mathcal{B}(n-\mathrm{Card}(B_j),p_\lambda)} - 1\right) \quad .$$

We here choose to take the second term equal to zero for every $\lambda\in\Lambda_m$ with $\widehat{p}_\lambda = 0$, as in the definition (5.21) of $\widetilde{p_1}^{(0)}(m)$. The advantage of this convention is that it does not involve any unknown quantities.

From Lemma 5.3, we deduce that if $n^{-1}\max_j \mathrm{Card}(B_j) \leq c_B < 1$, then

$$\frac{-(V-1)}{V\,(1-c_B\,)}\exp\left[-np_\lambda\,(1-c_B\,)\right] \leq \delta_{n,p_\lambda}^{(VF)} \leq \frac{\kappa_3(V-1)}{(1-c_B\,)^{5/4}\,V} \times (np_\lambda)^{-1/4}\ .$$

This implies $\lim_{np_\lambda\to\infty}\delta_{n,p_\lambda}^{(VF)}=0$. Another choice (*e.g.* conditioning on $\widehat{p}_\lambda>0$) would lead to the same asymptotics for $\delta_{n,p_\lambda}^{(VF)}$. □

Computations for $p_2(m)$ are similar to those for $p_1(m)$. We obtain the following lemma.

LEMMA 5.6. *If S_m is the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$, then,*

$$p_2(m) = \sum_{\lambda\in\Lambda_m}\widehat{p}_\lambda\left(\beta_\lambda-\widehat{\beta}_\lambda\right)^2 = \sum_{\lambda\in\Lambda_m}\frac{S_{\lambda,1}^2\mathbb{1}_{n\widehat{p}_\lambda>0}}{n^2\widehat{p}_\lambda} \tag{5.26}$$

Thus, if $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq 1$,

$$\mathbb{E}^{\Lambda_m}\left[p_2(m)\right] = \mathbb{E}\left[\widetilde{p_2}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(\left(\sigma_\lambda^d\right)^2+(\sigma_\lambda^r)^2\right)\ . \tag{5.27}$$

As a consequence, if $\min_{\lambda\in\Lambda_m}\{np_\lambda\}\geq B$, *then*

$$\left(1+\inf_{np\geq B}\delta_{n,p}^0\right)\mathbb{E}\left[\widetilde{p_2}(m)\right] \leq \mathbb{E}\left[\widetilde{p_1}^{(0)}(m)\right] \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] \leq \left(1+\sup_{np\geq B}\delta_{n,p}\right)\mathbb{E}\left[\widetilde{p_2}(m)\right] \tag{5.28}$$

$$\textit{with}\quad \delta_{n,p}^0 = e_{\mathcal{B}(n,p)}^0-1\geq e^{-np}\quad \textit{and}\quad \delta_{n,p}\leq e_{\mathcal{B}(n,p)}^0-1+npe^{-np}\ .$$

If $np\geq 1$, we have $\delta_{n,p}\leq L(np)^{-1/4}$. The same kind of inequality holds with $p_2(m)$ instead of $\widetilde{p_2}(m)$ inside the expectations.

V-fold penalties. We conclude this section by the computations related to V-fold penalties.

PROOF OF PROP. 5.2. First notice that

$$\mathbb{E}\left[P_n\gamma\left(\widehat{s}_m\right)+\mathrm{pen}(m)\right] = P\gamma(s_m)-\mathbb{E}\left[p_2(m)\right]+\mathbb{E}\left[\mathrm{pen}(m)\right]\ .$$

For $\mathbb{E}\left[p_2(m)\right]$, we use Lemma 5.6.

For $\mathbb{E}\left[\mathrm{pen}(m)\right]$, we first use a weight modification trick. Until the end of the proof, we work conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$. Let σ be a random permutation of $\{1,\ldots,n\}$, independent from W and D_n, and uniform over the permutations that leave $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$ unchanged (*i.e.* for every $\lambda\in\Lambda_m$, σ permutes the $X_i\in I_\lambda$ together). Define $\widetilde{W}=\left(W_{\sigma(i)}\right)_{1\leq i\leq n}$. Then,

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_W(m)\right] = \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\widetilde{W}}(m)\right]$$

since the penalty does not depend on the order of $(X_i,Y_i)_{X_i\in I_\lambda}$ and $(X_i,Y_i)_{X_i\in I_\lambda}$ is exchangeable. Moreover, for every $\lambda\in\Lambda_m$, $(\widetilde{W}_i)_{X_i\in I_\lambda}$ is exchangeable and independent from $(X_i,Y_i)_{X_i\in I_\lambda}$. We can thus use Lemma 5.7 to compute $\mathrm{pen}_{\widetilde{W}}(m)$. Noticing that

$$\mathbb{E}\left[S_{\lambda,2}\,|\,\widehat{p}_\lambda\right] = \mathbb{E}\left[S_{\lambda,1}^2\,\middle|\,\widehat{p}_\lambda\right] = n\widehat{p}_\lambda\left[(\sigma_\lambda^r)^2+\left(\sigma_\lambda^d\right)^2\right]\ ,$$

we derive

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] = \frac{C}{n}\sum_{\lambda\in\Lambda_m}\left(R_{1,\widetilde{W}}(n,\widehat{p}_\lambda)+R_{2,\widetilde{W}}(n,\widehat{p}_\lambda)\right)\left((\sigma_\lambda^r)^2+\left(\sigma_\lambda^d\right)^2\right)\ .$$

It now remains to compute $R_{1,\widetilde{W}}$ and $R_{2,\widetilde{W}}$. If V divides $n\widehat{p}_\lambda$, then $W_\lambda=1$ a.s. and $R_{1,\widetilde{W}}=R_{2,\widetilde{W}}=(V-1)^{-1}$. For the general case, since $(\widetilde{W}_i)_{X_i\in I_\lambda}$ is exchangeable and $\widetilde{W}_i$ takes

only two values,

$$W_\lambda = \mathbb{E}_W\left[W_i \,|\, W_\lambda\right] = \frac{V}{V-1}\mathbb{P}\left(W_i = \frac{V}{V-1}\,\middle|\, W_\lambda\right) \ .$$

Thus,

$$\mathcal{L}\left(W_i \,|\, W_\lambda\right) = \frac{V}{V-1}\mathcal{B}(\kappa^{-1}W_\lambda)$$

so that

$$R_{2,W}(n,\widehat{p}_\lambda) = \frac{1}{V-1} \qquad \text{and} \qquad R_{1,W}(n,\widehat{p}_\lambda) = \frac{V}{V-1}\mathbb{E}\left(\widetilde{W}_\lambda^{-1}\right) - 1 \ .$$

There exists $a, b \in \mathbb{N}$ such that $0 \leq b \leq V-1$ and $n\widehat{p}_\lambda = aV + b$. Then,

$$\mathbb{P}\left(\widetilde{W}_\lambda = \frac{V(a(V-1)+b)}{(V-1)(aV+b)}\right) = \frac{V-b}{V} \quad \text{and} \quad \mathbb{P}\left(\widetilde{W}_\lambda = \frac{V(a(V-1)+b-1)}{(V-1)(aV+b)}\right) = \frac{b}{V}$$

so that

$$\begin{aligned}\mathbb{E}\left[\widetilde{W}_\lambda^{-1}\right] &= \frac{V-b}{V}\frac{(V-1)(aV+b)}{V(a(V-1)+b)} + \frac{b}{V}\frac{(V-1)(aV+b)}{V(a(V-1)+b-1)} \\ &= 1 - \frac{b}{V(a(V-1)+b)} + \frac{(V-1)(aV+b)b}{V^2(a(V-1)+b-1)(a(V-1)+b)} \ .\end{aligned}$$

We deduce

$$R_{1,\widetilde{W}}(n,\widehat{p}_\lambda) = \frac{1}{V-1} - \frac{b}{(V-1)(a(V-1)+b)} + \frac{(aV+b)b}{V(a(V-1)+b-1)(a(V-1)+b)} \ .$$

The result follows with

$$\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})} = \frac{b}{n\widehat{p}_\lambda - a}\left(\frac{V-1}{V} \times \frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - a - 1} - 1\right) \in \left[0; \frac{2}{n\widehat{p}_\lambda - 2}\right] \ .$$

□

In the proof above, we used the following lemma. We state it for a general weight vector W, for which we may have $\mathbb{P}\left(W_\lambda = 0\right) > 0$. Hence, we cannot use definition (5.8) for the penalty, and we replace it by (5.9). We assume here that $n\widehat{p}_\lambda > 1$ for every $\lambda \in \Lambda_m$ and $m \in \mathcal{M}_n$, so that the last term disappears.

LEMMA 5.7. *Let S_m be the model of histograms adapted to some partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Let $W \in [0;\infty)^n$ be a random vector such that for every $\lambda \in \Lambda_m$, $(W_i)_{X_i\in I_\lambda}$ is exchangeable and independent from $(X_i, Y_i)_{X_i\in I_\lambda}$. Define the Resampling Penalty for histograms as* (5.9)*:*

$$\mathrm{pen}(m) = C\left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0\right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right]\right) \ .$$

Then,

$$\mathrm{pen}(m) = \frac{C}{n}\sum_{\lambda\in\Lambda_m}\left(R_{1,W}(n,\widehat{p}_\lambda) + R_{2,W}(n,\widehat{p}_\lambda)\right)\frac{n\widehat{p}_\lambda S_{\lambda,2} - S_{\lambda,1}^2}{n\widehat{p}_\lambda - 1} \tag{5.29}$$

$$\textit{with} \qquad R_{1,W}(n,\widehat{p}_\lambda) = \mathbb{E}\left[\frac{(W_1 - W_\lambda)^2}{W_\lambda^2}\,\middle|\, X_1 \in I_\lambda, W_\lambda > 0\right] \tag{5.30}$$

$$\textit{and} \qquad R_{2,W}(n,\widehat{p}_\lambda) = \mathbb{E}\left[\frac{(W_1 - W_\lambda)^2}{W_\lambda}\,\middle|\, X_1 \in I_\lambda\right] \ . \tag{5.31}$$

PROOF OF LEMMA 5.7. First, split the penalty (without the constant C) into these two terms:

$$\widehat{p}_1(m) = \sum_{\lambda\in\Lambda_m}\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0\right] \tag{5.32}$$

$$\widehat{p}_2(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] \ . \tag{5.33}$$

This split into two terms is the equivalent of the split of $\mathrm{pen}_{\mathrm{id}}$ into p_1 and p_2 (plus a centered term).

We first compute this quantity, which appears in both $\widehat{p}_1$ and $\widehat{p}_2$: let $\lambda \in \Lambda_m$ and $W_\lambda > 0$,

$$\begin{aligned}
\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda\right] &= \mathbb{E}_W\left[\widehat{p}_\lambda\left(\frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)\left(1-\frac{W_i}{W_\lambda}\right)\right)^2 \,\middle|\, W_\lambda\right] \\
&= \frac{1}{n^2\widehat{p}_\lambda}\left[\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)^2\,\mathbb{E}_W\left[\left(1-\frac{W_i}{W_\lambda}\right)^2 \,\middle|\, W_\lambda\right]\right. \\
&\left. + \frac{1}{n^2\widehat{p}_\lambda}\sum_{i\neq j, X_i\in I_\lambda, X_j\in I_\lambda}(Y_i-\beta_\lambda)(Y_j-\beta_\lambda)\mathbb{E}_W\left[\left(1-\frac{W_i}{W_\lambda}\right)\left(1-\frac{W_j}{W_\lambda}\right) \,\middle|\, W_\lambda\right]\right] \ .
\end{aligned} \tag{5.34}$$

Since the weights are exchangeable, $(W_i)_{X_i\in I_\lambda}$ is also exchangeable conditionally to W_λ and $(X_i)_{1\le i\le n}$. Thus, the "variance" term

$$R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) := \mathbb{E}_W\left[(W_i - W_\lambda)^2 \,\middle|\, W_\lambda\right]$$

does not depend from i (provided that $X_i \in I_\lambda$), and the "covariance" term

$$R_C(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) := \mathbb{E}_W\left[(W_i - W_\lambda)(W_j - W_\lambda) \mid W_\lambda\right]$$

does not depend from (i,j) (provided that $i \neq j$ and $X_i, X_j \in I_\lambda$). Moreover,

$$\begin{aligned}
0 &= \mathbb{E}_W\left[\left(\sum_{X_i\in I_\lambda}(W_i - W_\lambda)\right)^2 \,\middle|\, W_\lambda\right] \\
&= n\widehat{p}_\lambda R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) + n\widehat{p}_\lambda\left(n\widehat{p}_\lambda - 1\right) R_C(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W))
\end{aligned}$$

so that, if $n\widehat{p}_\lambda \ge 2$,

$$R_C(n, n\widehat{p}_\lambda, W_\lambda, W) = \frac{-1}{n\widehat{p}_\lambda - 1} R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) \qquad \text{and} \qquad R_V(n, 1, W_\lambda, \mathcal{L}(W)) = 0 \ . \tag{5.35}$$

Combining (5.34) and (5.35), we obtain

$$\begin{aligned}
\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda\right] &= \frac{R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W))}{W_\lambda n^2\widehat{p}_\lambda}\mathbb{1}_{n\widehat{p}_\lambda\ge 2} \\
&\quad \times\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - 1}S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda - 1}S_{\lambda,1}^2\right]
\end{aligned} \tag{5.36}$$

Combining (5.36) and (5.32) (resp. (5.36) and (5.33)), we have the following expressions for $\widehat{p}_1$ and $\widehat{p}_2$:

$$\widehat{p}_1(m) = \sum_{\lambda\in\Lambda_m}\frac{R_{1,W}(n,\widehat{p}_\lambda)\mathbb{1}_{n\widehat{p}_\lambda\ge 2}}{n^2\widehat{p}_\lambda}\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - 1}S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda - 1}S_{\lambda,1}^2\right] \tag{5.37}$$

$$\widehat{p}_2(m) = \sum_{\lambda\in\Lambda_m}\frac{R_{2,W}(n,\widehat{p}_\lambda)\mathbb{1}_{n\widehat{p}_\lambda\ge 2}}{n^2\widehat{p}_\lambda}\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - 1}S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda - 1}S_{\lambda,1}^2\right] \ . \tag{5.38}$$

Remark that the terms of the sum for which $n\widehat{p}_\lambda = 1$ are all equal to zero, which can be ensured with the convention $0\times\infty = 0$ since $R_{1,W}(n, n^{-1}) = R_{2,W}(n, n^{-1}) = 0$. The result follows. □

5.7.3. Proof of Thm. 5.1. In this section, $L_{(\mathbf{VF})}$ denotes a constant that depends only on the set of assumptions of Thm. 5.1, including V. For each $m \in \mathcal{M}_n$, we have

$$l(s, \widehat{s}_m) = P_n\gamma(\widehat{s}_m) + p_1(m) + p_2(m) - \delta(m) - P\gamma(s) \ .$$

By definition of $\widehat{m}$, for every $m \in \widehat{\mathcal{M}}_n$,

$$P_n\gamma(\widehat{s}_{\widehat{m}}) + \mathrm{pen}(\widehat{m}) \leq P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \ ,$$

so that

$$l(s, \widehat{s}_{\widehat{m}}) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m})\right) \leq l(s, \widehat{s}_m) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m)\right) \tag{5.39}$$

with $\mathrm{pen}'_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}_{\mathrm{id}}(m) + (P - P_n)\gamma(s)$. It is sufficient to control $\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}$ for every $m \in \mathcal{M}_n$. We will thus use the concentration inequalities of Sect. 5.7.4, which need to control the two following quantities:

$$P_m^{\ell}(q) := \frac{\sqrt{D_m \sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4}}{\sum_{\lambda\in\Lambda_m} m_{2,\lambda}^2} \leq \frac{\|Y - s_m(X)\|_\infty^2}{\min_{\lambda\in\Lambda_m}\left\{\left(\sigma_\lambda^r\right)^2\right\}} \leq \frac{4A^2}{\sigma_{\min}^2}$$

$$Q_m^{(p)} := \frac{1}{D_m}\sum_{\lambda\in\Lambda_m}\left[(\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right] \geq \sigma_{\min}^2 > 0 \ .$$

For every $m \in \mathcal{M}_n$, define

$$A_n(m) = \min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \qquad \text{and} \qquad B_n(m) = \min_{\lambda\in\Lambda_m}\{np_\lambda\} \ .$$

We now define the event Ω_n on which,

- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$ and $A_n(m) \geq 1$:

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{(\mathbf{VF})}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{5.43}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{(\mathbf{VF})}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{5.44}$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{(\mathbf{VF})}\frac{\ln(n)}{\sqrt{D_m}}\mathbb{E}\left[p_2(m)\right] \tag{5.45}$$

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s, s_m)}{\sqrt{D_m}} + L_{(\mathbf{VF})}\frac{\ln(n)}{\sqrt{D_m}}\mathbb{E}\left[p_2(m)\right] \tag{5.47}$$

$$\left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right| \leq L_{(\mathbf{VF})}\frac{\ln(n)}{\sqrt{D_m}}\mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \tag{5.49}$$

- for every $m \in \mathcal{M}_n$ such that $B_n(m) > 0$ and $A_n(m) \geq 1$:

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n(m)^{-1}\ln(n)} - \frac{L_{(\mathbf{VF})}\ln(n)^2}{\sqrt{D_m}}\right)\mathbb{E}\left[\widetilde{p_2}(m)\right] \ . \tag{5.46}$$

- for every $m \in \mathcal{M}_n$:

$$A_n(m) \geq \frac{B_n(m)}{2} - 2(3 + \alpha_{\mathcal{M}})\ln(n) \tag{5.58}$$

The equations above have the same tags as in Sect. 5.7.4 where they are accurately stated and proven. From Prop. 5.8 (for $\widetilde{p_1}$ and p_2), Lemma 5.9 (for $\overline{\delta}(m)$), Prop. 5.10 (for pen), Lemma 5.12 (for $A_n(m)$), we have

$$\mathbb{P}\left(\Omega_n\right) \geq 1 - L\sum_{m\in\mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L(c_{\mathcal{M}})n^{-2} \ .$$

For every $m \in \mathcal{M}_n$ such that $\ln(n)^7 \leq D_m \leq L_{\alpha_{\mathcal{M}},c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}$, $(\mathbf{Ar}^{\mathbf{X}}_{\ell})$ implies that

$$B_n(m) \geq [L \vee (1 + 4(\alpha_{\mathcal{M}} + 3))] \ln(n) \ .$$

As a consequence, on Ω_n:

$$A_n(m) \geq \ln(n)$$

$$\max\left\{ \left|\widetilde{p_1}(m) - \mathbb{E}\left[\widetilde{p_1}(m)\right]\right|, \left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right|, \left|\overline{\delta}(m)\right|, \left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \right\} \leq \frac{L_{(\mathbf{VF})}\mathbb{E}\left[l(s,s_m) + p_2(m)\right]}{\ln(n)}$$

Using Prop. 5.2, (5.28) (in Lemma 5.6) and the fact that $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq \ln(n)$,

$$\mathbb{E}\left[\mathrm{pen}(m)\right] = \frac{C}{V-1}\left(2 + \widetilde{\delta_n}\right)\mathbb{E}\left[\widetilde{p_1}(m) + p_2(m)\right]$$

with $\left|\widetilde{\delta_n}\right| \leq L\ln(n)^{-1/4}$. We deduce: if $n \geq L_{(\mathbf{VF})}$, for every $m \in \mathcal{M}_n$ such that $\ln(n)^7 \leq D_m \leq L_{\alpha_{\mathcal{M}},c^X_{\mathrm{r},\ell}} n\ln(n)^{-1}$, on Ω_n,

$$\frac{-L_{(\mathbf{VF})}}{\ln(n)^{1/4}} p_1(m) \leq (\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}})(m) \leq \left[2(\eta-1) + \frac{L_{(\mathbf{VF})}}{\ln(n)^{1/4}}\right] p_1(m) \ .$$

We need to assume that n is large enough in order to upper bound $\mathbb{E}\left[p_2(m)\right]$ in terms of $p_1(m)$, since we only have

$$p_1(m) \geq \left[1 - \frac{L_{(\mathbf{VF})}}{\ln(n)^{1/4}}\right]_+ \mathbb{E}\left[p_2(m)\right]$$

in general.

Combined with (5.39), this gives: if $n \geq L_{(\mathbf{VF})}$,

$$\begin{aligned} l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\ln(n)^7 \leq D_{\widehat{m}} \leq L_{\alpha_{\mathcal{M}},c^X_{\mathrm{r},\ell}} n\ln(n)^{-1}} \leq \left[2\eta - 1 + \frac{L_{(\mathbf{VF})}}{\ln(n)^{1/4}}\right] \\ \times \inf_{m\in\mathcal{M}_n \text{ s.t. } \ln(n)^7 \leq D_m \leq L_{\alpha_{\mathcal{M}},c^X_{\mathrm{r},\ell}} n\ln(n)^{-1}} \left\{ l(s,\widehat{s}_m) \right\} \ . \end{aligned} \tag{5.40}$$

Define the oracle model $m^\star \in \arg\min\{l(s,\widehat{s}_m)\}$. We prove below that for any $c > 0$, if $n \geq L_{(\mathbf{VF}),c}$, then, on Ω_n:

$$\ln(n)^7 \leq D_{\widehat{m}} \leq cn\ln(n)^{-1} \tag{5.41}$$

$$\ln(n)^7 \leq D_{m^\star} \leq cn\ln(n)^{-1} \ . \tag{5.42}$$

The result follows since $L_{(\mathbf{VF})}\ln(n)^{-1/4} \leq \epsilon_n = \ln(n)^{-1/5}$ for $n \geq L_{(\mathbf{VF})}$. We finally remove the condition $n \geq n_0 = L_{(\mathbf{VF})}$ by choosing $K_1 = L_{(\mathbf{VF})}$ such that $K_1 n_0^{-2} \geq 1$.

Proof of (5.41). By definition, $\widehat{m}$ minimizes $\mathrm{crit}(m)$ over $\widehat{\mathcal{M}}_n$. It thus also minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = l(s,s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $\widehat{\mathcal{M}}_n$.

(1) Lower bound on $\mathrm{crit}'(m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^7$. We then have $\mathrm{pen}(m) \geq 0$,

$$l(s,s_m) \geq C_{\mathrm{b}}^- (\ln(n))^{-7\beta_1}$$

$$p_2(m) \leq L_{(\mathbf{VF})}\frac{\ln(n)D_m}{n} \leq L_{(\mathbf{VF})}\frac{(\ln(n))^8}{n}$$

and from (5.48) (in Lemma 5.9),

$$\overline{\delta}(m) \geq -L_A \sqrt{\frac{l(s,s_m)\ln(n)}{n}} + L_A \frac{\ln(n)}{n} \geq -L_A \sqrt{\frac{\ln(n)}{n}} \ .$$

As a consequence,

$$\mathrm{crit}'(m) \geq L_{(\mathbf{VF})} \left(\ln(n)\right)^{-L(\beta_1)} \ .$$

(2) Lower bound for large models: let $m \in \widehat{\mathcal{M}}_n$ such that $D_m > cn\left(\ln(n)\right)^{-1}$. Since $A_n(m) \geq 3$, Prop. 5.2 shows that

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m) - p_2(m)\right] \geq \mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \ .$$

Then, on Ω_n, (5.45), (5.49) and (5.47) imply

$$\begin{aligned} \mathrm{pen}(m) - p_2(m) &\geq \left(1 - L_{(\mathbf{VF}),c} n^{-1/4}\right) \mathbb{E}\left[p_2(m)\right] \\ &\geq L_{c,\sigma_{\min}} \ln(n)^{-1} \quad \text{when} \quad n \geq L_{(\mathbf{VF}),c} \\ \text{and} \qquad \overline{\delta}(m) &\geq -L_{(\mathbf{VF}),c} \sqrt{\frac{\ln(n)}{n}} \ , \end{aligned}$$

so that

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) + \overline{\delta}(m) - p_2(m) \geq L_{c,(\mathbf{VF})} \ln(n)^{-1}$$

when $n \geq L_{(\mathbf{VF}),c}$.

(3) There exists a better model for $\mathrm{crit}(m)$: according to (**P2**), there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If $n \geq L_{c_{\mathrm{rich}},c}$,

$$\ln(n)^7 \leq \sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n} \leq \frac{cn}{\ln(n)} \ .$$

Using (**Ap**),

$$l(s,s_{m_0}) \leq C_{\mathrm{b}}^+ n^{-\beta_2/2}$$

so that, when $n \geq L_{(\mathbf{VF})}$,

$$\begin{aligned} \mathrm{crit}'(m_0) &\leq l(s,s_{m_0}) + \left|\overline{\delta}(m)\right| + \mathrm{pen}(m) \\ &\leq L_{(\mathbf{VF})} \left(n^{-\beta_2/2} + n^{-1/2}\right) \ . \end{aligned}$$

If $n \geq L_{(\mathbf{VF}),c}$, this upper bound is smaller than the previous lower bounds for small and large models. □

Proof of (5.42). Recall that $m^\star$ minimizes $l(s,\widehat{s}_m) = l(s,s_m) + p_1(m)$ over $m \in \mathcal{M}_n$, with the convention $l(s,\widehat{s}_m) = \infty$ if $A_n(m) = 0$.

(1) Lower bound on $l(s,\widehat{s}_m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < \left(\ln(n)\right)^7$. From (**Ap**), we have

$$l(s,\widehat{s}_m) \geq l(s,s_m) \geq C_{\mathrm{b}}^- \left(\ln(n)\right)^{-7\beta_1} \ .$$

(2) Lower bound on $l(s,\widehat{s}_m)$ for large models: let $m \in \mathcal{M}_n$ such that $D_m > cn(\ln(n))^{-1}$ and $A_n(m) = \min_{\lambda\in\Lambda_m}\left\{n\widehat{p}_\lambda\right\} \geq 1$. From (5.46), for $n \geq L_{(\mathbf{VF}),c}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)\left(c_{\mathrm{r},\ell}^X\right)^{-1}\ln(n)} - L(A,\sigma_{\min},\alpha_{\mathcal{M}},c)n^{-1/4}\right) \mathbb{E}\left[\widetilde{p_2}(m)\right] \geq \frac{L_{(\mathbf{VF}),c}}{\ln(n)^2}$$

$$\text{so that} \quad l(s,\widehat{s}_m) \geq L_{(\mathbf{VF}),c} \ln(n)^{-2} \ .$$

(3) There exists a better model for $l(s,\widehat{s}_m)$: let $m_0 \in \mathcal{M}_n$ be as in the proof of (5.41) and assume $n \geq L_{c_{\text{rich}},c}$. Then,

$$p_1(m_0) \leq L_{(\mathbf{VF})}\mathbb{E}\left[p_2(m)\right] \leq L_{(\mathbf{VF})}n^{-1/2}$$

and the arguments of the previous proof show that

$$l(s,\widehat{s}_{m_0}) \leq L_{(\mathbf{VF})}\left(n^{-\beta_2/2}+n^{-1/2}\right)$$

which is smaller than the previous upper bounds for $n \geq L_{(\mathbf{VF}),c}$.

Classical oracle inequality. Let Ω_n be the event on which (5.11) holds true. Then,

$$\begin{aligned}\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right] &= \mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_n}\right] + \mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_n^c}\right] \\ &\leq \left[2\eta-1+\epsilon_n\right]\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\{l(s,\widehat{s}_m)\right\}\right] + A^2\mathbb{P}\left(\Omega_n^c\right) \\ &\leq \left[2\eta-1+\epsilon_n\right]\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\{l(s,\widehat{s}_m)\right\}\right] + \frac{A^2K_1}{n^2}\end{aligned}$$

which proves (5.12). □

5.7.4. Concentration results. In the proof of Thm. 5.1, we used the following concentration results. We always assume that S_m is the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$. We also use some additional notations: for every $q>0$,

$$P_m^\ell(q) := \frac{\sqrt{D_m\sum_{\lambda\in\Lambda_m}m_{q,\lambda}^4}}{\sum_{\lambda\in\Lambda_m}m_{2,\lambda}^2}$$

$$Q_m^{(p)} := \frac{n\mathbb{E}\left[\widetilde{p}_2(m)\right]}{D_m} = \frac{1}{D_m}\sum_{\lambda\in\Lambda_m}\left[\left(\sigma_\lambda^r\right)^2+\left(\sigma_\lambda^d\right)^2\right] \ .$$

Ideal penalties.

PROPOSITION 5.8. *Let $\gamma>0$. Assume that $\min_{\lambda\in\Lambda_m}\left\{np_\lambda\right\} \geq B_n \geq 1$ and*

$$\forall q\geq 2, \qquad P_m^\ell(q) \leq a_\ell q^{\xi_\ell} \ . \tag{$\mathbf{A_{m},\ell}$}$$

Then, on an event of probability at least $1-Ln^{-\gamma}$,

$$\widetilde{p}_1(m) \geq \mathbb{E}\left[\widetilde{p}_1(m)\right] - L(a_\ell,\xi_\ell,\gamma)\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}}+e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{5.43}$$

$$\widetilde{p}_1(m) \leq \mathbb{E}\left[\widetilde{p}_1(m)\right] + L(a_\ell,\xi_\ell,\gamma)\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}}+\sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{5.44}$$

$$\left|p_2(m)-\mathbb{E}[p_2(m)]\right| \leq L(a_\ell,\xi_\ell,\gamma)D_m^{-1/2}\ln(n)^{\xi_\ell+1}\mathbb{E}\left[p_2(m)\right] \ . \tag{5.45}$$

If we only have a lower bound $B_n>0$, then, with probability at least $1-Ln^{-\gamma}$,

$$\widetilde{p}_1(m) \geq \left(\frac{1}{2+(\gamma+1)B_n^{-1}\ln(n)} - \frac{L(a_\ell,\xi_\ell,\gamma)\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}}\right)\mathbb{E}\left[\widetilde{p}_2(m)\right] \ . \tag{5.46}$$

REMARK 5.4. We focus here on the histogram case, whereas it is possible to prove moment inequalities for p_2 in general in the bounded case (see the proof of Prop. 3.4 in Sect. 3.5.4). However, to our knowledge, the lower deviations of the excess risk $p_1(m)$ have never been controlled non-asymptotically.

It now remains to control $\overline{\delta}(m) = (P-P_n)\left(\gamma(s_m)-\gamma(s)\right)$. Since the data is bounded, Bernstein inequality gives the following:

LEMMA 5.9 (Prop. 3.3, Sect. 3.3). *Assume that* $\|Y\|_\infty \leq A < \infty$. *Then for all* $x \geq 0$, *on an event of probability at least* $1 - 2e^{-x}$:

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s,s_m)}{\sqrt{D_m}} + \frac{20}{3}\frac{A^2}{Q_m^{(p)}}\frac{\mathbb{E}^{\Lambda_m}\left[\widetilde{p_2}(m)\right]}{\sqrt{D_m}}x \tag{5.47}$$

$$\text{and} \quad \forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta l(s,s_m) + \left(\frac{4}{\eta} + \frac{8}{3}\right)\frac{A^2 x}{n} \ . \tag{5.48}$$

V-fold penalties.

PROPOSITION 5.10. *Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Let* $\mathrm{pen}(m)$ *be defined by* (5.8) *with the weights W defined in algorithm 5.2. Let $\gamma > 0$ and assume that*

$$\forall q \geq 2, \qquad P_m^\ell(q) \leq a_\ell q^{\xi_\ell} \ . \tag{$\mathbf{A_{m},\ell}$}$$

Then, there exists an event of probability at least $1 - n^{-\gamma}$ such that

$$\begin{aligned} &\left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \mathbb{1}_{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq 1} \\ &\leq C\left(\frac{1}{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}} \vee \frac{1}{V}\right) L(a_\ell,\xi_\ell,\gamma) D_m^{-1/2}\ln(n)^{\xi_\ell+1}\mathbb{E}\left[p_2(m)\right] \ . \end{aligned} \tag{5.49}$$

Proofs.

PROOF OF PROP. 5.8. According to the explicit expressions (5.19) and (5.26), $\widetilde{p_1}(m)$ and $p_2(m)$ are both U-statistics of order 2 conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda)}$. Then, we use Prop. 5.5, with $\xi_{i,\lambda} = Y_i - \beta_\lambda$, $a_\lambda = 0$ and

$$b_\lambda = \frac{p_\lambda}{n^2\widehat{p}_\lambda^2} \quad \text{for } \widetilde{p_1} \quad \text{and} \quad b_\lambda = \frac{1}{n^2\widehat{p}_\lambda} \quad \text{for } p_2 \ .$$

This proves, for all $q \geq 2$,

$$\left\|\widetilde{p_1}(m) - \mathbb{E}^{\Lambda_m}[\widetilde{p_1}(m)]\right\|_q^{(\Lambda_m)} \leq \max_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\mathbb{1}_{\widehat{p}_\lambda>0}\right\} L(a_\ell,\xi_\ell) D_m^{-1/2} q^{\xi_\ell+1}\mathbb{E}\left[p_2(m)\right] \tag{5.50}$$

$$\left\|p_2(m) - \mathbb{E}[p_2(m)]\right\|_q^{(\Lambda_m)} \leq L(a_\ell,\xi_\ell) D_m^{-1/2} q^{\xi_\ell+1}\mathbb{E}\left[p_2(m)\right] \ . \tag{5.51}$$

We deduce conditional concentration inequalities with Lemma 8.10 (Sect. 8.6.2), taking $x = \gamma\ln(n)$. Since x is deterministic, this implies unconditional concentration inequalities. We then use Lemma 5.11 to replace $\mathbb{E}^{\Lambda_m}\left[p_1(m)\right]$ by $\mathbb{E}\left[\widetilde{p_1}(m)\right]$. Finally, we use the rough inequality below and the first result follows.

$$\mathbb{P}\left(\max_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\mathbb{1}_{\widehat{p}_\lambda>0}\right\} \leq L\times(\gamma+1)\ln(n)\right) \geq 1 - n^{-\gamma} \ . \tag{5.52}$$

In order to prove (5.46), we start from (5.50) combined with Lemma 8.10 with $x = \gamma\ln(n)$. Then, instead of using Lemma 5.11, we remark that

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}(m)\right] \geq \min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\right\}\mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \ ,$$

and the result follows since

$$\mathbb{P}\left(\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\right\} \geq \frac{1}{2+(\gamma+1)B_n^{-1}\ln(n)}\right) \geq 1 - n^{-\gamma} \ . \tag{5.53}$$

□

PROOF OF (5.52). For every $\lambda \in \Lambda_m$, Bernstein inequality gives

$$\forall \kappa > 0, \qquad \mathbb{P}\left(\frac{p_\lambda}{\widehat{p}_\lambda} \leq \frac{1}{\left(1 - \sqrt{2\kappa} - \frac{\kappa}{3}\right)_+} \right) \geq 1 - e^{-\kappa n p_\lambda} \ .$$

If $np_\lambda \geq 8(\gamma+1)\ln(n)$, take $\kappa = 1/8$ in this inequality. Otherwise, since $n\widehat{p}_\lambda \geq 1$, we have the deterministic upperbound $\mathbb{1}_{\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda > 0} p_\lambda \widehat{p}_\lambda^{-1} \leq np_\lambda \leq 8(\gamma+1)\ln(n)$. A simple union bound gives (5.52) since $D_m \leq n$. □

PROOF OF (5.53). For every $\lambda \in \Lambda_m$, Bernstein inequality gives

$$\forall \kappa > 0, \qquad \mathbb{P}\left(\frac{p_\lambda}{\widehat{p}_\lambda} \geq \frac{1}{1 + \sqrt{2\kappa} + \frac{\kappa}{3}} \right) \geq 1 - e^{-\kappa n p_\lambda} \ .$$

With $\kappa = B_n^{-1}(\gamma+1)\ln(n)$ and the union bound, the result follows. □

PROOF OF PROP. 5.10. By definition (5.8), $\mathrm{pen}(m) = \mathbb{E}_W[Z]$ with

$$\begin{aligned} Z &= \sum_{\lambda\in\Lambda_m} \left(\widehat{p}_\lambda + \widehat{p}_\lambda^W\right)\left(\widehat{\beta}_\lambda - \widehat{\beta}_\lambda^W\right)^2 \\ &= \sum_{\lambda\in\Lambda_m} \widehat{p}_\lambda (1+W_\lambda) \left[\frac{\sum_{X_i\in I_\lambda} \left(1 - \frac{W_i}{W_\lambda}\right)(Y_i - \beta_\lambda)}{n\widehat{p}_\lambda} \right]^2 \\ &= \sum_{\lambda\in\Lambda_m} \frac{1+W_\lambda}{n^2 \widehat{p}_\lambda W_\lambda^2} \left[\sum_{X_i\in I_\lambda} (W_\lambda - W_i)(Y_i - \beta_\lambda) \right]^2 \ . \end{aligned} \tag{5.54}$$

By Jensen inequality, for every $q \geq 1$,

$$\begin{aligned} \left\|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right\|_q^{(\Lambda_m)} &\leq \left\|Z - \mathbb{E}^{\Lambda_m}[Z \mid W]\right\|_q^{(\Lambda_m)} \\ &= \left(\mathbb{E}_W\left[\mathbb{E}^{\Lambda_m}\left[\left|Z - \mathbb{E}^{\Lambda_m}[Z \mid W]\right|^q \,\middle|\, W\right]\right]\right)^{1/q} \\ &\leq \sup_{W_0\in\mathrm{supp}(W)} \left\{ \mathbb{E}^{\Lambda_m}\left[\left|Z - \mathbb{E}^{\Lambda_m}[Z \mid W = W_0]\right|^q \,\middle|\, W = W_0\right] \right\}^{1/q} \\ &= \sup_{W_0\in\mathrm{supp}(W)} \left\{ \left\|Z - \mathbb{E}^{\Lambda_m}[Z \mid W = W_0]\right\|_q^{(W_0,\Lambda_m)} \right\} \end{aligned} \tag{5.55}$$

where $\mathrm{supp}(W)$ is the support of the resampling weight vector W and $\|\cdot\|_q^{(W_0,\Lambda_m)}$ denotes the q-th moment conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda)}$ and $W = W_0$.

Then, we can assume that $W \in \mathbb{R}^n$ is deterministic, and try to control $\left\|Z - \mathbb{E}^{\Lambda_m}[Z]\right\|_q^{(W,\Lambda_m)}$. According to (5.54), this can be done by Prop. 5.5. We denote by $X_{(1,\lambda)}, \ldots, X_{(n\widehat{p}_\lambda,\lambda)}$ the data such that $X_i \in I_\lambda$, and take

$$\xi_{i,\lambda} = \left(W_{(i,\lambda)} - W_\lambda\right)\left(Y_{(i,\lambda)} - \beta_\lambda\right) \qquad \text{so that} \quad m_{q,\lambda,i} = \left|W_\lambda - W_{(i,\lambda)}\right| m_{q,\lambda}$$

$$r_\lambda = n\widehat{p}_\lambda \qquad a_\lambda = 0 \qquad b_\lambda = \frac{1+W_\lambda}{n^2\widehat{p}_\lambda W_\lambda^2} \ .$$

We obtain, for every $W_0 \in \mathbb{R}^n$,

$$\left\|Z - \mathbb{E}^{\Lambda_m}[Z]\right\|_q^{(W,\Lambda_m)} \leq L\sqrt{q}\sqrt{\sum_{\lambda\in\Lambda_m} b_\lambda^2 m_{2q,\lambda}^4 \sum_{i=1}^{r_\lambda}\left(W_{(i,\lambda)} - W_\lambda\right)^4} + Lq\sqrt{\sum_{\lambda\in\Lambda_m} b_\lambda^2 m_{2q,\lambda}^4 \left(\sum_{i=1}^{r_\lambda}\left(W_{(i,\lambda)} - W_\lambda\right)^2\right)^2} .$$

Now fix some $\lambda \in \Lambda_m$ and write $n\widehat{p}_\lambda = aV + b$ with $a, b \in \mathbb{N}$ and $0 \leq b \leq V-1$. For any realization of W, there is some $\epsilon \in \{0,1\}$ such that

$$\{W_i \text{ s.t. } X_i \in I_\lambda\} = \left\{0 \text{ repeated } a+\epsilon \text{ times}, \frac{V}{V-1} \text{ repeated } r_\lambda - a - \epsilon \text{ times}\right\}$$

$$W_\lambda = \frac{V}{V-1} \times \frac{r_\lambda - a - \epsilon}{r_\lambda} = 1 + \frac{b - V\epsilon}{(V-1)(aV+b)}$$

so that, if $r_\lambda \geq 1$,

$$\begin{aligned}\sum_{i=1}^{r_\lambda}\left(W_{(i,\lambda)} - W_\lambda\right)^2 &= (a+\epsilon)W_\lambda^2 + (r_\lambda - a - \epsilon)\left(W_\lambda - \frac{V}{V-1}\right)^2 \\ &\leq (a+1)\left(1+\frac{1}{r_\lambda}\right)^2 + (r_\lambda - a)\left(\frac{1}{V-1} + \frac{1}{r_\lambda}\right)^2 \\ &\leq L \times \left[1 \vee \frac{r_\lambda}{V}\right] .\end{aligned}$$

Then, we have for every $q \geq 2$,

$$\begin{aligned}\left\|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right\|_q^{(\Lambda_m)} &\leq Lq\sqrt{\sum_{\lambda\in\Lambda_m}\left(\frac{1+W_\lambda}{n^2\widehat{p}_\lambda W_\lambda^2}\right)^2 m_{2q,\lambda}^4 \left[1 \vee \frac{(n\widehat{p}_\lambda)^2}{V^2}\right]} \\ &\leq La_\ell q^{\xi_\ell+1}\left[\frac{1}{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}} \vee \frac{1}{V}\right]\mathbb{E}^{\Lambda_m}[p_2(m)] .\end{aligned}$$

The result follows with the classical link between moment and concentration inequalities (*e.g.* Lemma 8.10 in Sect. 8.6.2). □

5.7.5. Technical lemmas.

For concentration results. We need the following technical lemma, in order to prove that $\mathbb{E}^{\Lambda_m}[p_1(m)]$ is close to $\mathbb{E}[p_1(m)]$ with several conventions for p_1.

LEMMA 5.11. *We assume that* $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n \geq 1$.

(1) *Lower deviations: let* $c_1 = 0.184$. *For all* $x \geq 0$, *with probability at least* $1 - e^{-x}$,

$$\mathbb{E}^{\Lambda_m}[\widetilde{p_1}(m)] \geq \mathbb{E}[\widetilde{p_1}(m)] - \theta^-(x, B_n, D_m, P_m^\ell(2)) \times \mathbb{E}[p_2(m)] \tag{5.56}$$

$$\text{with} \qquad \theta^- := L\left[\varphi_1(c_1 B_n) + P_m^\ell(2)\sqrt{e^{-c_1 B_n} + \frac{x}{D_m}}\right]$$

(2) *Upper deviations: let* $c_2 = 0.28$ *and* $c_4 = 0.09$. *For all* $T \in (0;1]$ *and* $x \geq 0$, *with probability at least* $1 - e^{-x}$,

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}^{(T)}(m)\right] \leq \mathbb{E}\left[\widetilde{p_1}^{(T)}(m)\right] + \theta^+(x, D_m, B_n, T)\mathbb{E}[p_2(m)] \tag{5.57}$$

$$\text{with} \qquad \theta^+ := L\left[\varphi_1(c_2 B_n) + P_m^\ell(2)\sqrt{xD_m^{-1} + e^{-c_4 B_n}}\left(1 \vee T\sqrt{x + D_m e^{-c_4 B_n}}\right)\right] .$$

PROOF. From (5.22) and (5.23), we have explicit expressions for $\widetilde{p_1}$ and $\widetilde{p_1}^{(T)}$. We then apply Lemma 5.4 in Sect. 5.6.2, with $a_\lambda = p_\lambda (\sigma_\lambda)^2 \geq 0$. For θ^+, we used the general upper bound

$$\max_{\lambda\in\Lambda_m} (\sigma_\lambda)^4 \left(\sum_{\lambda\in\Lambda_m} \sigma_\lambda^4 \right)^{-1} \leq 1 \ .$$

□

REMARK 5.5. If $B_n \geq \left(c_1^{-1} \vee c_4^{-1}\right) \ln(n)$, for every $\gamma > 0$ and $T \in (0;1)$,

$$\theta^- \vee \theta^+ \left(\gamma \ln(n), B_n, D_m, P_m^\ell(2) \right) \leq L_\gamma P_m^\ell(2) D_m^{-1/2} \ln(n)$$

since $D_m \leq n$.

Empirical and expected frequencies. Finally, we have to control empirical frequencies $n\widehat{p}_\lambda$ in the proof of Thm. 5.1.

LEMMA 5.12. *Let $(p_\lambda)_{\lambda\in\Lambda_m}$ be non-negative real numbers of sum 1, $(n\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$ a multinomial vector of parameters $(n;(p_\lambda)_{\lambda\in\Lambda_m})$. Then, for all $\gamma > 0$,*

$$\min_{\lambda\in\Lambda_m} \{ n\widehat{p}_\lambda \} \geq \frac{\min_{\lambda\in\Lambda_m} \{ np_\lambda \}}{2} - 2(\gamma+1)\ln(n) \tag{5.58}$$

with probability at least $1 - 2n^{-\gamma}$.

PROOF OF LEMMA 5.12. By Bernstein inequality ([**Mas07**], Prop. 2.9), for all $\lambda \in \Lambda_m$,

$$\mathbb{P}\left(n\widehat{p}_\lambda \geq (1-\theta)np_\lambda - \sqrt{2npx} - \frac{x}{3} \right) \geq 1 - e^{-x} \ .$$

Take $x = (\gamma+1)\ln(n)$ above, and remark that $\sqrt{2npx} \leq \frac{np}{2} + x$. The union bound gives the result since $\mathrm{Card}(\Lambda_m) \leq n$. □

5.7.6. Expectation of inverses of binomials. In this section, we prove Lemma 5.3. Let $Z \sim \mathcal{B}(n,p)$. We have

$$\mathbb{P}(Z > 0) = 1 - (1-p)^n \geq 1 - e^{-np}$$

so that the lower bound comes from (5.13). We introduce another interesting quantity closely related to e_Z^+:

$$e^0_{\mathcal{L}(Z)} := \mathbb{E}[Z]\,\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z>0}\right] = e_Z^+ \mathbb{P}(Z>0) \ , \tag{5.59}$$

so that we only have to give an upper bound on $e^0_{\mathcal{B}(n,p)}$.

Firstly, using Lemma 4.1 in [**GKKW02**], we have for every $n \in \mathbb{N}$ and $p \in [0;1]$

$$e^0_{\mathcal{B}(n,p)} \leq \frac{2np}{(n+1)p} \leq 2 \ . \tag{5.60}$$

The upper-bound by κ_4 follows, since $\kappa_4 \geq 2(1-e^{-1})^{-1}$.

Secondly, using that $\mathbb{P}(1 > Z > 0) = 0$, we have

$$\begin{aligned} \forall \alpha > 0, \quad e^0_{\mathcal{B}(n,p)} &= np\mathbb{E}\left[Z^{-1}\mathbb{1}_{\alpha\mathbb{E}[Z]>Z>0}\right] + np\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z\geq\alpha\mathbb{E}[Z]}\right] \\ &\leq np\mathbb{P}\left(\alpha np > Z > 0\right) + \alpha^{-1} \ . \end{aligned}$$

With Bernstein inequality ([**Mas07**], Prop. 2.9):

$$\forall \theta > 0, \quad \mathbb{P}\left(Z \leq \left(1 - \sqrt{2\theta} - \frac{\theta}{3}\right) np \right) \leq e^{-\theta np} \ ,$$

we obtain, for every $0 < \theta \leq h_+ = \frac{3\left(\sqrt{5}-\sqrt{3}\right)^2}{2}$,

$$e^0_{\mathcal{B}(n,p)} \leq \frac{1}{1-\sqrt{2\theta}-\frac{\theta}{3}} + \frac{1}{\theta}\left(\theta np e^{-\theta np}\right) \ .$$

As $\varphi : x \mapsto xe^{-x}$ is maximal on $\mathbb{R}$ at $x = 1$, and decreases on $[1, \infty)$, we have for every $A \geq 0$,

$$\sup_{np\geq A,\, n\in\mathbb{N},\, p\in[0;1]} \left\{ e^0_{\mathcal{B}(n,p)} \right\} \leq \inf_{0<\theta\leq h_+} \left\{ \frac{1}{1-\sqrt{2\theta}-\frac{\theta}{3}} + \frac{\varphi((\theta A)\vee 1)}{\theta} \right\} \ . \tag{5.61}$$

For every $A \geq A_0 > h_+^{-1/2}$, taking $\theta = A^{-1/2} < h_+$ in (5.61) leads to

$$\sup_{np\geq A} \{e^+_{\mathcal{B}(n,p)}\} \leq \left[\frac{1}{1-\sqrt{2}A^{-1/4}-\frac{1}{3}A^{-1/2}} + Ae^{-\sqrt{A}} \right] \frac{1}{1-e^{-A}} \ .$$

One can prove that if $A_0 = 29.17$, this upper bound is smaller than $1 + \kappa_3 A^{-1/4}$ with $\kappa_3 = 5.03$. When $A < A_0$, notice that $1 + \kappa_3 A^{-1/4} \geq \kappa_4$, so the upper bound in (5.14) still holds.

REMARK 5.6. (1) Taking $\theta = 3\ln(A)/A$ in (5.61) leads to an upper bound

$$1 + \kappa_5 \sqrt{\frac{\ln(A)}{A}} \geq \sup_{np\geq A} \left\{ e^+_{\mathcal{B}(n,p)} \right\}$$

for some numerical constant κ_5.

(2) We can also take $\theta = 0.16$ in (5.61) and obtain an absolute upper bound $\kappa_4 = 7.8$. Thus, the proof only needs $\mathbb{P}(0 < Z < c_Z) = 0$ for some $c_Z > 0$ and that Z satisfies Bernstein inequality or a similar concentration inequality. Such a result can be obtained for a quite large class of random variables Z.

CHAPTER 6

Resampling penalties

RÉSUMÉ. Nous étudions dans ce chapitre une nouvelle famille de pénalités par rééchantillonnage (RP), généralisant les pénalités bootstrap proposées par Efron [**Efr83**]. Celles-ci sont construites comme les pénalités V-fold définies au Chap. 5. Dans le cadre de la régression sur des modèles d'histogrammes, nous prouvons une inégalité-oracle non-asymptotique trajectorielle, avec une constante presque 1. On peut en particulier en déduire un résultat d'adaptation à la régularite hölder de la fonction de régression, en présence d'un bruit hétéroscedastique assez général. Ces résultats apportent également un nouvel éclairage aux résultats asymptotiques de Shibata [**Shi97**] sur les pénalités bootstrap dans un autre cadre, ainsi que sur la consistance de pénalités «m out of n» lorsque $m \ll n$ (Shao [**Sha96**]). De plus, leurs preuves reposent sur de nouveaux résultats non-asymptotiques nouveaux sur le rééchantillonnage à poids échangeable en général. Une étude de simulation illustre les bonnes performances de ces pénalités, notamment dans un cadre hétéroscédastique. Elle indique également un léger avantage à utiliser des poids échangeables «random hold-out» plutôt que des poids V-fold (étudiés au Chap. 5) ou bootstrap.

6.1. Introduction

Penalization is a classical tool in model selection theory. Basically, it states that a good choice between several algorithms can be made by minimizing the sum of the empirical risk (how do algorithms fit the data) and some complexity measure of the algorithms (called the penalty). The ideal penalty for prediction is of course the difference between the true and empirical risks of the output, but it is unknown in general. It is thus crucial to obtain tight estimates of such a quantity.

Many penalties or complexity measures have been proposed, both in the classification and regression frameworks. Consider for instance regression and least-square estimators on finite-dimensional vector spaces (the models). When the design is fixed and the noise-level constant equal to σ, Mallows' C_p penalty [**Mal73**] (equal to $2n^{-1}\sigma^2 D$ for a D-dimensional space, and it can be modified according to the number of models [**BM01, Sau06**]) has some optimality properties [**Shi81, Li87, Bar02**]. However, such a penalty linear in the dimension may be terrible in an heteroscedastic framework (as shown by (6.6) and Sect. 6.6.2).

In classification, the VC-dimension has the drawback of being independent of the underlying measure, so that it is adapted to the worst case. It has been improved with data-dependent complexity estimates, such as Rademacher complexities [**Kol01, BBL02**] (generalized by Fromont with resampling ideas [**Fro04**]), but they may be too large because they are still global complexity measures. The localization idea then led to local Rademacher complexities [**LW04, BBM05, Kol06**] which are tight estimates of the ideal penalty, but involve unknown (or much too large)

constants and may be very difficult to compute in practice. On the other hand, the V-fold cross-validation (VFCV) is very popular for such purposes, but it is still poorly understood from the non-asymptotic viewpoint. Some results about VFCV and hold-out, mainly asymptotic, are given by Györfi *et al.* [**GKKW02**] in the regression case. More general cross-validation schemes are studied by van der Laan, Dudoit and Keles [**vdLDK04**] in the density estimation framework. For references about cross-validation in regression or density estimation, see Chap. 5 and the references therein (in particular [**vdLDK04, Yan06, CR06**]). For the classification case, see the recent results of Yang [**Yan07**] and references therein.

In this chapter, we propose a new family of model selection algorithms by penalization, called Resampling Penalization (RP). RP is based on Efron's bootstrap heuristics [**Efr79**] (generalized to exchangeable weighted bootstrap, *i.e.* resampling, by Mason and Newton [**MN92**] and Præstgaard and Wellner [**PW93**]) and generalizes Efron's bootstrap penalty [**Efr83**] (which is quite similar to EIC proposed by Ishiguro, Sakamoto and Kitagawa [**ISK97**] and studied by Shibata [**Shi97**]). It is a localized version of Fromont's penalties [**Fro04**], which does not involve any unknown constant, and is easy to compute in its leave-one-out version. For similar algorithms with a smaller computational complexity, we refer to Chap. 5, where we defined a V-fold version of RP. There are many resampling-based model selection algorithms in the literature, for prediction error (or variance) estimation (Wu [**Wu86**], Efron and Tibshirani [**ET97**], Molinaro, Simon and Pfeiffer [**MSP05**]), model identification (Shao [**Sha96**]) or correction of AIC (Shibata [**Shi97**] and references therein) to name but a few. Though, to our knowledge, RP has never been proposed with such a generality, neither in theoretical studies nor in practical situations. We also want to emphasize that RP is defined in a much general framework, so that it may have a wide range of practical applications.

As a first theoretical step, we prove the efficiency of these algorithms in the case example of least-square regression on histograms. In this framework, the i.i.d. data $(X_i, Y_i)_{1 \leq i \leq n}$ can be written

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i$$

where s is the regression function, $\sigma(\cdot)$ the *heteroscedastic* noise-level and ϵ_i a noise term with unit variance. Assuming that the models $(S_m)_{m \in \mathcal{M}}$ are histogram models, RP satisfies oracle inequalities with constant almost one and is asymptotically optimal (Thm. 6.1), under several reasonable sets of assumptions. Moreover, we prove some kind of adaptivity to the regularity of the regression function (Thm 6.2), even when the noise is highly heteroscedastic. These results come from explicit computations that allow us to deeply understand why these penalties are working well. A major advance of this chapter is its non-asymptotic theoretical approach, which is unusual in the resampling literature. Although our proofs are restricted to a particular case, we believe that RP has a similar behaviour in a far more general framework. We explain why in Sect. 6.6 and Chap. 7. So, our extensive study of the toy model of histograms is made to derive heuristics for the general case. Our main goal here is to help practical users.

As already noticed, several similar results for other algorithms exist in the literature, for instance for Mallows' (for homoscedastic noise only) and classical cross-validation algorithm (*i.e.* leave-one-out). The interest of RP is both its generality (contrary to Mallows' C_p) and its flexibility.

We conduct an extensive simulation experiment (Sect. 6.5) with small sample sizes. RP is shown to be competitive with Mallows' C_p for "easy" problems, and much better for some harder ones (*e.g.* with a variable noise-level). On the other hand, a well-calibrated RP has almost always better performances than classical VFCV. Thus, RP may be of great interest in situations where

no *a priori* information is known about the data. It is an efficient alternative to VFCV, which is able to deal with difficult problems, while being close to the best procedures that are fitted for easier problems.

This chapter is organized as follows. The general Resampling Penalization algorithm (RP) is defined in Sect. 6.2. We focus on the histogram regression case in Sect. 6.3, for which we state some theorems in Sect. 6.4. We then present an extensive simulation experiment in Sect. 6.5. A discussion about practical implementation of RP and a comparison with other model selection procedures is made in Sect. 6.6. The remaining sections are devoted to probabilistic tools (Sect. 6.7) and proofs (Sect. 6.8).

6.2. A general model selection algorithm

We consider the following general setting: $\mathcal{X} \times \mathcal{Y}$ is a measurable space, P an unknown probability measure on it and $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ some data of common law P. Let $\mathcal{S}$ be the set of predictors (measurable functions $\mathcal{X} \mapsto \mathcal{Y}$) and $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$ a contrast function. Given a family $(\widehat{s}_m)_{m \in \mathcal{M}_n}$ of data-dependent predictors, our goal is to find the one minimizing the prediction loss $P\gamma(t)$. We will extensively use this functional notation $Q\gamma(t) := \mathbb{E}_{(X,Y)\sim Q}[\gamma(t,(X,Y))]$, for any probability measure Q on $\mathcal{X} \times \mathcal{Y}$. Notice that the expectation here is only taken w.r.t. (X, Y), so that $Q\gamma(t)$ is random when $t = \widehat{s}_m$ is random. Assuming that there exists a minimizer $s \in \mathcal{S}$ of the loss (the Bayes predictor), we will often consider the excess loss $l(s,t) = P\gamma(t) - P\gamma(s) \geq 0$ instead of the loss.

Assume that each predictor $\widehat{s}_m$ may be written as a function $\widehat{s}_m(P_n)$ of the empirical distribution of the data $P_n = n^{-1} \sum_{i=1}^n \delta_{(X_i,Y_i)}$. The ideal choice for $\widehat{m}$ is the one which minimizes over $\mathcal{M}_n$ the true prediction risk $P\gamma(\widehat{s}_m(P_n)) = P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}_{\mathrm{id}}(m)$ where the ideal penalty is equal to

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m(P_n)) \ . \tag{6.1}$$

The *resampling heuristics* (introduced by Efron [**Efr79**]) states that the expectation of any functional $F(P, P_n)$ is close to its resampling counterpart $\mathbb{E}_W F(P_n, P_n^W)$, where

$$P_n^W = \frac{1}{n} \sum_{i=1}^n W_i \delta_{(X_i,Y_i)}$$

is the empirical distribution P_n weighted by an independent random vector $W \in [0; +\infty)^n$, with $\sum_i \mathbb{E}[W_i] = n$. The expectation $\mathbb{E}_W[\cdot]$ means that we only integrate w.r.t. the weights W. We suggest here to use this heuristics for estimating $\mathrm{pen}_{\mathrm{id}}(m)$, and plug it into the penalized criterion $P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m)$. This defines $\widehat{m} \in \mathcal{M}_n$ as follows.

Algorithm 6.1 (Resampling penalization).

(1) Choose a resampling scheme, *i.e.* the law $\mathcal{L}(W)$ of a weight vector W.

(2) Choose a constant $C \geq C_{W,\infty} \approx \left(n^{-1} \sum_{i=1}^n \mathbb{E}\left(W_i - 1\right)^2 \right)^{-1}$.

(3) Compute the following resampling penalty for each $m \in \mathcal{M}_n$:

$$\mathrm{pen}(m) = C\mathbb{E}_W \left[P_n\gamma\left(\widehat{s}_m\left(P_n^W\right)\right) - P_n^W\gamma\left(\widehat{s}_m\left(P_n^W\right)\right) \right] \ . \tag{6.2}$$

(4) Minimize the penalized empirical criterion to choose $\widehat{m}$ and thus $\widehat{s}_{\widehat{m}}$:

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}(m) \right\} \ .$$

REMARK 6.1. (1) Applying the resampling heuristics as we do in Sect. 6.2, with the bootstrap resampling scheme, leads to Efron's bootstrap penalty [**Efr83**]. In the log-likelihood framework, this is also called the EIC procedure by Ishiguro, Sakamoto and Kitagawa [**ISK97**].

(2) There is a constant $C \neq 1$ in front of the penalty, although there isn't any in Efron's heuristics, because we did not normalize W. The asymptotical value of the right normalizing constant $C_{W,\infty}$ may be derived from Thm. 3.6.13 of van der Vaart and Wellner [**vdVW96**]. In the case example of histograms, we give a non-asymptotic expression for it in Tab. 6.1. In general, we suggest to use some data-driven method to choose C, whereas the resampling penalty only estimates the shape of the ideal one (see algorithm 11.1).

(3) We allowed C to be larger than $C_{W,\infty}$ because overpenalizing may be fruitful in a non-asymptotic viewpoint, *e.g.* when there is few noisy data. The simulation study of Sect. 6.5 provides experimental evidence for this fact. See also Sect. 11.3.3.

(4) Since $\widehat{m}$ is computed through a plug-in method, algorithm 6.1 seems to be reasonable only if $\mathcal{M}_n$ is not too large. Otherwise, we can for instance group the models of similar complexities and reduce $\mathcal{M}_n$ to a polynomial family.

6.3. The histogram regression case

As studying algorithm 6.1 in general is a rather difficult question, we focus in this chapter on the case example of least-square regression on histograms. Although we do not consider histograms as a final goal, this first theoretical step is useful to derive heuristics making the general algorithm 6.1 work.

6.3.1. A modified algorithm for histograms. We first precise the framework and some notations. The data $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ are i.i.d. of common law P. Denoting by s the regression function, we have

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{6.3}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms, possibly dependent from X_i, but with variance 1 conditionally to X_i. Throughout this chapter, we always assume that there is some noise:

$$\|\sigma\|_2^2 = \|\sigma(X)\|_2^2 = \mathbb{E}\left[\sigma(X)^2\right] = \mathbb{E}\left[\epsilon^2\right] > 0 \ .$$

The feature space $\mathcal{X}$ is typically a compact subset of $\mathbb{R}^d$. We use the least-square contrast $\gamma : (t, (x, y)) \mapsto (t(x) - y)^2$ to measure the quality of a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$. As a consequence, the Bayes predictor is the regression function s, and the excess loss is $l(s, t) = \mathbb{E}_{(X,Y)\sim P}\left(t(X) - s(X)\right)^2$. To each model S_m, we associate the empirical risk minimizer $\widehat{s}_m = \widehat{s}_m(P_n) = \arg\min_{t \in S_m}\{P_n\gamma(t)\}$ (when it exists and is unique).

Each model in $(S_m)_{m\in\mathcal{M}_n}$ is the set of piecewise constant functions (histograms) on some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$. It is thus a vector space of dimension $D_m = \mathrm{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda\in\Lambda_m}$. As this basis is orthogonal in $L^2(\mu)$ for any probability measure μ on $\mathcal{X}$, we can make explicit computations in order to understand algorithm 6.1. The following notations will be useful throughout this chapter.

$$p_\lambda := P(X \in I_\lambda) \qquad \widehat{p}_\lambda := P_n(X \in I_\lambda) \qquad \widehat{p}_\lambda^W = \widehat{p}_\lambda W_\lambda := P_n^W(X \in I_\lambda)$$

$$s_m := \arg\min_{t\in S_m} P\gamma(t) = \sum_{\lambda\in\Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \qquad \beta_\lambda = \mathbb{E}_P\left[Y \mid X \in I_\lambda\right]$$

$$\widehat{s}_m := \arg\min_{t\in S_m} P_n\gamma(t) = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{n\widehat{p}_\lambda} \sum_{X_i\in I_\lambda} Y_i$$

$$\widehat{s}_m^W := \arg\min_{t\in S_m} P_n^W\gamma(t) = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda^W \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda^W = \frac{1}{n\widehat{p}_\lambda^W} \sum_{X_i\in I_\lambda} W_i Y_i$$

Remark that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i, and the same problem arises for $\widehat{s}_m^W$. This is why we will slightly modify the general algorithm for histograms.

Assuming that $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda > 0$ (otherwise, the model m should clearly not be chosen), we can compute the ideal penalty (see (5.19) and (5.26) in Sect. 5.7.2):

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m) = \sum_{\lambda\in\Lambda_m} (p_\lambda + \widehat{p}_\lambda)\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2 + (P - P_n)\gamma(s_m) \ .$$

According to the resampling heuristics, $\mathrm{pen}_{\mathrm{id}}(m)$ can be estimated (up to some normalizing constant) by

$$\begin{aligned}
\mathrm{pen}(m) &= \mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m^W)\right] \\
&= \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\left(\widehat{p}_\lambda + \widehat{p}_\lambda^W\right)\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] + \mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m)\right] \\
&= \sum_{\lambda\in\Lambda_m} \left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right]\right) \qquad (6.4)
\end{aligned}$$

since $\sum_i \mathbb{E}[W_i] = 1$ ($\mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m)\right]$ indeed estimates the expectation of $(P - P_n)\gamma(s_m)$ which is centered). Uniqueness issues[1] for $\widehat{s}_m^W$ make the first term badly defined: with positive probability, we have $\widehat{p}_\lambda^W = 0$ and $\widehat{\beta}_\lambda^W$ undefined, even if $\widehat{p}_\lambda > 0$ and $\widehat{\beta}_\lambda$ is well-defined. This is why we modified step 3 in algorithm 6.1, which leads to the following.

ALGORITHM 6.2 (Resampling penalization for histograms).

0. Replace $\mathcal{M}_n$ by
$$\widehat{\mathcal{M}}_n = \left\{m \in \mathcal{M}_n \text{ s.t. } \min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq 3\right\} \ .$$
1. Choose a resampling scheme $\mathcal{L}(W)$.
2. Choose a constant $C \geq C_{W,\infty}$ where $C_{W,\infty}$ is defined in Tab. 6.1.

3'. Compute the following resampling penalty for each $m \in \widehat{\mathcal{M}}_n$:
$$\mathrm{pen}(m) = C\sum_{\lambda\in\Lambda_m}\left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0\right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right]\right) \ . \qquad (6.5)$$

4. Minimize the penalized empirical criterion to choose $\widehat{m}$ and thus $\widehat{s}_{\widehat{m}}$:
$$\widehat{m} \in \arg\min_{m\in\widehat{\mathcal{M}}_n}\left\{P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)\right\} \ .$$

6.3.2. Explicit formulas. When the resampling weights are exchangeable (see definition in Sect. 6.3.3), we are able to compute pen explicitly (see Lemma 5.7 in Sect. 5.7.2). It is enlightening to compare it with $\mathrm{pen}_{\mathrm{id}}$ in expectation, conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$ (we

[1] This question is studied more deeply in Sect. 8.1.

$\mathcal{L}(W)$	Efr(q)	Rad(p)	Poi(μ)	Rho(q)	Loo
$R_{2,W}(n,\widehat{p}_\lambda)$	$\frac{n}{q}\left(1-\frac{1}{n\widehat{p}_\lambda}\right)$	$\frac{1}{p}-1$	$\frac{1}{\mu}\left(1-\frac{1}{n\widehat{p}_\lambda}\right)$	$\frac{n}{q}-1$	$\frac{1}{n-1}$
$C_{W,\infty}$	q/n	$p/(1-p)$	μ	$q/(n-q)$	n

TABLE 6.1. $C_{W,\infty}$ for several resampling schemes.

denote by $\mathbb{E}^{\Lambda_m}[\cdot]$ this conditional expectation):

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(1+\frac{p_\lambda}{\widehat{p}_\lambda}\right)\left((\sigma_\lambda^r)^2+\left(\sigma_\lambda^d\right)^2\right)\mathbb{1}_{n\widehat{p}_\lambda>0} \tag{6.6}$$

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] = \frac{C}{n}\sum_{\lambda\in\Lambda_m}\left(R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)\right)\left((\sigma_\lambda^r)^2+\left(\sigma_\lambda^d\right)^2\right) \tag{6.7}$$

$$\text{with}\quad (\sigma_\lambda^r)^2 := \mathbb{E}\left[\sigma(X)^2 \,\middle|\, X\in I_\lambda\right]\quad;\quad \left(\sigma_\lambda^d\right)^2 := \mathbb{E}\left[(s(X)-s_m(X))^2 \,\middle|\, X\in I_\lambda\right]$$

$$\text{and}\quad R_{1,W}(n,\widehat{p}_\lambda) = \mathbb{E}\left[\frac{(W_1-W_\lambda)^2}{W_\lambda^2}\,\middle|\, X_1\in I_\lambda, W_\lambda>0\right] \tag{6.8}$$

$$R_{2,W}(n,\widehat{p}_\lambda) = \mathbb{E}\left[\frac{(W_1-W_\lambda)^2}{W_\lambda}\,\middle|\, X_1\in I_\lambda\right]\ . \tag{6.9}$$

One can thus choose $C=C_{W,\infty}$ such that when $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}$ is large, pen is close to $\mathrm{pen}_{\mathrm{id}}$ in expectation:

$$\begin{aligned}\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] &= \frac{2}{n}\sum_{\lambda\in\Lambda_m}\left(1+\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penW})}\right)\left((\sigma_\lambda^r)^2+\left(\sigma_\lambda^d\right)^2\right)\\ &\approx \frac{2}{n}\sum_{\lambda\in\Lambda_m}\left(1+\delta_{n,p_\lambda}\right)\left((\sigma_\lambda^r)^2+\left(\sigma_\lambda^d\right)^2\right) = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]\end{aligned} \tag{6.10}$$

with $\lim_{np\to\infty}\delta_{n,p}=0$ (Lemma 5.6 in Sect. 5.7.2) and $\lim_{np\to\infty}\delta_{n,p}^{(\mathrm{penW})}=0$ for several resampling weights (Prop. 6.5). Hence, contrary to Mallows' penalty $2\sigma^2 D_m n^{-1}$, Resampling Penalties really take into account the heteroscedasticity of the noise (σ_λ^r depends on λ) and the bias terms $\left(\sigma_\lambda^d\right)^2$. A more detailed comparison with Mallows' C_p is made in Sect. 6.6.2.

6.3.3. Examples of resampling weights. In this chapter, we consider resampling weights $W=(W_1,\dots,W_n)\in[0;+\infty)^n$ such that $\mathbb{E}[W_i]=1$ for all i and $\mathbb{E}[W_i^2]<\infty$. We mainly consider the following exchangeable weights (*i.e.* such that for any permutation τ, $(W_{\tau(1)},\dots,W_{\tau(n)}) \overset{(d)}{=} (W_1,\dots,W_n)$).

(1) *Efron* (q), $q\in\mathbb{N}\backslash\{0\}$ (Efr): $((q/n)W_i)_{1\le i\le n}$ is a multinomial vector with parameters $(q;n^{-1},\dots,n^{-1})$. A classical choice is $q=n$.
(2) *Rademacher* (p), $p\in(0;1)$ (Rad): (pW_i) are independent, with a Bernoulli (p) distribution. A classical choice is $p=1/2$.
(3) *Poisson* (μ), $\mu\in(0,\infty)$ (Poi): (μW_i) are independent, with a Poisson (μ) distribution. A classical choice is $\mu=1$.
(4) *Random hold-out* (q), $q\in\{1,\dots,n\}$ (Rho): $W_i=(n/q)\mathbb{1}_{i\in I}$ with I uniform random subset (of cardinality q) of $\{1,\dots,n\}$. A classical choice is $q=n/2$.
(5) *Leave-one-out* (Loo) = Rho $(n-1)$.

In the following, when the parameter is not mentioned, it has its "classical" value.

In each case, we can compute $R_{2,W}(n,\widehat{p}_\lambda)$ (see Tab. 6.1) and show that $R_{1,W} \approx R_{2,W}$ (see Prop. 6.5; we assumed that $nq^{-1} = \mathcal{O}(1)$ for Efr, and $n(n-q)^{-1} = \mathcal{O}(1)$ for Rho). We then define $C_{W,\infty} \sim R_{2,W}^{-1}$ as in Tab. 6.1.

REMARK 6.2. The terminology above is made to give explicit links with some classical resampling schemes. See [**MN92, HM94, vdVW96**] for more details about classical resampling weights names.

- The name "Efron" comes from the classical choice $q = n$ for which Efron weights actually are Efron's bootstrap weights.
- The name "Rademacher" for the i.i.d. Bernoulli weights comes from the classical choice $p = 1/2$ for which $(W_i - 1)_i$ are i.i.d. Rademacher random variables. As noticed by Fromont [**Fro04**], using the resampling heuristics to estimate the left-hand side of
$$\sup_{t\in S_m} \{(P - P_n)\gamma(t)\} \geq (P - P_n)\gamma(\widehat{s}_m) = \mathrm{pen}_{\mathrm{id}}(m)$$
leads to Rademacher complexities. Remark that this upper-bound is infinite in the unbounded regression case. Its use is only appropriate when γ is bounded, *e.g.* in the binary classification case with the 0-1 loss.
- Poisson weights are often used as approximations to Efron weights, via the so-called "Poissonization" technique (*cf.* [**vdVW96**], Chap. 3.5, and [**Fro04**]). They are known to be efficient for estimating several non-smooth functionals (Barbe and Bertail [**BB95**], Chap. 3; see also Mammen [**Mam92**], Sect. 1.4).
- The Leave-one-out weights also refer to the jackknife (sometimes called cross-validation). The Random hold-out (q) weights can also be called "delete-$(n-q)$ jackknife". They are both resampling schemes without replacement ([**vdVW96**], example 3.6.14), more often called *subsampling weights* (see *e.g.* Politis, Romano and Wolf [**PRW99**] on subsampling). They are thus very close to the idea of splitting the data into a training set and a validation set (*e.g.* leave-one-out, hold-out and cross-validation). Indeed, if one defines the training set as
$$\{(X_i, Y_i) \text{ s.t. } W_i \neq 0\}$$
and the validation set as its complementary, there is a one-to-one correspondence between the two ideas.
Nevertheless, we do not mean that the Loo penalization algorithm is identical to the classical cross-validation model selection algorithm. According to Chap. 5, when $C = n - 1$, it is identical to Burman's n-fold corrected cross-validation [**Bur89**].

REMARK 6.3. With the explicit computation of $R_{1,W}$ and $R_{2,W}$ for several resampling weights, we can enlighten several known results.

- In the maximum log-likelihood framework, Shibata [**Shi97**] showed the asymptotical equivalence of two bootstrap penalization methods. The first penalty, denoted by B_1, is Efron's bootstrap penalty [**Efr83**]. It is defined by (6.2) with $C = 1$ and Efron (n) weights. The second penalty, denoted B_2, was proposed by Cavanaugh and Shumway [**CS97**]. It is the equivalent of
$$2\widehat{p}_1(m) = 2\mathbb{E}_W\left[P_n\left(\gamma(\widehat{s}_m^W) - \gamma(\widehat{s}_m)\right)\right]$$

in the log-lokelihood framework. In the least-square regression framework (with histogram models), we have just shown that

$$\mathbb{E}^{\Lambda_m}\left[2\widehat{p}_1(m)\right] = \frac{2}{n}\sum_{\lambda\in\Lambda_m} R_{1,W}(n,\widehat{p}_\lambda)\left((\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right) \approx \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]$$

for several resampling schemes, including Efron's bootstrap (for which $C_{W,\infty}=1$). The concentration results of Sect. 6.8 show that this remains true without expectations. Our result is thus a non-asymptotic version of Shibata's [**Shi97**], for general resampling weights, in the least-square regression framework.

- With Efron (q_n) weights (and a bootstrap selection procedure close to RP), Shao [**Sha96**] showed that $q_n = n$ leads to an inconsistent model selection procedure for identification. On the contrary, when $q_n \to \infty$ and $q_n \ll n$, the bootstrap selection procedure becomes consistent. Considering that identification needs overpenalization within a factor that goes to infinity, (6.7) gives a simple explanation to this phenomenon since $R_{2,W} \approx n/q_n$.

6.4. Main results

In this section, we prove that RP (algorithm 6.2) has some optimality properties. We assume the existence of some non-negative constants $\alpha_{\mathcal{M}}$, $c_{\mathcal{M}}$, c_{rich}, η such that:

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n) \le c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.
(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} \in [\sqrt{n}; c_{\mathrm{rich}}\sqrt{n}]$.
(**P3**) The constant C is well chosen: $\eta C_{W,\infty} \ge C \ge C_{W,\infty}$.
(**P4**) The weight vector W is chosen among Efr, Rad, Poi, Rho and Loo (defined in Sect. 6.3.3).

6.4.1. Oracle inequalities.

THEOREM 6.1. *Assume that the (X_i, Y_i)'s satisfy the following:*

(**Ab**) *Bounded data:* $\|Y_i\|_\infty \le A < \infty$.
(**An**) *Noise-level bounded from below:* $\sigma(X_i) \ge \sigma_{\min} > 0$ *a.s.*
(**Ap**) *Polynomial decreasing of the bias: there exists* $\beta_1 \ge \beta_2 > 0$ *and* $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ *such that*

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \le l(s, s_m) \le C_{\mathrm{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for* $\mathcal{L}(X)$: $D_m \min_{\lambda\in\Lambda_m} p_\lambda \ge c_{\mathrm{r},\ell}^X$.

Let $\widehat{m}$ be the model chosen by algorithm 6.2 (under restrictions (**P1** $-$ **4**)*). Then, there exists a constant K_1 and a sequence ϵ_n converging to zero at infinity such that, with probability at least $1 - K_1 n^{-2}$,*

$$l(s, \widehat{s}_{\widehat{m}}) \le \left[2\eta - 1 + \epsilon_n\right] \inf_{m\in\mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\} \quad . \tag{6.11}$$

Moreover, we have the oracle inequality

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \le \left[2\eta - 1 + \epsilon_n\right] \mathbb{E}\left[\inf_{m\in\mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\}\right] + \frac{A^2 K_1}{n^2} \quad . \tag{6.12}$$

The constant K_1 may depend on constants in (**Ab**), (**An**), (**Ap**), ($\mathbf{Ar}_\ell^{\mathbf{X}}$) *and* (**P1** $-$ **4**)*, but not on n. The small term ϵ_n depends only on n (it can for instance be upperbounded by $\ln(n)^{-1/5}$).*

Thm. 6.1 implies the a.s. asymptotic optimality of algorithm 6.2 in this framework, when $C \sim_{n\to\infty} C_{W,\infty}$. This means that if $\mathcal{M}_n$ contains a model that takes well into account the smoothness of s and the shape of the noise $\sigma(X)$, the Resampling Penalization algorithm does as well as this oracle model for estimation. Since this does not require any knowledge about the

smoothness of s, the heteroscedasticity of σ or any other property satisfied by P, it is a *naturally adaptive algorithm*.

REMARK 6.4 (Resampling penalty *vs.* ideal deterministic penalty). It follows from the proof of Thm. 6.1 (see Prop. 6.8 and Remark 6.9) that the resampling penalties are much closer to $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ than $\mathrm{pen}_{\mathrm{id}}(m)$ itself. This means that the (ideal) penalization algorithm which uses $\mathrm{pen}(m) = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ as a penalty does not outperform much the resampling penalization algorithm. Up to the constant $1+\epsilon_n$ (that is close to 1 when n is large enough; we assumed here that $C = C_{W,\infty}$), they perform equally well on a set of probability $1 - K_1 n^{-2}$.

REMARK 6.5 (Assumptions of Thm. 6.1). Let us emphasize that no constant in the assumptions of Thm. 6.1 has to be known when computing $\widehat{m}$. This is the main reason why these assumptions are not quite restrictive. We now give a few more comments.

- (**Ab**) forbids gaussian noises, but this is not a main concern from the practical viewpoint, since practical data are often truncated, thus bounded.
- (**An**) ensures that the noise contributes for the main part of the penalty, so that the fluctuations of the penalty are negligible in front of its expectation. It is not at all necessary, and we give several alternatives for this condition below.
- ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied if we consider "almost regular" histograms and if X has a lower bounded density w.r.t. Leb. For instance, all the simulation experiments of Sect. 6.5 satisfy this assumption, even S2 or HSd2 in which the histograms are quite irregular. In general, it means that models in $\mathcal{M}_n$ should be chosen harmoniously with the law $\mathcal{L}(X)$ of the design. This is possible if one has some *a priori* knowledge on the design. Notice that we only use this assumption to prove that $D_{\widehat{m}}$ and $D_{m^\star}$ are neither very small nor very large (see Sect. 6.8.5).
- (**Ap**) on the bias may seem strange, in particular the lower bound. It is related to our way of proving the oracle inequality (6.11). Indeed, the keystone of our proof is that resampling penalties are close to the ideal ones for all sufficiently large models (*i.e.* with $D_m \geq \ln(n)^\chi$, for some $\chi > 0$ depending on the assumptions). We thus have to prove that the selected model (and the oracle) are large, which needs that s does not belong to any model in $\mathcal{M}_n$. This is a quite classical assumption, used for instance by [**Shi81, Li87, BM06c**] for proving the asymptotic optimality of Mallows' C_p.

  In our non-asymptotic framework, we need an explicit lower bound on $D_{\widehat{m}}$ and $D_{m^\star}$ (which have to go to infinity at the speed $\ln(n)^\chi$). The polynomial decreasing (**Ap**) is a convenient sufficient condition for such a lower bound. For the same kind of reasons, Stone [**Sto85**] used this assumption in the density estimation framework (see also Burman [**Bur02**], Lemma 3, for the multidimensional case). Other sufficient conditions may be derived from a careful look at the proof of Thm 6.1.

  Non-constant hölderian functions satisfy (**Ap**) with

$$\beta_1 = k^{-1} + \alpha^{-1} - (k-1)k^{-1}\alpha^{-1} \qquad \text{and} \qquad \beta_2 = 2\alpha k^{-1}$$

  when X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$ (*cf.* Thm. 6.2 and Sect. 8.10 for more details). This is why (**Ap**) is not too restrictive.

The result of Thm. 6.1 may also be proved under other assumptions, which are detailed in Sect. 8.3. Actually, if our proof works for $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ under some set of assumption, then it still works for RP. Then, RP may work under any "reasonable" assumption set.

For instance, one can remove $\sigma(X) \geq \sigma_{\min} > 0$ (**An**) and add that $\mathcal{X} \subset \mathbb{R}^k$ is bounded, equipped with $\|\cdot\|_\infty$, and

($\mathbf{Ar_u^d}$) Upper regularity of the partitions:

$$\max_{\lambda\in\Lambda_m}\{\operatorname{diam}(I_\lambda)\}\le c_{\mathrm{r,u}}^d D_m^{-\alpha_d}\operatorname{diam}(X)\ .$$

($\mathbf{Ar_u}$) Upper regularity of the partitions for Leb:

$$\max_{\lambda\in\Lambda_m}\{\operatorname{Leb}(I_\lambda)\}\le c_{\mathrm{r,u}} D_m^{-1}\operatorname{Leb}(X)\ .$$

($\mathbf{A}\sigma$) σ is piecewise K_σ-Lipschitz with at most J_σ jumps.

The boundedness assumption (**Ab**) can also be removed, at the price of adding the following: $\mathcal{X}\subset\mathbb{R}$ is bounded measurable and

($\mathbf{A}_{\mathrm{gauss}}$) The noise is sub-gaussian: there exists $c_{\mathrm{gauss}}>0$ such that

$$\forall q\ge 2,\ \forall x\in\mathcal{X},\quad \mathbb{E}\left[|\epsilon|^q\mid X=x\right]^{1/q}\le c_{\mathrm{gauss}}\sqrt{q}\ .$$

($\mathbf{A}\sigma_{\max}$) Noise-level bounded from above: $\sigma^2(X)\le\sigma_{\max}^2<+\infty$ a.s.

($\mathbf{As}_{\max}$) Bound on the target function: $\|s\|_\infty\le A$.

(**Al**) s is B-Lipschitz, piecewise C^1 and non-constant: $\pm s'\ge B_0>0$ on some interval $J\subset\mathcal{X}$ with $\operatorname{Leb}(J)\ge c_J\operatorname{Leb}(\mathcal{X})$ and $c_J>0$.

($\mathbf{Ar}_{\ell,\mathbf{u}}$) Regularity of the partitions for Leb:

$$\forall\lambda\in\Lambda_m,\quad c_{\mathrm{r},\ell}D_m^{-1}\operatorname{Leb}(\mathcal{X})\le\operatorname{Leb}(I_\lambda)\le c_{\mathrm{r,u}}D_m^{-1}\operatorname{Leb}(\mathcal{X})\ .$$

($\mathbf{Ad}_\ell$) Density bounded from below: $\exists c_X^{\min}>0,\ \forall I\subset\mathcal{X},\ \mathbb{P}(X\in I)\ge c_X^{\min}\operatorname{Leb}(I)\operatorname{Leb}(\mathcal{X})^{-1}$.

Finally, it is possible to remove simultaneously (**An**) and (**Ab**). See Sect. 8.3 for more details. Thus, for most "reasonably" difficult problems, the results of Thm. 6.1 hold true.

6.4.2. Adaptivity to the hölderian regularity. The main example where assumption (**Ap**) is satisfied is when s is non-constant and hölderian. The following result states that resampling penalization has some adaptivity to the hölderian smoothness of s in an heteroscedastic framework, since it attains the minimax rate of convergence $n^{-2\alpha/(2\alpha+k)}$ (see Stone [**Sto80**] for the homoscedastic case ; the heteroscedastic one with $k=1$ and $\alpha=1$ is studied by Galtchouk and Pergamenschikov [**GP05**], Thm. 2.2 with a fixed-design).

ALGORITHM 6.3 (Resampling penalization for histograms, regular case).
Let $(S_m)_{m\in\mathcal{M}_n}:=\left(S_{m(D)}\right)_{1\le D\le\operatorname{Leb}(\mathcal{X})^{-1/k}n^{1/k}}$, where $S_{m(D)}$ is the model of regular[2] histograms with pace D^{-1}.

0. Replace $\mathcal{M}_n$ by
$$\widehat{\mathcal{M}}_n=\left\{m\in\mathcal{M}_n\text{ s.t. }\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\ge 3\right\}\ .$$
1. Choose a resampling scheme $\mathcal{L}(W)$ among Efr, Rad, Poi, Rho and Loo.
2. Choose a constant $C\in[C_{W,\infty};\eta C_{W,\infty}]$.
3. For each $m\in\widehat{\mathcal{M}}_n$, compute $\operatorname{pen}(m)$ defined by (6.5).
4. Minimize the penalized empirical criterion to choose $\widehat{m}$ and thus $\widehat{s}_{\widehat{m}}$:
$$\widehat{m}\in\arg\min_{m\in\widehat{\mathcal{M}}_n}\left\{P_n\gamma(\widehat{s}_m)+\operatorname{pen}(m)\right\}\ .$$

THEOREM 6.2. *Let $\mathcal{Y}\subset\mathbb{R}$ and $\mathcal{X}$ be some non-empty closed ball of $(\mathbb{R}^k,\|\cdot\|_\infty)$. Assume that (X_i,Y_i) satisfy*

[2] the "regular" partition of $\mathcal{X}$ and pace D^{-1} is the collection of non-empty intersections between $\mathcal{X}$ and the family $\left(\prod_{i=1}^k[\frac{j_i}{D};\frac{j_i+1}{D})\right)_{j_1,\dots,j_k\in\mathbb{Z}}$. It has a dimension $\approx\operatorname{Leb}(\mathcal{X})D^k$ if $\mathcal{X}\subset\mathbb{R}^k$.

(**Ab**) *Bounded data:* $\|Y_i\|_\infty \leq A < \infty$.

(**An**) *Noise-level bounded from below:* $\sigma(X_i) \geq \sigma_{\min} > 0$ *a.s.*

(**Ad**$_\ell$) *Density bounded from below:* $\exists c_X^{\min} > 0$, $\forall I \subset \mathcal{X}$, $P(X \in I) \geq c_X^{\min} \operatorname{Leb}(I) \operatorname{Leb}(\mathcal{X})^{-1}$.

(**Ah**) *Hölderian regression function: there exists* $\alpha \in (0;1]$ *and* $R > 0$ *s.t.*

$$s \in \mathcal{H}(\alpha, R) \quad \text{i.e.} \quad \forall x_1, x_2 \in \mathcal{X},\ |s(x_1) - s(x_2)| \leq R \|x_1 - x_2\|_\infty^\alpha \ .$$

Let $\widehat{m}$ *be the model chosen by algorithm 6.3. Denote* $\sigma_{\max} = \sup_{\mathcal{X}} |\sigma| \leq 2A$. *Then, there exists constants* K_2 *and* K_3 *and a sequence* ϵ_n *converging to zero at infinity such that,*

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \leq K_2 \left(1 + \epsilon_n\right) R^{\frac{2k}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \sigma_{\max}^{\frac{4\alpha}{2\alpha+k}} + K_3 n^{-1} \ . \tag{6.13}$$

Assume moreover that the noise-level is smooth:

(**A**σ) σ *is piecewise* K_σ*-Lipschitz with at most* J_σ *jumps.*

Then, (6.13) *holds with* $\|\sigma\|_{L^2(\mathrm{Leb})}$ *instead of* $\sigma_{\max}$.

The constant K_2 *depends only on* η, α *and* k. *The constant* K_3 *depends only on* k, η, *constants in* (**Ab**), (**An**), (**Ad**$_\ell$), (**Ah**) *(and* (**A**σ) *for the last result) and* s *through its variation over* $\mathcal{X}$*:*

$$\operatorname{varia}_{\mathcal{X}}(s) := \sup_{\mathcal{X}} s - \inf_{\mathcal{X}} s \ . \tag{6.14}$$

The small term ϵ_n *depends only on* n, *and can for instance be upperbounded by* $\ln(n)^{-1/5}$.

Remark 6.6.

(1) The minimax rate of estimation under some heteroscedastic noise has already been adressed by Galtchouk and Pergamenschikov [**GP05**] (Thm. 2.2) in the fixed design case, with $\mathcal{X} \subset \mathbb{R}$ (see also Efromovich and Pinsker [**EP96**]). They assume lower and upper bounds on the noise-level ((**An**)+(**A**$\sigma_{\max}$)), some regularity condition (σ^2 continuous, with Riemann sums that converges to its L^2(Leb) norm), and consider more regular target functions ($s \in \mathcal{H}(\alpha, R)$ with $\alpha \in \mathbb{N} \setminus \{0\}$, where we consider $\alpha \in (0;1]$). Then, they show that the minimax rate of estimation over $\mathcal{H}(\alpha, R)$ is indeed

$$R^{\frac{2}{2\alpha+1}} n^{\frac{-2\alpha}{2\alpha+1}} \|\sigma\|_{L^2(\mathrm{Leb})}^{\frac{4\alpha}{2\alpha+1}} \ .$$

As a consequence, when $\alpha = 1$ and $k = 1$, the estimation rate in (6.13) is optimal up to some factor independent from n, R and σ.

(2) In (6.13), the constant K_3 is not uniform over the hölderian ball $\mathcal{H}(\alpha, R)$, but only on its subsets

$$\mathcal{H}_\epsilon(\alpha, R) := \{\, s \in \mathcal{H}(\alpha, R) \quad \text{s.t. } \operatorname{varia}_{\mathcal{X}}(s) \geq \epsilon \,\} \ .$$

for $\epsilon > 0$. Indeed, the lower bound in (**Ap**) cannot be made uniform without ensuring that s is sufficiently non-constant.

This is probably a technical restriction, since we also prove (6.13) with K_3 uniform over the set of constant functions. Moreover, in the minimax viewpoint, the harder functions to estimate are certainly not the almost constant ones. If this issue was solved, then $\widehat{s}_{\widehat{m}}$ would be proved to attain the minimax rate over $\mathcal{H}(\alpha, R)$ with an heteroscedastic noise $\|\sigma\|_{L^2(\mathrm{Leb})}$, up to some factor independent from n, R and σ.

(3) If s is only piecewise α-hölder, with at most J_s jumps (of height bounded by $2A$), then the same results hold. This requires the additional assumption that s is non-constant on some ball B of $(\mathbb{R}^k, \|\cdot\|_\infty)$ on which it is continuous. The constant K_3 then also depends on J_s, $\operatorname{varia}_B(s)$ and $\operatorname{Leb}(B)/\operatorname{Leb}(\mathcal{X})$.

(4) As for Thm. 6.1, the boundedness of the data and the lower bound on the noise level can be replaced by other assumptions.

6.5. Simulations

To illustrate the results of Sect. 6.4, we compare the performances of algorithm 6.2 (with several resampling schemes), Mallows' C_p and VFCV on some simulated data.

6.5.1. Experimental setup. In the following simulation study, we consider the same data sets as in Sect. 4.4.2 and 5.4. We briefly describe them again below. First, we focus on four main experiments, called S1, S2, HSd1 and HSd2. Data are generated according to (6.3) with X_i i.i.d. uniform on $\mathcal{X} = [0;1]$ and $\epsilon_i \sim \mathcal{N}(0,1)$ independent from X_i. The experiments differ from the regression function s (smooth for S, see Fig. 4.3; smooth with jumps for HS, see Fig. 4.4), the noise type (homoscedastic for S1 and HSd1, heteroscedastic for S2 and HSd2), the number n of data and the families of models (see the top of Tab. 6.2; "regular" refers to the family of algorithm 6.3; "with two bin sizes" means regular on $[0;1/2]$ and on $(1/2;1]$; "dyadic" means that we limit ourselves to bin sizes of the form 2^{-k}, $k \in \mathbb{N}$). Instances of data sets are given in Fig. 4.5 to 4.8 (in Sect. 4.4.2).

We compare the following algorithms:

- Mal Mallows' C_p penalty: $\mathrm{pen}(m) = 2\widehat{\sigma}^2 D_m n^{-1}$ where $\widehat{\sigma}^2$ is the variance estimator (6.18) used in [**Bar00**], Sect. 6.
- VFCV Classical V-fold cross-validation, with $V \in \{2, 5, 10, 20\}$ (defined by (5.1) in Sect. 5.2.1).
- LOO Classical Leave-one-out (*i.e.* VFCV with $V = n$).
- penEfr Efron (n) penalty, $C = C_{W,\infty} = 1$.
- penRad Rademacher penalty, $C = C_{W,\infty} = 1$.
- penRho Random hold-out ($n/2$) penalty, $C = C_{W,\infty} = 1$.
- penLoo Leave-one-out penalty, $C = C_{W,\infty} = n - 1$.

For each of these, we also consider the same penalties multiplied by 5/4 (denoted by a + symbol added after its shortened name). This intends to test for overpenalization.

In each experiment, for each simulated data set, we first remove the models with less than 2 data points in one piece of their associated partition. Then, we compute the least-square estimators $\widehat{s}_m$ for each $m \in \widehat{\mathcal{M}}_n$. Finally, we select $\widehat{m} \in \widehat{\mathcal{M}}_n$ using each algorithm and compute its true excess risk $l(s, \widehat{s}_{\widehat{m}})$ (and the excess risk of each model $m \in \mathcal{M}_n$). Since we simulate N data sets ($N = 1000$ in the four main experiments), we can then estimate the two following benchmarks:

$$C_{\mathrm{or}} = \frac{\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)\right]} \qquad C_{\mathrm{path-or}} = \mathbb{E}\left[\frac{l(s, \widehat{s}_{\widehat{m}})}{\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)}\right]$$

Basically, C_{or} is the constant that should appear in an oracle inequality like (6.11), and $C_{\mathrm{path-or}}$ corresponds to a pathwise oracle inequality like (6.12). As C_{or} and $C_{\mathrm{path-or}}$ approximatively give the same rankings between algorithms, we only report C_{or} in Tab. 6.2.

6.5.2. Results and comments. We observe that penRad and penRho are always competitive with Mallows and much better for more "difficult" problems (S2 is heteroscedastic; jumps in HSd1[3] and HSd2 induce much bias). Notice that in the case of HSd2, penRad and penRho

[3] In the particular case of HSd1, we must add that Mallows' C_p performs quite better when σ^2 is known ($C_{\mathrm{or}} \approx 1.044 \pm 0.004$ for Mal, which is still worse than VFCV and penRP; and $C_{\mathrm{or}} \approx 1.606 \pm 0.015$ for Mal+). This is mainly due to the difficulty of estimating σ^2 accurately when even large models can have a large bias. However, this is no longer the case for HSd2, in which the knowledge of σ^2 does not improve Mal and Mal+.

TABLE 6.2. Accuracy indexes C_{or} for each algorithm in four experiments, $\pm$ a rough estimate of uncertainty of the value reported (*i.e.* the empirical standard deviation divided by $\sqrt{N}$; $N = 1000$). In each column, the more accurate algorithms (taking the uncertainty into account) are bolded.

Experiment	S1	S2	HSd1	HSd2
s	sin	sin	HeaviSine	HeaviSine
$\sigma(x)$	1	x	1	x
n (data)	200	200	2048	2048
$\mathcal{M}_n$	regular	2 bin sizes	dyadic, regular	dyadic, 2 bin sizes
Mal	1.928 ± 0.04	3.864 ± 0.02	1.606 ± 0.015	1.487 ± 0.011
Mal+	$\mathbf{1.800 \pm 0.03}$	4.047 ± 0.02	1.606 ± 0.015	1.487 ± 0.011
2−FCV	2.078 ± 0.04	2.542 ± 0.05	$\mathbf{1.002 \pm 0.003}$	1.184 ± 0.004
5−FCV	2.137 ± 0.04	2.582 ± 0.06	1.014 ± 0.003	1.115 ± 0.005
10−FCV	2.097 ± 0.05	2.603 ± 0.06	1.021 ± 0.003	1.109 ± 0.004
20−FCV	2.088 ± 0.04	2.578 ± 0.06	1.029 ± 0.004	1.105 ± 0.004
LOO	2.077 ± 0.04	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penEfr	2.597 ± 0.07	3.152 ± 0.07	1.067 ± 0.005	1.114 ± 0.005
penRad	1.973 ± 0.04	2.485 ± 0.06	1.018 ± 0.003	$\mathbf{1.102 \pm 0.004}$
penRho	1.982 ± 0.04	2.502 ± 0.06	1.018 ± 0.003	1.103 ± 0.004
penLoo	2.080 ± 0.05	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penEfr+	2.016 ± 0.05	2.605 ± 0.06	1.011 ± 0.003	$\mathbf{1.097 \pm 0.004}$
penRad+	$\mathbf{1.799 \pm 0.03}$	$\mathbf{2.137 \pm 0.05}$	$\mathbf{1.002 \pm 0.003}$	$\mathbf{1.095 \pm 0.004}$
penRho+	$\mathbf{1.798 \pm 0.03}$	$\mathbf{2.142 \pm 0.05}$	$\mathbf{1.002 \pm 0.003}$	$\mathbf{1.095 \pm 0.004}$
penLoo+	$\mathbf{1.844 \pm 0.03}$	$\mathbf{2.215 \pm 0.05}$	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.096 \pm 0.004}$

do better than any linear penalty (possibly with a slope that depends on both the data and the unknown law P; see Sect. 6.6.2). On the other hand, VFCV is a little worse than Mal for easy problems (S1) and better for more difficult ones, but never better than penRad or penRho.

The best resampling schemes (not taking overpenalization into account) are Rad and Rho, in view of S1 and S2 (dyadic models do not induce much differences between them in HSd1 and HSd2). Looking more carefully at the values of the penalties, it appears that Loo is slightly underpenalizing and Efr strongly overfits. The comparison penRad $\approx$ penRho $>$ penLoo $\gg$ penEfr can also be derived from Sect. 6.3 and 6.6.1. Moreover, when variability is considered (it is measured through the empirical variance of $l(s, \widehat{s}_{\widehat{m}}) / \inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)$ and reported in Tab. 6.2 to 6.4), the order of the weights is unchanged. Contrary to what is usually stated, the bootstrap (penEfr) can thus be more variable than the leave-one-out (penLoo).

In the four experiments, overpenalizing within a factor 5/4 leads to better results, mainly because n is quite small for the noisy (S1, S2) or irregular (HSd1, HSd2) signals observed. This is no longer the case for some larger n or smaller σ.

Finally, we report the results of eight more experiments in Tab. 6.3-6.4. They are quite similar to the first four ones, since we only changed a few parameters among n, σ and s (instances of data sets and regression functions are plotted on Fig. 4.9 to 4.18; see Sect. 4.4.2). Remark that we simulated only $N = 250$ data sets. The comparison between Mallows', VFCV and Resampling Penalization is quite the same: in "easy" homoscedastic frameworks (S1000, S$\sqrt{0.1}$, S0.1), their

TABLE 6.3. Accuracy indexes C_{or} for more experiments ($N = 250$).

Experiment	S1000	S$\sqrt{0.1}$	S0.1	Svar2
s	sin	sin	sin	sin
$\sigma(x)$	1	$\sqrt{0.1}$	0.1	$\mathbb{1}_{x\geq 1/2}$
n (data)	1000	200	200	200
$\mathcal{M}_n$	regular	regular	regular	2 bin sizes
Mal	$\mathbf{1.667 \pm 0.04}$	$\mathbf{1.611 \pm 0.03}$	$\mathbf{1.400 \pm 0.02}$	3.520 ± 0.03
Mal+	$\mathbf{1.619 \pm 0.03}$	$\mathbf{1.593 \pm 0.03}$	1.426 ± 0.02	3.672 ± 0.03
2−FCV	$\mathbf{1.668 \pm 0.04}$	1.663 ± 0.04	$\mathbf{1.394 \pm 0.02}$	2.960 ± 0.15
5−FCV	1.756 ± 0.07	1.693 ± 0.04	$\mathbf{1.393 \pm 0.02}$	2.950 ± 0.16
10−FCV	1.746 ± 0.04	1.664 ± 0.04	$\mathbf{1.385 \pm 0.02}$	2.681 ± 0.14
20−FCV	1.774 ± 0.05	1.645 ± 0.03	$\mathbf{1.382 \pm 0.02}$	2.742 ± 0.16
LOO	1.768 ± 0.05	1.639 ± 0.04	$\mathbf{1.379 \pm 0.02}$	2.641 ± 0.15
penEfr	1.813 ± 0.05	1.888 ± 0.05	1.417 ± 0.02	3.451 ± 0.20
penRad	1.748 ± 0.05	$\mathbf{1.609 \pm 0.03}$	$\mathbf{1.405 \pm 0.02}$	2.510 ± 0.15
penRho	1.748 ± 0.05	$\mathbf{1.619 \pm 0.03}$	$\mathbf{1.404 \pm 0.02}$	2.518 ± 0.15
penLoo	1.776 ± 0.05	1.641 ± 0.04	$\mathbf{1.379 \pm 0.02}$	2.656 ± 0.15
penEfr+	$\mathbf{1.636 \pm 0.04}$	1.670 ± 0.04	$\mathbf{1.407 \pm 0.02}$	2.614 ± 0.16
penRad+	$\mathbf{1.619 \pm 0.03}$	$\mathbf{1.574 \pm 0.03}$	1.417 ± 0.02	$\mathbf{2.232 \pm 0.12}$
penRho+	$\mathbf{1.619 \pm 0.03}$	$\mathbf{1.578 \pm 0.03}$	1.417 ± 0.02	$\mathbf{2.243 \pm 0.12}$
penLoo+	$\mathbf{1.626 \pm 0.03}$	$\mathbf{1.587 \pm 0.03}$	$\mathbf{1.401 \pm 0.02}$	$\mathbf{2.349 \pm 0.13}$

performances are similar. An harder problem such as Svar2 (which is heteroscedastic but different from S2) make Mallows' fail whereas the two others only get a bit worse.

As expected, taking n larger (S1000) or σ smaller (S$\sqrt{0.1}$ and S0.1) make the constant C_{or} closer to 1. Notice also that the overpenalization factor 5/4 is generally not optimal, and even not always better than 1. We have for instance $C_{\text{or}}(\text{penLoo}) < C_{\text{or}}(\text{penRho}) < C_{\text{or}}(\text{penRho+})$ in S0.1 (with only small differences), although penLoo may slightly underpenalize.

In Tab. 6.4, we consider several other target functions s (plotted on Fig. 4.13, 4.15 and 4.17; see Sect. 4.4.2). In Sqrt, s is not Lipschitz but only 1/2-hölderian. In His6, s is even not continuous, since it belongs to the model of regular histograms with 6 pieces. The results obtained with these two functions strengthen the fact that the assumptions of our theorem are not actual restrictions for algorithm 6.2.

On the other hand, with DopReg and Dop2bin, we use the classical function Doppler of the thresholding literature (see [**DJ95**]). Dyadic histograms approximate it with a large bias, and this bias is not homogeneous in space (since Doppler is much more variable on the left of the unit interval). In such a framework, taking into account the σ_λ^d terms (see (6.6)) in the penalty seems necessary (in particular in Dop2bin, where the histograms have two different bin sizes), so that Mallows' C_p can fail even with homoscedastic data. This fault is avoided by both VFCV and resampling penalties which behave quite well.

6.6. Discussion

6.6.1. Practical implementation.

Computation time. An exact computation (without using our formulas (5.37) and (5.38) for histograms) would be either impossible or very greedy. So, we recommend to make a Monte-Carlo

TABLE 6.4. Accuracy indexes C_{or} for more experiments ($N = 250$).

Experiment	Sqrt	His6	DopReg	Dop2bin
s	$\sqrt{\cdot}$	His_6	Doppler	Doppler
$\sigma(x)$	1	1	1	1
n (data)	200	200	2048	2048
$\mathcal{M}_n$	regular	regular	dyadic, regular	dyadic, 2 bin sizes
Mal	2.295 ± 0.11	1.969 ± 0.11	1.130 ± 0.011	1.469 ± 0.013
Mal+	$\mathbf{1.989 \pm 0.08}$	$\mathbf{1.799 \pm 0.09}$	1.130 ± 0.011	1.459 ± 0.014
2−FCV	2.489 ± 0.12	2.788 ± 0.13	1.097 ± 0.005	1.165 ± 0.009
5−FCV	2.777 ± 0.16	2.316 ± 0.12	1.064 ± 0.005	$\mathbf{1.049 \pm 0.006}$
10−FCV	2.571 ± 0.13	2.074 ± 0.11	1.043 ± 0.005	$\mathbf{1.051 \pm 0.006}$
20−FCV	2.561 ± 0.12	2.071 ± 0.11	$\mathbf{1.034 \pm 0.005}$	$\mathbf{1.053 \pm 0.006}$
LOO	2.695 ± 0.14	2.059 ± 0.12	$\mathbf{1.026 \pm 0.005}$	$\mathbf{1.058 \pm 0.006}$
penEfr	3.468 ± 0.22	2.721 ± 0.16	$\mathbf{1.030 \pm 0.007}$	1.064 ± 0.006
penRad	2.396 ± 0.11	$\mathbf{1.884 \pm 0.10}$	1.043 ± 0.006	$\mathbf{1.055 \pm 0.006}$
penRho	2.448 ± 0.12	$\mathbf{1.907 \pm 0.11}$	1.043 ± 0.006	$\mathbf{1.055 \pm 0.006}$
penLoo	2.695 ± 0.14	2.063 ± 0.12	$\mathbf{1.026 \pm 0.005}$	$\mathbf{1.058 \pm 0.006}$
penEfr+	2.205 ± 0.11	1.924 ± 0.11	1.056 ± 0.006	$\mathbf{1.057 \pm 0.006}$
penRad+	$\mathbf{2.036 \pm 0.09}$	$\mathbf{1.746 \pm 0.09}$	1.092 ± 0.004	$\mathbf{1.058 \pm 0.007}$
penRho+	$\mathbf{2.053 \pm 0.09}$	$\mathbf{1.747 \pm 0.09}$	1.091 ± 0.004	$\mathbf{1.059 \pm 0.007}$
penLoo+	$\mathbf{2.152 \pm 0.10}$	$\mathbf{1.858 \pm 0.10}$	1.082 ± 0.005	$\mathbf{1.048 \pm 0.006}$

simulation of a few number of weight vectors, in order to approach the exact resampling penalty we are dealing with here. Practical methods for this are addressed by Hall [**Hal92**], appendix II. In addition, we proved in Sect. 10.2.5 a non-asymptotic estimation of the accuracy of Monte-Carlo approximation (Prop. 10.5). In our framework, a similar proof (based upon McDiarmid's inequality, Prop. 8.7 in Sect. 8.5) would give a practical way of choosing the number of weight vectors to consider (at least for Rad, Rho and Loo weights).

An alternative to Monte-Carlo approximation would be the use of V-fold cross-validation weights. The resulting *V-fold penalties* are defined and studied in Chap. 5.

Choice of the weights. According to the simulations of Sect. 6.5, the best weights (for accuracy of prediction and for the variability of this accuracy) are Rho and Rad. On the other hand, Loo weights are much better from the computational viewpoint (without Monte-Carlo approximation), and they induce only a small underpenalization. Notice that from both accuracy and variability viewpoints, Efron's bootstrap weights appear to perform worse than Rho, Rad, and even Loo.

Thus, if an exact computation of Loo penalties is possible, we suggest to choose these weights and enlarge the constant C (we explain why below). Otherwise, Monte-Carlo approximation with Leave-one-out weights is known to be quite variable. Then, Rho or Rad weights should be prefered to Loo.

In a more general framework, this analysis may be slightly changed: when the empirical minimization algorithm $\xi_{1\dots n} \mapsto \widehat{s}_m$ is unstable, the leave-one-out is known to be highly variable. Then, Rho and Rad weights are likely to be much better.

About the choice of the weights, we also refer to Barbe and Bertail [**BB95**], Chap. 2, where an asymptotic analysis of a large family of exchangeable weights is proposed, based upon Edgeworth expansions.

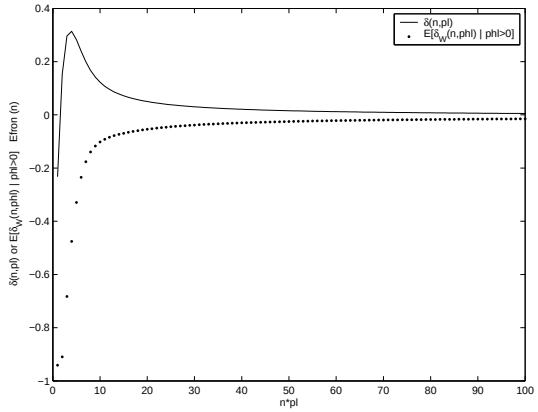


FIGURE 6.1. $\delta_{n,p_\lambda} > 0 > \delta^{(\mathrm{penEfr(n)})}_{n,p_\lambda}$.

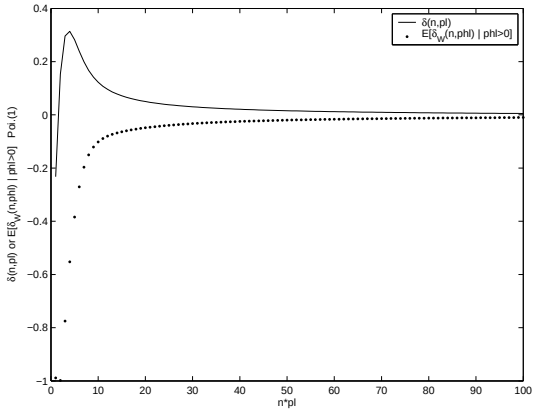


FIGURE 6.2. $\delta_{n,p_\lambda} > 0 > \delta^{(\mathrm{penPoi(1)})}_{n,p_\lambda}$.

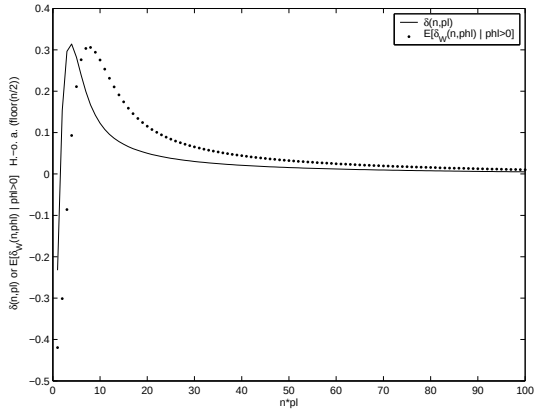


FIGURE 6.3. $\delta_{n,p_\lambda} > \delta^{(\mathrm{penRho}(n/2))}_{n,p_\lambda}$ for $np_\lambda \geq 6$.

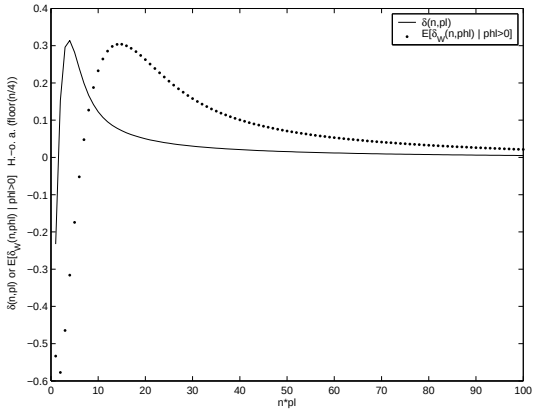


FIGURE 6.4. $\delta_{n,p_\lambda} > \delta^{(\mathrm{penRho}(n/4))}_{n,p_\lambda}$ for $np_\lambda \geq 9$.

Second-order terms. In this paragraph, we intend to understand the comparison

$$\mathrm{penRad} \approx \mathrm{penRho} > \mathrm{penLoo} >> \mathrm{penEfr} \tag{6.15}$$

observed in the simulations of Sect. 6.5. This is done by a more accurate computation of the expectation of $\mathrm{pen}(m)$, taking into account second order terms. We then compare numerically these terms for the classical weights.

In Sect. 6.3, we showed that $\mathrm{pen}_{\mathrm{id}}(m)$ and $\mathrm{pen}(m)$ have the same expectation, up to small terms δ_{n,p_λ} and $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$. We deduce that

$$\mathbb{E}\left[\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(\delta^{(\mathrm{penW})}_{n,p_\lambda} - \delta_{n,p_\lambda}\right)(\sigma_\lambda)^2 \tag{6.16}$$

$$\text{with} \qquad \delta^{(\mathrm{penW})}_{n,p_\lambda} := \mathbb{E}\left[\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} \,\middle|\, \widehat{p}_\lambda > 0\right] .$$

We computed δ_{n,p_λ} and $\delta^{(\mathrm{penW})}_{n,p_\lambda}$ for several resampling schemes, when $n = 200$. The results are given on Fig. 6.1 to 6.6 (with straight lines for δ_{n,p_λ} and dots for $\delta^{(\mathrm{penW})}_{n,p_\lambda}$). Loo is the most accurate. On the contrary, Rho $(n/2)$ and Rad give overestimations of δ_{n,p_λ}. The bias of Rho (q) is a decreasing function of q, as shown by Figure 6.4. Finally, Efr and Poi are strongly underestimating the ideal penalty, because of the $1-(n\widehat{p}_\lambda)^{-1}$ term.

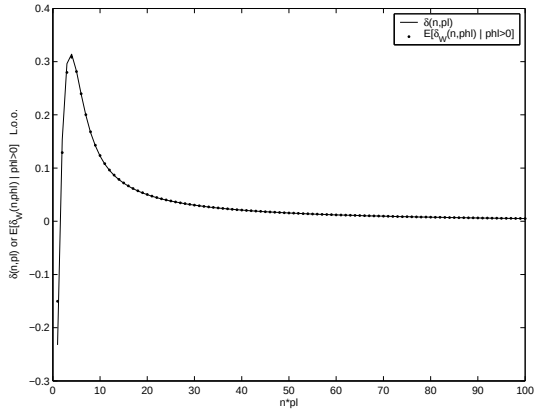


FIGURE 6.5. $\delta_{n,p_\lambda} \approx \delta^{(\mathrm{penLoo})}_{n,p_\lambda}$.

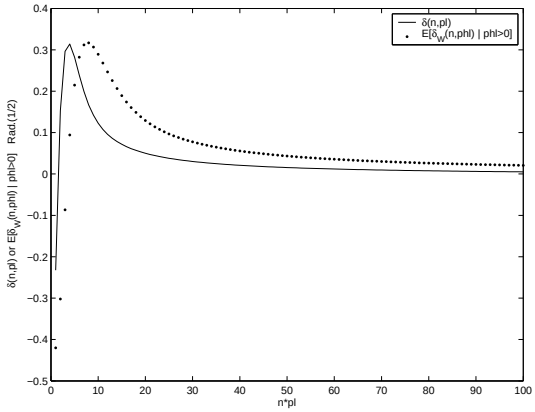


FIGURE 6.6. $\delta_{n,p_\lambda} > \delta^{(\mathrm{penRad}(1/2))}_{n,p_\lambda}$ for $np_\lambda \geq 6$.

This may explain (6.15): Efr and Poi are clearly underpenalizing. The strong accuracy of Loo in expectation make it underpenalize a bit. Indeed, the fluctuations of pen and $\mathrm{pen}_{\mathrm{id}}$ around their expectations are probably as important as these second-order terms. Then, being too accurate in expectation may lead to underestimation with positive probability. Finally, Rad and Rho overpenalize in expectation, so that they underpenalize less often in our simulations.

We also observe that $\delta^{(\mathrm{penRho})}_{n,p_\lambda} \propto \delta_{n,p_\lambda}$ when np_λ is large enough. Then, Loo and Rho are almost equivalent, up to the choice of C. If a wise tuning of C is possible, we just have to choose between Loo and Rho according to computation issues (see the discussion above).

Choice of the constant C.

Optimal constant. The optimal constant $C^\star$ is the one for which pen is an unbiased estimator of the ideal penalty $\mathrm{pen}_{\mathrm{id}}$ (at least for the "reasonable" models). Thus, $C = C_{W,\infty}$ is asymptotically optimal. In the histogram case, we even proved non-asymptotic oracle inequalities for $C = C_{W,\infty}$. However, a careful look at the proofs shows that such a result holds for any constant $C > C^\star/2$. Then, $C_{W,\infty}$ may not be exactly equal to $C^\star$ when the sample size n is small. Moreover, we do not have exact non-asymptotic expressions for $C_{W,\infty}$ for the general algorithm 6.1. Using the asymptotical value of $C_{W,\infty}$ may lead to an uncorrect algorithm if the size of $\mathcal{M}_n$ depends on n or is infinite, even for large n!

In order to solve this issue, we suggest to choose C with the so-called "slope heuristics", proposed by Birgé and Massart [**BM06c**] for penalties linear in dimension. Their claim is that the optimal penalty is twice the minimal penalty, *i.e.* the one under which the selected model is obviously too large. With Massart, we extended this result to any shape of the ideal penalty (*cf.* Chap. 3). This leads to estimating the shape of $\mathrm{pen}_{\mathrm{id}}$ by resampling, and the constant C with the slope heuristics. The resulting algorithm 11.1 is described in Sect. 11.3.2.

Overpenalization. When n is small, σ is large, or when s is non-smooth, it may be necessary to overpenalize. The simulations of Sect. 6.5 showed that overpenalization may greatly improve the quality of prediction (*e.g.* in the experiments of Tab. 6.2).

This problem would appear even if we knew the "optimal" constant $C^\star$ such that pen is non-asymptotically unbiased. Indeed, $C^\star$ does not take into account the deviations of $\mathrm{pen}(m)$ around its expectation $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$, neither the deviations of $\mathrm{pen}_{\mathrm{id}}(m)$ around its expectation. To avoid the possible overfit induced by these fluctuations, we have to slightly enlarge C.

The factor 5/4 taken in our simulations is obviously not a universal one (*e.g.* it is worse than 1 in experiments S0.1 and DopReg). Our proposal is to modify definition (6.2) of pen by replacing

the expectation by an α-quantile, as follows:

$$\mathrm{pen}(m) = C \inf \left\{ t \in \mathbb{R} \text{ s.t. } P_n^W \left[P_n \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) - P_n^W \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) > t \right] \leq \alpha \right\} \quad . \tag{6.17}$$

Of course, the level α remains to be chosen. With $\alpha = 0.5$ (*i.e.* take the median), we obtain almost the same penalty as (6.2). We do not have any theoretical result nor heuristics for the choice of α: the amount of overpenalization corresponding to any particular α depends on how the variance of $\mathrm{pen}_{\mathrm{id}}(m)$ increases or decreases with m. The constant C can still be taken equal to $C_{W,\infty}$, or estimated by the slope heuristics.

Resampling quantiles of the form (6.17) can also be used to derive a (two-sided) confidence region for the prediction error $\left(P\gamma\left(\widehat{s}_m\right) \right)_{m \in \mathcal{M}_n}$. Then, we may deduce a "confidence set for $m^\star$" instead of a single model $\widehat{m}$, and choose the more parcimonious model as $\widehat{m}$. Such an algorithm would "overpenalize" more and more when the level α of the confidence region goes to zero.

Such modifications are of course impossible with Mallows' penalty. This shows one more drawback of Mallows' C_p for difficult problems. On the contrary, the classical V-fold cross-validation can be modified in such a way. Further comments on overpenalization are given in Sect. 2.4.1 and 11.3.3.

6.6.2. Comparison with other procedures. In this chapter, we have shown that Resampling Penalization should work well in almost all "reasonable" frameworks. This robustness is a key property of RP, which may be used in almost every situation. However, computing the resampling penalties may be quite long, even with a Monte-Carlo approximation, when minimizing $P_n^W \gamma(t, \cdot)$ over $t \in S_m$ is hard. In such cases, we need some clues for choosing between simple procedures (*e.g.* Mallows') and RP. In particular, for "easy" problems, RP can behave worse than Mallows', simply because it is more general. We would like to know what are those "easy" problems, for which we can avoid long computations.

Mallows' C_p. Mallows' penalty is equal to $2\sigma^2 D_m n^{-1}$ for a model m of dimension D_m. Non-asymptotic results about Mallows'-like penalties can be found in [**BBM99, Bar00, Bar02**]. They imply that Mallows' penalty is asymptotically optimal in the homoscedastic framework, when $\mathcal{M}_n$ is not too large.

When the (constant) noise-level σ is unknown, one has to estimate it. Introducing artificially a model $S_{\lfloor n/2 \rfloor}$ of dimension $\lfloor n/2 \rfloor$, Baraud [**Bar00, Bar02**] suggests to estimate σ^2 with

$$\widehat{\sigma}^2 = \frac{d^2 \left(Y_{1\ldots n}, S_{\lfloor n/2 \rfloor} \right)}{n - \lfloor n/2 \rfloor} \quad , \tag{6.18}$$

where $Y_{1\ldots n} = (Y_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ and d the Euclidean distance on $\mathbb{R}^n$. He then showed that this data-driven model selection procedure has some adaptivity property in both the fixed-design and random-design frameworks.

Assume for the sake of simplicity that n is even, and choose $S_{n/2}$ such that each piece of the associated partition contains exactly two data points. Reordering the (X_i, Y_i) according to X_i, we then have

$$\widehat{\sigma}^2 = \frac{2}{n} \sum_{i=1}^{n/2} \left(\left(Y_{2i-1} - \frac{Y_{2i-1} + Y_{2i}}{2} \right)^2 + \left(Y_{2i} - \frac{Y_{2i-1} + Y_{2i}}{2} \right)^2 \right) = \frac{1}{n} \sum_{i=1}^{n/2} \left(Y_{2i} - Y_{2i-1} \right)^2$$

$$\text{hence,} \qquad \mathrm{pen}_{\mathrm{Mallows}}(m) = \frac{2 D_m}{n^2} \sum_{i=1}^{n/2} \left(Y_{2i} - Y_{2i-1} \right)^2 \quad .$$

Since

$$\mathbb{E}\left[(Y_{2i}-Y_{2i-1})^2 \,\middle|\, X_{2i}, X_{2i-1}\right] = \sigma(X_{2i})^2 + \sigma(X_{2i-1})^2 + (s(X_{2i}) - s(X_{2i-1}))^2$$

we obtain that

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{Mallows}}(m)\right] = \frac{2}{n}\sum_{\lambda\in\Lambda_m} (D_m \widehat{p}_\lambda)\,(\sigma_\lambda^r)^2 + \frac{2D_m}{n^2}\sum_{i=1}^{n/2} (s(X_{2i}) - s(X_{2i-1}))^2 \quad .$$

When s is smooth, the second term is negligible in front of $2n^{-1}\sum_{\lambda\in\Lambda_m}\left(\sigma_\lambda^d\right)^2$ (see (6.6) in Sect. 6.3.2). This means that Mallows' penalty does not take into account the bias component of $\mathrm{pen}_{\mathrm{id}}$.

On the other hand, the variance component of $\mathrm{pen}_{\mathrm{id}}$ (which is the main one in general) is deformed in $\mathrm{pen}_{\mathrm{Mallows}}$. The penalty part corresponding to I_λ is multiplied by $D_m\widehat{p}_\lambda$, which may not be close to 1 when the model m is not regular w.r.t. $\mathcal{L}(X)$. This happens for instance in the experiments S2 and HSd2 in Sect. 6.5.

These two main differences between Mallows' C_p and Resampling Penalization enlightens several "hard" problems:

- heteroscedastic noise, with irregular histograms and X uniform (*e.g.* S2, HSd2 or Svar2 of Sect. 6.5),
- heteroscedastic noise, with regular histograms and X highly non-uniform on $\mathcal{X}$,
- regression function s with jumps (*e.g.* HeaviSine) or non-smooth areas (*e.g.* Doppler).

In either of those cases, one should avoid the use of Mallows'-like penalties, and we suggest RP as an efficient alternative. We show below that the first class of problems can make any linear penalty unefficient.

Linear penalties. The simplicity of Mallows' C_p comes from the fact that its shape is fixed *a priori* as linear in the dimension D_m of the models. Thus,

$$\mathrm{pen}(m) = \widehat{K} D_m$$

and there is only one constant $\widehat{K}$ to determine. In the case of Mallows', we take

$$\widehat{K}_{\mathrm{Mallows}} = 2\sigma^2 n^{-1} \quad \text{or} \quad 2\widehat{\sigma}^2 n^{-1}$$

if the mean variance level σ is unknown. Following the slope heuristics of Birgé and Massart [**BM06c**], one can also define a data-dependent constant $\widehat{K}_{\mathrm{slope}}$ with steps 3 and 4 of algorithm 3.1 (and $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$).

However, in view of (6.6), the ideal penalty is not linear in general, even in expectation. We show in Chap. 4 that these linear penalties can fail, in the sense that they can not satisfy an oracle inequality with a constant smaller than some $\kappa > 1$. See Sect. 4.3 for a theoretical result, and Sect. 4.4 for experimental evidence. In particular, we consider in Sect. 4.4.2 the experiments of Sect. 6.5, and it appears that in experiment HSd2, even a linear penalty using both the data and the unknown distribution P is less efficient than all the V-fold and Resampling Penalties. In HSd2, linear penalties are even worse than the classical V-fold cross-validation with $V \in \{5, 10, 20\}$.

Thus, in difficult situations such as HSd2, Resampling Penalization is an efficient alternative to linear penalization.

Refined versions of Mallows'. In least-square regression and other frameworks, several penalties have been defined as refinements of Mallows' C_p, in Gaussian frameworks (Barron, Birgé and Massart [**BBM99**]) as in non-Gaussian ones (Baraud [**Bar02**]). Basically, when $\mathrm{Card}(\mathcal{M}_n)$ is

polynomial in n, these penalties are linear in D_m. So, they have at least the same drawbacks as the optimal linear penalty above.

When $\mathrm{Card}(\mathcal{M}_n)$ is larger (*e.g.* exponential in n), one has to take a larger penalty of the form

$$\mathrm{pen}(m) = K D_m \left(1 + c \ln \left(\frac{n}{D_m} \right) \right) \ ,$$

as in Birgé and Massart [**BM06c**] or Sauvé [**Sau06**]. With such a family of models, one can not use Resampling Penalization without modifications. Indeed, uniform deviations for $\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)$ derived from the union bound may be too large, so that the model selection procedure can fail.

In order to solve this issue, we propose to apply algorithm 6.1 to $(\widetilde{S}_D)_{1 \leq D \leq n}$ instead of $(S_m)_{m \in \mathcal{M}_n}$, with

$$\widetilde{S}_D := \bigcup_{D_m = D} S_m \ .$$

This new model selection problem satisfies the polynomial complexity assumption. By grouping models according to D_m (or any other natural index of complexity of S_m), we allow the Resampling procedure to detect the complexity of $\mathcal{M}_n$ through the complexity of each $\widetilde{S}_D$. However, our results for histograms cannot be extended to this case since $\widetilde{S}_D$ is not an histogram model, but only a union of histogram models with the same number of pieces. Results in this framework would be very interesting, since they could be applied to CART algorithm (defined by Breiman *et al.* [**BFOS84**]; *cf.* also [**ST06**]).

Ad hoc *procedures.* One of the main points of Thm. 6.1 and 6.2 is that Resampling Penalization works in an heteroscedastic framework, contrary to Mallows' C_p. However, it is possible to adapt Mallows' penalty to heteroscedasticity, for instance by splitting $\mathcal{X}$ into disjoint subsets $(\mathcal{X}_k)_{1 \leq k \leq K_n}$. Then, replace $\sigma^2 D_m$ by $\sum_{k=1}^{K_n} \sigma_k^2 D_{m,k}$, where σ_k and $D_{m,k}$ are local indexes for the noise and the complexity of S_m. Choosing K_n such that both K_n and $n K_n^{-1}$ go to infinity with n, we obtain a procedure that is (asymptotically) optimal in the histogram case if σ is Lipschitz with a finite number of jumps. In the least-square regression framework, Galtchouk and Pergamenschikov [**GP05**] defined another procedure that is minimax in the heteroscedastic case.

Those two procedures may perform a little better than resampling penalization. We call them "*ad hoc*" because they are specially designed for the heteroscedastic case and a particular family of estimators. On the contrary, Resampling Penalization is a general-purpose device. It was neither built to be adaptive to heteroscedastic noises, nor to take advantage of a specific model (regression, histograms).

When no information is available on the data, or when no known algorithm can make use of such informations, we suggest the use of RP. Moreover, it may happen that informations available are partial or wrong. Then, using an *ad hoc* procedure may be catastrophic, whereas a general device like RP would still work. In a nutshell, choose RP if you have no useful information or if you do not trust them.

Other model selection procedures by resampling. The most well-known resampling-based model selection procedure is cross-validation. For practical reasons, it is often used in its V-fold version, which may have some tricky behaviors, in particular when V has to be chosen [**Yan07**]. This can also be shown in our simulation experiments (Sect. 6.5, Tab. 6.2): in HSd1, $V = 2$ is better than $V \in \{5, 10, 20\}$. In Chap. 5, we explain this phenomenon by some bias of the V-fold criterion, that strongly depends on V. We also use Resampling Penalization for defining an alternative to V-Fold Cross-Validation which does not have this drawback. This enlightens one of the main benefit of using a penalization method like RP, that is flexibility.

There also exist some bootstrap model selection procedures [**Sha96, Shi97**]. As noticed in Remark 6.3, the ones studied in [**Shi97**] are quite close to RP, although stated in a less general form. In particular, they are restricted to Efron(n) weights, which have been shown to be the worst ones in our simulations. A second main improvement of RP is the use of a parameter C in front of the penalty. This allows to disconnect the choice of the weights from the overpenalization problem. For instance, inconsistency results of Shao [**Sha96**] with Efron(q) weights can be prevented by a wise choice of C, without changing q.

6.6.3. General frameworks, including classification. Our results on Resampling Penalization are restricted to the histogram case, so that we can wonder whether it still works in general frameworks. First, notice that algorithm 6.1 can be applied (maybe up to some little modifications, like for histograms) to any model selection problem. Then, it relies on the resampling idea, which is known to be quite robust in a wide variety of situations. Our Thm 6.1 and 6.2 show that RP is actually robust to heteroscedasticity in regression, whereas it has not been built for this. These are the main reasons why we expect RP to have such robustness properties in many other frameworks: least-square regression on general models, binary classification (*e.g.* with margin conditions), and so on.

In the classification case, there is another reason why RP should work. Indeed, when some margin condition holds (introduced by Mammen and Tsybakov [**MT99**]), penalization methods based upon global penalties such as Rademacher complexities are much too large, because they estimate

$$\mathrm{pen}_{\mathrm{id,g}}(m) := \sup_{t \in S_m} \{ (P - P_n)\gamma(t) \} \quad \text{instead of} \quad \mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m) \ .$$

This is no longer the case with localized penalties (*e.g.* local Rademacher complexities), that take into account the closeness of $\widehat{s}_m$ and s. Recent results [**LW04, BBM05, Kol06**] have shown that these localized penalties really estimate $\mathrm{pen}_{\mathrm{id}}(m)$ and not its global upper bound $\mathrm{pen}_{\mathrm{id,g}}(m)$, so that the resulting algorithms can benefit of the margin condition.

With RP, we precisely try to improve Fromont's bootstrap penalties [**Fro04**] by estimating $\mathrm{pen}_{\mathrm{id}}(m)$ instead of $\mathrm{pen}_{\mathrm{id,g}}(m)$. Then, RP can be considered as local penalties. One major drawback of local Rademacher complexities is their dependence on huge or unknown constants (see Sect. 2.2.1, page 77; in particular, they are not margin adaptive, as noticed by Koltchinskii [**Kol06**]). This calibration problem seems much easier for RP, since C can be chosen as described in Sect. 6.6.1. As a consequence, we can conjecture that RP (*e.g.* combined with the slope heuristics, as in Algorithm 11.1) is adaptive to the margin condition. A rigorous proof of this fact would of course be of much interest. We draw in Chap. 7 some possible ways towards such a proof.

Remark also that the computational cost of RP is much smaller that the one of local Rademacher complexities. Considering that the V-fold penalties introduced in Chap. 5 also belong to the RP family, we have built some local penalties which greatly improves on local Rademacher penalties from the computational viewpoint.

6.6.4. Conclusion. This chapter intends to help the practical user to answer the following question: when shall we use Resampling Penalization? To sum up, we list below the advantages and drawbacks of RP *vs.* the classical methods.

Advantages of RP.

- generality: well-defined in almost any framework.
- robustness and versatility: perfect for the cautious user.

- adaptivity: to several properties, *e.g.* heteroscedasticity and smoothness of the target.
- flexibility: possibility of overpenalization, either for non-asymptotic prediction or for identification.

Drawbacks of RP.

- computation time: one may prefer V-fold algorithms, VFCV or penVFCV (see Chap. 5).
- possibly suboptimal in easy cases (against Mallows' C_p) or in some particular frameworks (against *ad hoc* procedures).

6.7. Probabilistic tools: expectations of inverses

In this section, we give some results that may be interesting independently from the resampling penalization method. When computing the resampling penalty for some classical resampling schemes, the quantity

$$e_Z^+ = e_{\mathcal{L}(Z)}^+ := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\,\middle|\, Z>0\right]$$

appears (in $R_{1,W}$), for some random variables Z with Binomial, Poisson or Hypergeometric laws. The binomial case also appears for a comparison between

$$\mathbb{E}\left[P\left(\gamma\left(\widehat{s}_m\right)-\gamma\left(s_m\right)\right)\right] \quad \text{and} \quad \mathbb{E}\left[P_n\left(\gamma\left(s_m\right)-\gamma\left(\widehat{s}_m\right)\right)\right] \ .$$

Such quantities have been considered several times (for instance [**Lew76**] consider general Z, and [**JZ04, Žni05**] investigate the case of Binomial and Poisson random variables). However, these results are either asymptotic or too general to be accurate. In this section, we give some non-asymptotic bounds on e_Z^+, from which we can recover some of the well-known asymptotic results.

We first introduce another interesting quantity closely related to e_Z^+:

$$e_Z^0 = e_{\mathcal{L}(Z)}^0 := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z>0}\right] = e_Z^+\mathbb{P}(Z>0) \ . \tag{6.19}$$

It is useful to notice the following general lower bound, which is a straightforward consequence of Jensen's inequality: if $\mathbb{P}(Z>0)>0$, then

$$e_Z^+ \geq \mathbb{P}\left(Z>0\right) \ . \tag{6.20}$$

We used non-asymptotic concentration inequalities to derive the following upper bounds.

6.7.1. Binomial case.

LEMMA 6.1 (Lemma 5.3 in Sect. 5.6.1). *For any $n \in \mathbb{N}\backslash\{0\}$ and $p \in (0;1]$, $\mathcal{B}(n,p)$ denotes the binomial law with parameters (n,p). Denote $\kappa_3 = 5.1$ and $\kappa_4 = 3.2$. Then, if $np \geq 1$,*

$$\kappa_4 \wedge \left(1+\kappa_3(np)^{-1/4}\right) \geq e_{\mathcal{B}(n,p)}^+ \geq 1-e^{-np} \ . \tag{6.21}$$

As a consequence,

$$\lim_{np\to+\infty} e_{\mathcal{B}(n,p)}^+ = 1 \ . \tag{6.22}$$

Notice that when $p=1/2$, we can improve a little this result (see Lemma 8.14 in Sect. 8.7). This is useful for estimating $R_{1,\mathrm{Rad}}$.

6.7.2. Hypergeometric case. Recall that an hypergeometric random variable $X \sim \mathcal{H}(n,r,q)$ is defined by

$$\forall k \in \{0,\ldots,q\wedge r\}, \quad \mathbb{P}(X=k) = \frac{\binom{r}{k}\binom{n-r}{q-k}}{\binom{n}{q}} \ .$$

LEMMA 6.2. *Let $n,r,q \in \mathbb{N}$ such that $n \geq r \geq 1$ and $n \geq q \geq 1$.*

(1) *General lower-bound:*

$$e^+_{\mathcal{H}(n,r,q)} \geq 1 - \mathbb{1}_{r\leq n-q} \exp\left(-\frac{qr}{n}\right) \ . \tag{6.23}$$

(2) *General upper-bound: Let* $\epsilon \in (0;1)$ *and* $\kappa_5(\epsilon) = 0.9 + 1.4 \times \epsilon^{-2}$.

$$\textit{If} \qquad r \geq 2 \quad \textit{and} \quad \frac{n}{q} \leq (1-\epsilon)\frac{2r}{2+\sqrt{3(r+1)\ln(r)}}$$

$$\textit{Then,} \qquad e^+_{\mathcal{H}(n,r,q)} \leq 1 + \kappa_5(\epsilon)\frac{n}{q}\sqrt{\frac{\ln(r)}{r}} \ . \tag{6.24}$$

(3) *"Rho" case: if* $n \geq 2$,

$$\sup_{r\geq 1}\left\{e^+_{\mathcal{H}(n,r,\lfloor \frac{n}{2}\rfloor)}\right\} \leq 14.3 \qquad \textit{and} \qquad \sup_{r\geq 26}\left\{e^+_{\mathcal{H}(n,r,\lfloor \frac{n}{2}\rfloor)}\right\} \leq 3 \ . \tag{6.25}$$

(4) *"Loo" case:*

$$1 + \frac{\mathbb{1}_{r\geq 2}}{n(r-1)} \geq e^+_{\mathcal{H}(n,r,n-1)} = 1 + \frac{1}{n}\left(\frac{(n-1)r}{n(r-1)}\mathbb{1}_{r\geq 2} - 1\right) \geq 1 - \frac{\mathbb{1}_{r=1}}{n} \ . \tag{6.26}$$

(5) *"Lpo" case: if* $n \geq r \geq n-q+1 \geq 2$,

$$\frac{rn^{n-q}}{(r-n+q)n\cdots(q+1)} \geq e^+_{\mathcal{H}(n,r,q)} \geq 1 \ . \tag{6.27}$$

In particular, if $\sup_k \left\{ n_k q_k^{-1} \wedge (n_k - q_k)\right\} < +\infty$ and $n_k \geq r_k \to +\infty$, we have

$$\lim_{k\to+\infty}\left\{e^+_{\mathcal{H}(n_k,r_k,q_k)}\right\} = 1 \ . \tag{6.28}$$

6.7.3. Poisson case.

LEMMA 6.3. *For every* $\mu > 0$, $\mathcal{P}(\mu)$ *denotes the Poisson law with parameter* μ. *Then,*

$$\mathbb{1}_{\mu\geq 1.61} \vee \left(1 - e^{-\mu}\right) \leq e^+_{\mathcal{P}(\mu)} \leq \left(2 - 2e^{-2\mu}\right) \wedge \left(1 + \frac{2(1+e^{-3})}{(\mu-2)_+}\right) \ . \tag{6.29}$$

As a consequence,

$$\lim_{\mu\to+\infty} e^+_{\mathcal{P}(\mu)} = 1 \ . \tag{6.30}$$

6.7.4. Numerical illustration. For those three cases, we computed numerically e^+_Z for several values of the parameters. The results are given on Fig. 6.7 (binomial case), 6.8 (hypergeometric case) and 6.9 (Poisson case). It seems that the true absolute upper bounds are quite close to 1 (maybe lower than 1.4), and the asymptotic behaviour appears for rather small values of n.

6.8. Proofs

In the following, when we do not want to write explicitly some constants, we use the letter L. It means "some positive absolute constant, possibly different from a line to another, or even within the same line". When L is not numerical, but depends on some parameters $p_1, \ldots, p_k$, it is written $L_{p_1,\ldots,p_k}$ of $L(p_1,\ldots,p_k)$. When L depends on the constants that appear in a set (**A**) of assumptions, it is written $L_{(\mathbf{A})}$.

6.8.1. General framework. We first need to introduce some notations and assumptions.

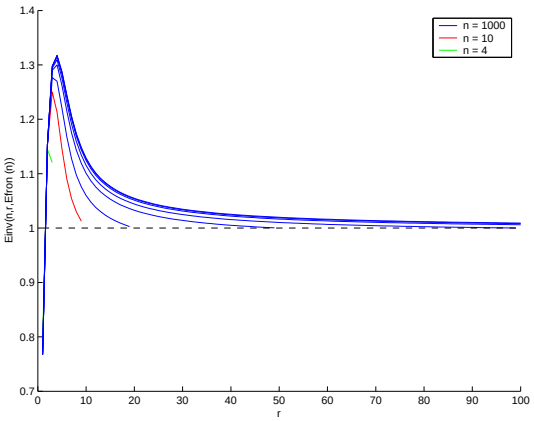


FIGURE 6.7. $e^+_{\mathcal{B}(n,p)}$ as a function of $r = np$, for $n \in \{4; 10; 20; 50; 100; 200; 10^3\}$.

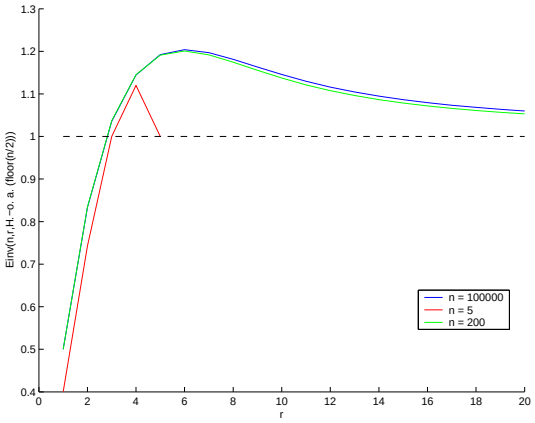


FIGURE 6.8. $e^+_{\mathcal{H}(n,\lfloor \frac{n}{2} \rfloor,r)}$ as a function of r, for $n \in \{5; 200; 10^5\}$.

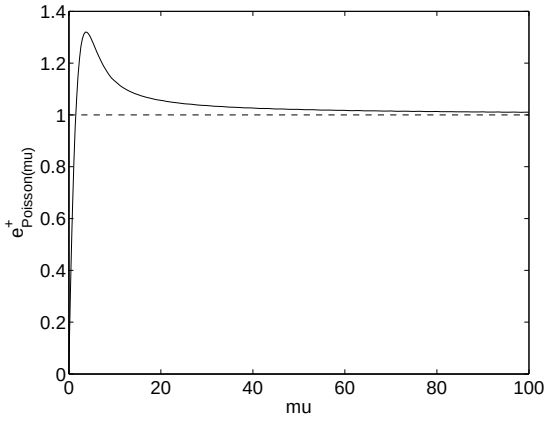


FIGURE 6.9. Inverse moment $e^+_{\mathcal{P}(\mu)}$ of a Poisson variable for $\mu \in [0; 100]$.

Notations. When $m \in \mathcal{M}_n$, $q > 0$ and Z is an arbitrary random variable, we define

$$p_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right) \qquad p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$$

$$m_{q,\lambda} := \|Y - s_m(X)\|_{q,\lambda} := \left(\mathbb{E}\left[|Y - s_m(X)|^q \mid X \in I_\lambda\right]\right)^{1/q}$$

$$\|Z\|_q^{(\Lambda_m)} := \mathbb{E}^{\Lambda_m}\left[|Z|^q\right]^{1/q} = \mathbb{E}\left[|Z|^q \mid \left(\mathbb{1}_{X_i \in I_\lambda}\right)_{1 \le i \le n,\, \lambda \in \Lambda_m}\right]^{1/q}$$

$$\text{and } \forall x \ge 0, \quad \varphi(x) = xe^{-x} \qquad \varphi_1(x) = \varphi(x \vee 1) \ .$$

In the histogram case, it is convenient to replace $p_2(m)$ by

$$\widetilde{p_2}(m) := p_2(m) + \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(\left(\sigma_\lambda^d\right)^2 + \left(\sigma_\lambda^r\right)^2 \right) \mathbb{1}_{n\widehat{p}_\lambda = 0} \ .$$

Then, if $\min_{\lambda \in \Lambda_m} \{n\widehat{p}_\lambda\} \ge 1$ (which will always be assumed in practice),

$$p_2(m) = \widetilde{p_2}(m) \qquad \text{and} \qquad \mathbb{E}^{\Lambda_m}\left[p_2(m)\right] = \mathbb{E}^{\Lambda_m}\left[\widetilde{p_2}(m)\right] = \mathbb{E}\left[\widetilde{p_2}(m)\right] \ .$$

Inside expectations, we will often write p_2 instead of $\widetilde{p_2}(m)$ by convention. When $\min_{\lambda \in \Lambda_m} \{np_\lambda\}$ is large, this does not make much difference.

In a similar way, we define $\widetilde{p_1}(m)$ as an alternative definition to $p_1(m)$ (see Sect. 5.7.2, where other conventions $\widetilde{p_1}^{(T)}$, $T \geq 0$, are also defined). This is more crucial than for $p_2(m)$, since $p_1(m)$ is not well-defined when $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} = 0$.

Bounded assumption set (**Bg**).

There is some noise: $\|\sigma(X)\|_2 > 0$.

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} \in [\sqrt{n}; c_{\mathrm{rich}}\sqrt{n}]$.

(**P3**) The constant C is well chosen: $\eta C_{W,\infty} \geq C \geq C_{W,\infty}$.

(**P4**) The weights are exchangeable, among Efr, Rad, Poi, Rho and Loo.

(**Ab**) Bounded data: $\|Y_i\|_\infty \leq A < \infty$.

($\mathbf{A_{m,\ell}}$) Local moment assumption: there exists $a_\ell, \xi_\ell \geq 0$ such that for every $q \geq 2$, for every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$,

$$P_m^\ell(q) := \frac{\sqrt{D_m \sum_{\lambda \in \Lambda_m} m_{q,\lambda}^4}}{\sum_{\lambda \in \Lambda_m} m_{2,\lambda}^2} \leq a_\ell q^{\xi_\ell} \ .$$

(**Ap**) Polynomial decreasing of the bias: there exists $\beta_1 \geq \beta_2 > 0$ and $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ such that, for every $m \in \mathcal{M}_n$,

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \leq l(s, s_m) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \ .$$

($\mathbf{A_Q}$) For every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$,

$$Q_m^{(p)} := \frac{n\mathbb{E}\left[p_2(m)\right]}{D_m} = \frac{1}{D_m} \sum_{\lambda \in \Lambda_m} \left[(\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2 \right] \geq c_Q^- > 0$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) Lower regularity of the partitions for $\mathcal{L}(X)$: there exists $c_{\mathrm{r},\ell}^X > 0$ such that for every $m \in \mathcal{M}_n$, $D_m \min_{\lambda \in \Lambda_m} p_\lambda \geq c_{\mathrm{r},\ell}^X$.

Unbounded assumption set (**Ug**). We remove (**Ab**) from (**Bg**), and add

($\mathbf{A}\sigma_{\max}$) Noise-level bounded from above: $\sigma^2(X) \leq \sigma_{\max}^2 < +\infty$ a.s.

($\mathbf{As}_{\max}$) Bound on the target function: $\|s\|_\infty \leq A$.

($\mathbf{A_{m,\ell}^+}$) Upper bound on the local moments: there exists $a_\ell^+, \xi_\ell^+ \geq 0$ such that for every $m \in \mathcal{M}_n$

$$\max_{\lambda \in \Lambda_m} \{ m_{q,\lambda} \} \leq a_\ell^+ q^{\xi_\ell^+} \ .$$

($\mathbf{A_{g,\epsilon}}$) Global moment assumption for the noise: there exists $a, \xi \geq 0$ such that for every $q \geq 2$,

$$P^{g\epsilon}(q) := \|\epsilon\|_q \leq a_{g\epsilon} q^{\xi_{g\epsilon}}$$

($\mathbf{A}\delta$) Global moment assumption for the bias: there is a constant $c_{\Delta,m}^g > 0$ such that, for every $m \in \mathcal{M}_n$ of dimension $D_m \geq D_0$,

$$\|s - s_m\|_\infty \leq c_{\Delta,m}^g \|s(X) - s_m(X)\|_2$$

General result.

LEMMA 6.4. *Let $n \in \mathbb{N} \backslash \{0\}$, $\gamma_0 > 0$ and $\widehat{m}$ given by algorithm 6.2. Assume that either* (**Bg**) *or* (**Ug**) *holds with constants independent from n.*

Then, there exists a constant K_1 (that depends on γ_0 and all the constants in (**Bg**) *(resp.* (**Ug**)*), but not on n) such that*

$$l(s, \widehat{s}_{\widehat{m}}) \leq \left[2\eta - 1 + \ln(n)^{-1/5} \right] \inf_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) \} \tag{6.31}$$

with probability at least $1 - K_1 n^{-\gamma_0}$.

The proof of this lemma is made in Sect. 6.8.5.

REMARK 6.7. In the infimum in (6.31), there may be some $m \in \mathcal{M}_n$ such that $\widehat{s}_m$ is not well defined. We took by convention $l(s, \widehat{s}_m) = \infty$ in those cases.

From the proof, there is a constant $c > 0$ (that depend on $\alpha_{\mathcal{M}}$, γ_0 and $c^X_{\mathrm{r},\ell}$) such that every model of dimension smaller than $cn\,(\ln(n))^{-1}$ belongs to $\widehat{\mathcal{M}}_n$ on the event where (6.31) holds. For each of these models,

$$l(s, \widehat{s}_m) = l(s, s_m) + \widetilde{p_1}^{(0)}(m) = l(s, s_m) + \widetilde{p_1}(m) = l(s, s_m) + \widetilde{p_1}^{(A^{-1})}(m)$$

so that we can then restrict the infimum to models of dimension lower than $cn\,(\ln(n))^{-1}$ with any of these conventions for $l(s, \widehat{s}_m)$.

6.8.2. Proof of Thm. 6.1. We apply Lemma 6.4 with $\gamma_0 = 2$. In order to deduce (6.11), it remains to show that $(\mathbf{A_{m},}_\ell)$ and $(\mathbf{A_Q})$ are satisfied. This is true with $D_0 = 1$ since for every $m \in \mathcal{M}_n$,

$$P^\ell_m(q) := \frac{\sqrt{D_m \sum_{\lambda\in\Lambda_m} m^4_{q,\lambda}}}{\sum_{\lambda\in\Lambda_m} m^2_{2,\lambda}} \le \frac{\|Y - s_m(X)\|^2_\infty}{\min_{\lambda\in\Lambda_m}\left\{\left(\sigma^r_\lambda\right)^2\right\}} \le \frac{4A^2}{\sigma^2_{\min}}$$

$$Q^{(p)}_m := \frac{1}{D_m}\sum_{\lambda\in\Lambda_m}\left[(\sigma^r_\lambda)^2 + \left(\sigma^d_\lambda\right)^2\right] \ge \sigma^2_{\min}\ .$$

Let Ω_n be the event on which (6.11) holds true. Then,

$$\begin{aligned}\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] &= \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_n}\right] + \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega^c_n}\right] \\ &\le \left[2\eta - 1 + \epsilon_n\right]\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\{l(s, \widehat{s}_m)\right\}\right] + A^2 K_1 \mathbb{P}\left(\Omega^c_n\right)\end{aligned}$$

which proves (6.12).

Remark that (6.12) also holds with $\mathcal{M}_n$ replaced by

$$\left\{m \in \mathcal{M}_n \text{ s.t. } D_m \le c(\alpha_{\mathcal{M}}, c^X_{\mathrm{r},\ell}) n \ln(n)^{-1}\right\}$$

and the convention $p_1(m) = \widetilde{p_1}^{(0)}(m)$.

6.8.3. Proof of Thm. 6.1: alternative assumptions.

Without (**An**). When $\sigma(X)$ is allowed to be zero, we only need another proof for $(\mathbf{A_{m},}_\ell)$ and $(\mathbf{A_Q})$.

$$P^\ell_m(q) = \frac{\sqrt{\sum_{\lambda\in\Lambda_m} m^4_{q,\lambda}}}{\sqrt{D_m} Q^{(p)}_m} \le \frac{\|Y - s_m(X)\|^2_\infty}{Q^{(p)}_m} \le \frac{4A^2}{Q^{(p)}_m}$$

$$Q^{(p)}_m \ge \frac{\|\sigma\|^2_{L^2(\mathrm{Leb})}}{2c_{\mathrm{r,u}}} - \frac{K^2_\sigma\left(c^d_{\mathrm{r,u}}\right)^2 \mathrm{diam}(\mathcal{X})^2}{D^{2\alpha_d}_m} - \frac{J_\sigma \|\sigma(X)\|^2_\infty}{2D_m} \quad (\textit{cf.}\ \text{Lemma 6.13}).$$

Thus, $(\mathbf{A_{m},}_\ell)$ and $(\mathbf{A_Q})$ hold true uniformly on models $m \in \mathcal{M}_n$ such that $D_m \ge D_0 = L_{(\mathbf{Bg})}$.

Unbounded case. We still use Lemma 6.4, but the proof is a little longer.

Pathwise oracle inequality. We prove it for a general γ_0 (since we need it for the classical oracle below). We have to prove $(\mathbf{A_{m},}_\ell)$, $(\mathbf{A_Q})$, $(\mathbf{A^+_{m},}_\ell)$, $(\mathbf{A_{g},}_\epsilon)$ and $(\mathbf{A}\delta)$. The four first ones are

almost straightforward: for every $m \in \mathcal{M}_n$,

$$P_m^\ell(q) = \frac{\sqrt{\sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4}}{\sqrt{D_m} Q_m^{(p)}} \le \frac{\left(2A + c_{\text{gauss}}\sqrt{q}\sigma_{\max}\right)^2}{Q_m^{(p)}} \le \frac{L_{c_{\text{gauss}},\sigma_{\max},A} q^2}{Q_m^{(p)}}$$
$$Q_m^{(p)} \ge \sigma_{\min}^2$$
$$\max_{\lambda\in\Lambda_m} \{m_{q,\lambda}\} \le \|s - s_m\|_\infty + \sigma_{\max} c_{\text{gauss}}\sqrt{q} \le \left(2A + c_{\text{gauss}}\sigma_{\max}\right)\sqrt{q}$$
$$P^{g\epsilon}(q) \le \sigma_{\max} c_{\text{gauss}}\sqrt{q} \ .$$

For the last one, we use Lemma 6.14 (with (**Al**), ($\mathbf{Ar}_{\ell,\mathbf{u}}$) and ($\mathbf{Ad}_\ell$)) which shows that

$$c_{\Delta,m}^g \le L_{(\mathbf{Ug})} \qquad \text{if} \quad D_m \ge D_0 = L_{(\mathbf{Ug})} \ .$$

Classical oracle inequality. Let Ω_n be the event on which (6.11) holds true with $\gamma_0 = 6 + \alpha_{\mathcal{M}}$. As in the bounded case, we only have to upper bound

$$\mathbb{E}^{\Lambda_m}\left[l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_n^c}\right] \le \sqrt{\mathbb{P}(\Omega^c)}\sqrt{\mathbb{E}^{\Lambda_m}\left[l(s,\widehat{s}_{\widehat{m}})^2\right]} \qquad \text{by Cauchy-Schwartz}$$
$$\le \sqrt{K_1} n^{-\gamma_0/2}\sqrt{\mathbb{E}^{\Lambda_m}\left[2\|s\|_\infty^2 + 2p_1(\widehat{m})^2\right]}$$
$$\le L_{(\mathbf{Ug})} n^{-\gamma_0/2}\left[1 + \sqrt{\mathbb{E}^{\Lambda_m}\left[\sum_{m\in\mathcal{M}_n} p_1(m)^2 \mathbb{1}_{m\in\widehat{\mathcal{M}}_n}\right]}\right] \ .$$

For every $m \in \widehat{\mathcal{M}}_n$, we have to compute $\mathbb{E}^{\Lambda_m}[(p_1(m))^2]$ (and derive a bound on it, even very poor). Starting from (5.19) in Sect. 5.7.2, we have

$$\mathbb{E}^{\Lambda_m}\left[p_1(m)^2\right] = \frac{1}{n^2}\sum_{\lambda\in\Lambda_m}\left(\frac{p_\lambda}{\widehat{p}_\lambda}\right)^2 \mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] + \frac{1}{n^2}\sum_{\lambda\neq\lambda'}\left[\frac{p_\lambda}{\widehat{p}_\lambda}\frac{p_{\lambda'}}{\widehat{p}_{\lambda'}} m_{2,\lambda}^2 m_{2,\lambda'}^2\right]$$
$$\le \sum_{\lambda\in\Lambda_m}\mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] + \sum_{\lambda\neq\lambda'}\left(\sigma_{\max}^2 + (2A)^2\right)^2 \le D_m^2 L_{(\mathbf{Ug})} \le n^2 L_{(\mathbf{Ug})}$$

since

$$\mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] = \mathbb{E}^{\Lambda_m}\left[\frac{\left(\sum_{X_i\in I_\lambda}(Y_i - \beta_\lambda)\right)^4}{(n\widehat{p}_\lambda)^2}\right] = \frac{m_{4,\lambda}^4}{n\widehat{p}_\lambda} + \frac{6(n\widehat{p}_\lambda - 1)m_{2,\lambda}^4}{n\widehat{p}_\lambda}$$
$$\text{and} \quad D_m \sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4 \le \left(a_\ell q^{\xi_\ell}\right)^2 \left(\sigma_{\max}^2 + (2A)^2\right)^2 \ .$$

Using that $\text{Card}(\mathcal{M}_n) \le c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$, we obtain

$$\mathbb{E}^{\Lambda_m}\left[l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_n^c}\right] \le L_{(\mathbf{Ug})} n^{1+(\alpha_{\mathcal{M}}-\gamma_0)/2}$$

which proves (6.12).

6.8.4. Proof of Thm. 6.2.

Non-constant regression function. First assume that s is non-constant, *i.e.* $\text{varia}_{\mathcal{X}}\, s > 0$. We start from the classical oracle inequality of Thm. 6.1, with $\mathcal{M}_n$ reduced to models such that $D_m \le cn\ln(n)^{-1}$ (*cf.* the remark at the end of the proof of Thm. 6.1). This is possible since (**P2**) holds with $c_{\text{rich}} = 2^k$, ($\mathbf{Ar}_\ell^{\mathbf{X}}$) holds with $c_{\text{r},\ell}^X = c_{\min}^X$ and (**Ap**) holds with

$$C_{\text{b}}^- = L(\alpha, R, \text{diam}(\mathcal{X}), k, \text{varia}_{\mathcal{X}}(s)) > 0 \qquad C_{\text{b}}^+ = R^2$$

$$\beta_1 = k^{-1} + \alpha^{-1} - (k-1)k^{-1}\alpha \qquad \beta_2 = 2\alpha k^{-1}$$

(see Lemma 8.20 in Sect. 8.10). Thus,

$$\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right] \leq \left[2\eta - 1 + \epsilon_n\right] \inf_{m\in\mathcal{M}_n, D_m \leq \frac{cn}{\ln(n)}} \left\{ l(s,s_m) + \mathbb{E}\left[\widetilde{p_1}^{(0)}(m)\right]\right\} + A^2 K_1 n^{-2} \ .$$

Let $T \in \mathbb{N}$ and $m = m(T) \in \mathcal{M}_n$ be the model of dimension $D_m \approx \mathrm{Leb}(\mathcal{X})T^k$. Then,

$$\begin{aligned} l(s,s_m) &\leq R^2 T^{-2\alpha} \\ \mathbb{E}\left[\widetilde{p_1}^{(0)}(m)\right] &\leq \sup_{np\geq 0} e^0_{\mathcal{B}(n,p)} \times \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left((\sigma^r_\lambda)^2 + \left(\sigma^d_\lambda\right)^2\right) \\ &\leq \frac{2}{n}\left(R^2T^{1-2\alpha} + \sum_{\lambda\in\Lambda_m}(\sigma^r_\lambda)^2\right) \leq \frac{2R^2T^{1-2\alpha}}{n} + \frac{2\sigma^2_{\max}D_m}{n} \end{aligned}$$

where the bound on $e^0_{\mathcal{B}(n,p)}$ comes from (5.60).

We now take

$$T^k \approx R^{\frac{2k}{2\alpha+k}} n^{\frac{k}{2\alpha+k}} \sigma_{\max}^{\frac{-2k}{2\alpha+k}}$$

which is smaller than $cn(\ln(n))^{-1}$ if $n \geq L_{k,\alpha,\sigma_{\max},c}$. Since $\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right] \leq A^2$, we obtain (6.13) by choosing $K_3 = K_1 \vee (A^2 L_{k,\alpha,\sigma_{\max},c})$.

When $(\mathbf{A}\sigma)$ holds, when $m = m(T)$, for every $\lambda \in \Lambda_m$ such that there is no jump of σ on I_λ,

$$\begin{aligned} (\sigma^r_\lambda)^2 \leq \max_{I_\lambda}\sigma^2 &\leq \left(\frac{K_\sigma}{T} + \sqrt{\frac{1}{\mathrm{Leb}(\mathcal{X})}\int_{\mathcal{X}}\sigma^2(t)\,\mathrm{Leb}(dt)}\right)^2 \\ &\leq \left(1+\theta^{-1}\right)\frac{K_\sigma^2}{T^2} + \frac{1+\theta}{\mathrm{Leb}(\mathcal{X})}\int_{\mathcal{X}}\sigma^2(t)\,\mathrm{Leb}(dt) \end{aligned}$$

for every $\theta > 0$.

If σ jumps on I_λ (and there are at most J_σ such λ), we simply bound $\max_{I_\lambda}\sigma^2$ by $\sigma^2_{\max}$. As a consequence, taking $\theta = T^{-1}$, we get

$$\begin{aligned} \mathbb{E}\left[\widetilde{p_1}^{(0)}(m(T))\right] &\leq \frac{2}{n}\left(R^2T^{1-2\alpha} + \sum_{\lambda\in\Lambda_m}(\sigma^r_\lambda)^2\right) \\ &\leq \frac{2R^2T^{1-2\alpha}}{n} + \frac{2D_m\left\|\sigma\right\|^2_{L^2(\mathrm{Leb})}}{n} + \frac{L_{(\mathbf{Bg})}}{n} \end{aligned}$$

and the end of the proof does not change.

Constant functions. The case of constant functions has to be considered separately, since $(\mathbf{Ap})$ is no longer satisfied. Indeed, constant functions are too well approximated by histograms and thus lead to select models of rather small dimension (the optimal dimension being one). We will show that they are estimated at the rate $(\ln(n))^\chi n^{-1}$ for some $\chi > 0$, which is faster than $n^{-\beta}$ for any $\beta < 1$.

Upper bound on $D_{\widehat{m}}$. Consider Ω_{n,γ_0} defined in the proof of Lemma 6.4 (see Sect. 6.8.5) with $\gamma_0 = 2$. As in the proof of (6.34), we have for every $m \in \widehat{\mathcal{M}}_n$ such that $D_m \geq \ln(n)^{\xi+1}$,

$$\begin{aligned} \mathrm{crit}(m) &= \mathrm{pen}(m) - p_2(m) \\ &\geq \left(\frac{1}{4} - L_{(\mathbf{Bg})}\ln(n)^{-\xi}\right) Q_m^{(p)} \ln(n)^{\xi+1} n^{-1} \end{aligned}$$

$$\text{and} \quad \overline{\delta}(m) \geq \frac{-L_A \gamma \ln(n)}{n} \ .$$

Let m_1 be the model of dimension 1 (it belongs to $\widehat{\mathcal{M}}_n$ on Ω_{n,γ_0}). Then,

$$\mathrm{crit}(m_1) = \mathrm{pen}(m_1) - p_2(m_1) \leq L_{(\mathbf{Bg})} \ln(n)^{\xi_\ell+1} \mathbb{E}\left[p_2(m_1)\right] \leq L_{(\mathbf{Bg})} \ln(n)^{\xi} \sigma_{\max} n^{-1} \ .$$

Then, if $n \geq L_{(\mathbf{Bg})}$, $D_{\widehat{m}} \leq \ln(n)^{\xi+1}$.

We now have

$$\begin{aligned}
\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right] &= \mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_{n,\gamma_0}}\right] + \mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\Omega_{n,\gamma_0}^c}\right] \\
&\leq \sum_{1\leq D_m \leq \ln(n)^{\xi+1}} \mathbb{E}\left[\widetilde{p_1}^{(0)}(m)\mathbb{1}_{\Omega_{n,\gamma_0}}\right] + A^2 \mathbb{P}(\Omega_{n,\gamma_0}^c) \\
&= \sum_{1\leq D_m \leq \ln(n)^{\xi+1}} \frac{L D_m \left(A^2 + \sigma_{\max}^2\right)}{n} + K_1 A^2 n^{-2} \\
&\leq L_{(\mathbf{Bg})} \ln(n)^{\xi+1} n^{-1} \leq R^{\frac{2k}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \sigma_{\max}^{\frac{4\alpha}{2\alpha+k}} + L_{(\mathbf{Bg})} n^{-1} \ .
\end{aligned}$$

if $n \geq L_{(\mathbf{Bg})}$. Otherwise, we enlarge the constant K_3 so that $K_3 n^{-1} \geq A^2$.

6.8.5. Proof of Lemma 6.4. We first give the complete proof in the bounded case. Then, we will explain how it can be extended to the unbounded case.

Bounded case. For each $m \in \mathcal{M}_n$, we have

$$l(s,\widehat{s}_m) = P_n\gamma(\widehat{s}_m) + p_1(m) + p_2(m) - \delta(m) - P\gamma(s) \ .$$

By definition of $\widehat{m}$, for every $m \in \widehat{\mathcal{M}}_n$,

$$P_n\gamma(\widehat{s}_{\widehat{m}}) + \mathrm{pen}(\widehat{m}) \leq P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \ ,$$

so that

$$l(s,\widehat{s}_{\widehat{m}}) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m})\right) \leq l(s,\widehat{s}_m) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m)\right) \ . \tag{6.32}$$

with $\mathrm{pen}'_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}(m) + (P - P_n)\gamma(s)$.

Let $\gamma = \gamma_0 + \alpha_{\mathcal{M}}$. For every $m \in \mathcal{M}_n$, define

$$A_n(m) = \min_{\lambda\in\Lambda_m} \{n\widehat{p}_\lambda\} \qquad \text{and} \qquad B_n(m) = \min_{\lambda\in\Lambda_m} \{np_\lambda\} \ .$$

We now define the event Ω_{n,γ_0} on which the following hold:

- for every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$, $A_n(m) \geq 1$ and $B_n(m) \geq 1$:

$$\left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right| \leq L_{\gamma_0,(\mathbf{Bg})} \frac{\ln(n)^{\xi_\ell+1}}{\sqrt{A_n(m) D_m}} \mathbb{E}\left[p_2(m)\right] \tag{6.57}$$

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{\gamma_0,(\mathbf{Bg})} \left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right] \mathbb{E}\left[p_2(m)\right] \tag{6.58}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{\gamma_0,(\mathbf{Bg})} \left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + \sqrt{D_m} e^{-LB_n}\right] \mathbb{E}\left[p_2(m)\right] \tag{6.59}$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{\gamma_0,(\mathbf{Bg})} \frac{\ln(n)^{\xi_\ell+1}}{\sqrt{D_m}} \mathbb{E}\left[p_2(m)\right] \tag{6.60}$$

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s,s_m)}{\sqrt{D_m}} + L_{\gamma_0,(\mathbf{Bg})} \frac{\ln(n)}{\sqrt{D_m}} \mathbb{E}\left[p_2(m)\right] \tag{6.62}$$

- for every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$, $A_n(m) \geq 1$ and $B_n(m) > 0$:

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1) B_n(m)^{-1} \ln(n)} - \frac{L(a_\ell, \xi_\ell, \gamma) \ln(n)^{\xi_\ell + 2}}{\sqrt{D_m}} \right) \mathbb{E}\left[\widetilde{p_2}(m)\right] \tag{6.61}$$

- for every $m \in \mathcal{M}_n$,

$$A_n(m) \geq \frac{B_n(m)}{2} - 2(\gamma+1)\ln(n) \tag{6.67}$$

From Prop. 6.9 (for $\widetilde{p_1}$ and p_2), Lemma 6.10 (for $\overline{\delta}(m)$), Prop. 6.8 (for pen), Lemma 6.12 (for $A_n(m)$), we have

$$\mathbb{P}\left(\Omega_{n,\gamma_0}\right) \geq 1 - L \sum_{m \in \mathcal{M}_n} n^{-\gamma_0 - \alpha_{\mathcal{M}}} \geq 1 - L(c_{\mathcal{M}}) n^{-\gamma_0} \ .$$

If $D_m \leq L_{\gamma, c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}$, then, $(\mathbf{Ar}^{\mathbf{X}}_\ell)$ implies that $B_n(m) \geq L^{-1} \vee (1 + 4(\gamma+1)) \ln(n)$. As a consequence, on Ω_{n,γ_0}, for every $m \in \mathcal{M}_n$ such that $D_0 \vee (\ln(n))^{2\xi_\ell + 7} \leq D_m \leq L_{\gamma, c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}$:

$$\begin{aligned} \max\left\{ \left|\widetilde{p_1}(m) - \mathbb{E}\left[\widetilde{p_1}(m)\right]\right|, \left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right|, \left|\overline{\delta}(m)\right|, \left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \right\} \\ \leq \frac{L_{\gamma_0,(\mathbf{Bg})} \mathbb{E}\left[l(s, s_m) + p_2(m)\right]}{\ln(n)} \end{aligned}$$

and $A_n(m) \geq \ln(n)$. Using Prop. 6.5 (and the non-asymptotic bounds given within its proof, since $\min_{\lambda \in \Lambda_m} \{ n \widehat{p}_\lambda \} \geq \ln(n)$) and (6.10) (*cf.* Lemma 5.6 in Sect. 5.7.2),

$$\begin{aligned} \left(2 - L\ln(n)^{-1/4}\right) \mathbb{E}\left[p_2(m)\right] \leq \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] \leq \left(2\eta + L_\eta \ln(n)^{-1/4}\right) \mathbb{E}\left[p_2(m)\right] \\ \left(1 - Ln^{-1}\right) \mathbb{E}\left[\widetilde{p_1}(m)\right] \leq \mathbb{E}\left[p_2(m)\right] \leq \left(1 + L\ln(n)^{-1/4}\right) \mathbb{E}\left[\widetilde{p_1}(m)\right] \ . \end{aligned}$$

We deduce: if $n \geq L_{\gamma_0,(\mathbf{Bg})}$, for every $m \in \mathcal{M}_n$ such that $\ln(n)^{2\xi_\ell+7} \leq D_m \leq L_{\gamma, c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}$, on Ω_{n,γ_0},

$$\frac{-L_{\gamma_0,(\mathbf{Bg})}}{\ln(n)^{1/4}} p_1(m) \leq (\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}})(m) \leq \left[2(\eta - 1) + \frac{L_{\gamma_0,(\mathbf{Bg})}}{\ln(n)^{1/4}} \right] p_1(m) \ .$$

We need to assume that n is large enough in order to upper bound $\mathbb{E}\left[p_2(m)\right]$ in terms of $p_1(m)$, since we only have

$$p_1(m) \geq \left[1 - \frac{L_{\gamma_0,(\mathbf{Bg})}}{\ln(n)} \right]_+ \mathbb{E}\left[p_2(m)\right]$$

in general.

Combined with (6.32), this gives: if $n \geq L_{\gamma_0,(\mathbf{Bg})}$,

$$\begin{aligned} l(s, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\ln(n)^{2\xi_\ell+7} \leq D_{\widehat{m}} \leq L_{\gamma, c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}} \leq \left[2\eta - 1 + \frac{L_{\gamma_0,(\mathbf{Bg})}}{\ln(n)^{1/4}} \right] \\ \times \inf_{m \in \mathcal{M}_n \text{ s.t. } \ln(n)^{2\xi_\ell+7} \leq D_m \leq L_{\gamma, c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}} \left\{ l(s, \widehat{s}_m) \right\} \ . \end{aligned} \tag{6.33}$$

Define the oracle model $m^\star \in \arg\min \{ l(s, \widehat{s}_m) \}$. We prove below that for any $c > 0$, if $n \geq L_{\gamma_0,(\mathbf{Bg}),c}$, then, on an event Ω'_{n,γ_0} of probability at least $1 - L(c_{\mathcal{M}}) n^{-\gamma_0}$,

$$\ln(n)^\xi \leq D_{\widehat{m}} \leq cn \ln(n)^{-1} \tag{6.34}$$

$$\ln(n)^\xi \leq D_{m^\star} \leq cn \ln(n)^{-1} \qquad \text{with} \quad \xi = 2\xi_\ell + 7 \ . \tag{6.35}$$

The result follows since $L_{\gamma_0,(\mathbf{Bg})} \ln(n)^{-1/4} \leq \epsilon_n = \ln(n)^{-1/5}$ for $n \geq L_{\gamma_0,(\mathbf{Bg})}$. We finally remove the condition $n \geq n_0 = L_{\gamma_0,(\mathbf{Bg})}$ by choosing $K_1 = L_{\gamma_0,(\mathbf{Bg})}$ such that $K_1 n_0^{-\gamma} \geq 1$.

Bounded case: control of $D_{\widehat{m}}$ and $D_{m^\star}$. We first state three additional concentration inequalities:

(1) From the proof of Prop. 6.9 (*cf.* Sect. 5.7.4), we have for every $m \in \mathcal{M}_n$ and $q \geq 2$,

$$\left\| p_2(m) - \mathbb{E}^{\Lambda_m}[p_2(m)] \right\|_q^{(\Lambda_m)} \leq \frac{Lq}{n} \sqrt{\sum_{\lambda \in \Lambda_m} m_{2q,\lambda}^4} \leq \frac{L\sqrt{D_m}q}{n} \max_{\lambda \in \Lambda_m} \left\{ m_{2q,\lambda}^2 \right\} \leq \frac{L_A \sqrt{D_m} q}{n} \ .$$

Hence, on an event of probability at least $1 - Ln^{-\gamma}$,

$$p_2(m) \leq \mathbb{E}^{\Lambda_m}\left[p_2(m)\right] + \frac{L_{A,\gamma}\sqrt{D_m}\ln(n)}{n} \leq \frac{L_{A,\gamma} D_m \ln(n)}{n} \ . \tag{6.36}$$

(2) From (6.63) with $\eta = \sqrt{\ln(n)/n}$, for every $m \in \mathcal{M}_n$, there is an event of probability at least $1 - Ln^{-\gamma}$ on which

$$\left|\overline{\delta}(m)\right| \leq \left(l(s, s_m) + L\gamma A^2 \right) \sqrt{\frac{\ln(n)}{n}} \leq L_{A,\gamma} \sqrt{\frac{\ln(n)}{n}} \ . \tag{6.37}$$

We then define Ω'_{n,γ_0} the subset of Ω_{n,γ_0} on which (6.36) and (6.37) hold for every $m \in \mathcal{M}_n$. Since $\gamma = \gamma_0 + \alpha_{\mathcal{M}}$ and $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$, $\mathbb{P}(\Omega'_{n,\gamma_0}) \geq 1 - L(c_{\mathcal{M}}) n^{-\gamma_0}$.

PROOF OF (6.34). By definition, $\widehat{m}$ minimizes $\mathrm{crit}(m)$ over $\widehat{\mathcal{M}}_n$. It thus also minimize

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = l(s, s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $\widehat{\mathcal{M}}_n$.

(1) Lower bound on $\mathrm{crit}'(m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^\xi$. We then have, on Ω'_{n,γ_0},

$$\begin{aligned} l(s, s_m) &\geq C_{\mathrm{b}}^- \left(\ln(n)\right)^{-\beta_1 \xi} && \text{by } (\mathbf{Ap}) \\ \mathrm{pen}(m) &\geq 0 \\ p_2(m) &\leq L_{\gamma_0,(\mathbf{Bg})} \frac{\ln(n)^{\xi+1}}{n} && \text{by (6.36)} \\ \overline{\delta}(m) &\geq -L_{\gamma_0,(\mathbf{Bg})} \sqrt{\frac{\ln(n)}{n}} && \text{by (6.37)} \end{aligned}$$

so that

$$\mathrm{crit}'(m) \geq L_{(\mathbf{Bg}),\gamma_0} \left(\ln(n)\right)^{-L(\beta_1,\xi)} \qquad \text{if} \quad n \geq L_{\gamma_0,(\mathbf{Bg})} \ .$$

(2) Lower bound for large models: let $m \in \widehat{\mathcal{M}}_n$ such that $D_m > cn\left(\ln(n)\right)^{-1}$. Since $A_n(m) \geq 3$, Lemma 6.6 shows that

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m) - p_2(m)\right] \geq \frac{\mathbb{E}\left[p_2(m)\right]}{4} \ .$$

Then, on Ω'_{n,γ_0}, (6.57), (6.60) and (6.62) imply

$$\begin{aligned} \mathrm{pen}(m) - p_2(m) &\geq \left(\frac{1}{4} - L_{\gamma_0,(\mathbf{Bg}),c} n^{-1/4} \right) \mathbb{E}\left[p_2(m)\right] \\ &\geq L_{c,c_Q^-} \ln(n)^{-1} \quad \text{when} \quad n \geq L_{\gamma_0,(\mathbf{Bg}),c} \\ \text{and} \qquad \overline{\delta}(m) &\geq -L_{(\mathbf{Bg}),c} \sqrt{\frac{\ln(n)}{n}} \ , \end{aligned}$$

so that

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) + \overline{\delta}(m) - p_2(m) \geq L_{c,(\mathbf{Bg})} \ln(n)^{-1}$$

when $n \geq L_{\gamma_0,(\mathbf{Bg}),c}$.

(3) There exists a better model for $\mathrm{crit}(m)$: From (**P2**), there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If moreover $n \geq L_{c_{\mathrm{rich}},c,\xi}$,

$$\ln(n)^{\xi} \leq \sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n} \leq \frac{cn}{\ln(n)} \quad .$$

Using (**Ap**),

$$l(s, s_{m_0}) \leq C_{\mathrm{b}}^{+} n^{-\beta_2/2}$$

so that, when $n \geq L_{\gamma_0,(\mathbf{Bg})}$,

$$\begin{aligned}\mathrm{crit}'(m_0) &\leq l(s, s_{m_0}) + \left|\overline{\delta}(m)\right| + \mathrm{pen}(m) \\ &\leq L_{\gamma_0,(\mathbf{Bg})} \left(n^{-\beta_2/2} + n^{-1/2} \right) \quad .\end{aligned}$$

If $n \geq L_{\gamma_0,(\mathbf{Bg}),c}$, this upper bound is smaller than the previous lower bounds for small and large models.

□

PROOF OF (6.35). Recall that $m^\star$ minimizes $l(s, \widehat{s}_m) = l(s, s_m) + p_1(m)$ over $m \in \mathcal{M}_n$, with the convention $l(s, \widehat{s}_m) = \infty$ if $A_n(m) = 0$.

(1) Lower bound on $l(s, \widehat{s}_m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^{\xi}$. From (**Ap**), we have

$$l(s, \widehat{s}_m) \geq l(s, s_m) \geq C_{\mathrm{b}}^{-} \left(\ln(n)\right)^{-\beta_1 \xi} \quad .$$

(2) Lower bound on $l(s, \widehat{s}_m)$ for large models: let $m \in \mathcal{M}_n$ such that $D_m > cn(\ln(n))^{-1}$ and $A_n(m) = \min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} \geq 1$. From (6.61), for $n \geq L_{\gamma,(\mathbf{Bg}),c}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)\left(c_{\mathrm{r},\ell}^{X}\right)^{-1} \ln(n)} - L(a_\ell, \xi_\ell, \gamma, c) n^{-1/4} \right) \mathbb{E}\left[\widetilde{p_2}(m)\right] \geq \frac{L_{(\mathbf{Bg}),c}}{\ln(n)^2}$$

$$\text{so that} \quad l(s, \widehat{s}_m) \geq L_{(\mathbf{Bg}),c} \ln(n)^{-2} \quad .$$

(3) There exists a better model for $l(s, \widehat{s}_m)$: let $m_0 \in \mathcal{M}_n$ be as in the proof of (6.34) and assume that $n \geq L_{c_{\mathrm{rich}},c,\xi}$. Then,

$$p_1(m_0) \leq L_{(\mathbf{Bg}),\gamma_0} \mathbb{E}\left[p_2(m)\right] \leq L_{(\mathbf{Bg}),\gamma_0} n^{-1/2}$$

and the arguments of the previous proof show that

$$l(s, \widehat{s}_{m_0}) \leq L_{(\mathbf{Bg}),\gamma_0} \left(n^{-\beta_2/2} + n^{-1/2} \right)$$

which is smaller than the previous upper bounds for $n \geq L_{(\mathbf{Bg}),\gamma_0,c}$.

□

Unbounded case. The proof of the bounded case has to be slightly modified. In the definition of Ω_{n,γ_0}, we replace (6.62) by

$$\left|\overline{\delta}(m)\right| \leq \frac{L\left(a_{g\epsilon}, \xi_{g\epsilon}, c_{\Delta,m}^{g}\right) x^{\xi_{g\epsilon}+1/2}}{\sqrt{D_m}} \left[l(s, s_m) + \frac{\sigma_{\max}^2}{Q_m^{(p)}} \mathbb{E}[p_2(m)] \right] \tag{6.64}$$

which holds with probability at least $1 - Ln^{-\gamma}$ because of Lemma 6.11.

In the proof of (6.34), we replace (6.36) by

$$p_2(m) \leq \mathbb{E}^{\Lambda_m}\left[p_2(m)\right] + \frac{L_{a_\ell^+,\xi_\ell^+,\gamma}\sqrt{D_m}\ln(n)^{2\xi_\ell^+ +1}}{n} \leq \frac{L_{(\mathbf{Ug})}D_m\ln(n)^{2\xi_\ell^+ +1}}{n} \ . \tag{6.38}$$

To prove that (6.38) holds with the same probability, we use $(\mathbf{A}^+_{\mathbf{m},\ell})$ instead of $(\mathbf{Ab})$ to upper bound

$$\max_{\lambda\in\Lambda_m}\left\{m^2_{2q,\lambda}\right\} \leq \left(a^+_{loc}\right)^2 2^{2\xi_\ell^+} q^{2\xi_\ell^+}$$

and $(\mathbf{A}\sigma_{\max})$ and $(\mathbf{As}_{\max})$ instead of $(\mathbf{Ab})$ to upper bound

$$\mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \leq \frac{D_m\left(\sigma^2_{\max}+4A^2\right)}{n} \ .$$

Then, we replace (6.64) by

$$\left|\overline{\delta}(m)\right| \leq L\left(a_{g\epsilon},\xi_{g\epsilon},A,\sigma_{\max},\gamma\right)\ln(n)^{\xi_{g\epsilon}+1/2} \tag{6.65}$$

which comes from Lemma 6.11 (and does not use $(\mathbf{A}\delta)$).

The proof of (6.35) remains unchanged.

6.8.6. Resampling constants. In this section, we prove the statements of Sect. 6.3.3, in particular the ones of Tab. 6.1. Using Lemma 5.7 of Sect. 5.7.2, this gives completely explicit formulas for the penalty in the histogram case:

$$\mathrm{pen}(m) = \frac{C}{n}\sum_{\lambda\in\Lambda_m}\left(R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)\right)\frac{n\widehat{p}_\lambda S_{\lambda,2}-S^2_{\lambda,1}}{n\widehat{p}_\lambda-1} \tag{6.39}$$

$$\text{with} \quad S_{\lambda,k} = \sum_{X_i\in I_\lambda}\left(Y_i-\beta_\lambda\right)^k \quad \text{for } k=1,2 \ . \tag{6.40}$$

PROPOSITION 6.5. *Let W be an exchangeable resampling weight vector among Efr(q_n), Rad(p), Poi(μ), Rho(q_n) and Loo and $C_{W,\infty}$ defined as in Tab. 6.1. Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$ and* $\mathrm{pen}(m)$ *defined as in* (6.5).

Then, there exist real numbers $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ (depending on the resampling scheme chosen) such that

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] = \frac{C}{C_{W,\infty}n}\sum_{\lambda\in\Lambda_m}\left(2+\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}\right)\left(\left(\sigma^d_\lambda\right)^2+\left(\sigma^r_\lambda\right)^2\right) \ . \tag{6.41}$$

If $\limsup_{n\to\infty} q_n n^{-1} < \infty$ *(Efr),* $p\in(0;1)$ *(Rad),* $\mu>0$ *(Poi) or* $0<\liminf q_n n^{-1} \leq \limsup q_n n^{-1} < 1$ *(Rho), then*

$$\lim_{n\geq n\widehat{p}_\lambda\to\infty}\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} = 0$$

and explicit non-asymptotic bounds are given by (6.42) *to* (6.47).

PROOF OF PROP. 6.5. From (6.39), we obtain (6.41) with

$$\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} = C_{W,\infty}\left(R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)\right)-2 \ .$$

Explicit formulas for $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ in each case come from Lemma 6.7 below. Combining it with Lemma 6.1 (for Efr and Rad), Lemma 8.14 in Sect. 8.7 (for Rad(1/2)), Lemma 6.2 (for Rho and Loo) and Lemma 6.3 (for Poi), we obtain the following non-asymptotic bounds:

(1) Efron (q_n): let $\kappa_3 = 5.1$ and $\kappa_4 = 3.2$,

$$(\kappa_4-1)\wedge\left(\frac{\kappa_3}{\left(q\widehat{p}_\lambda\right)^{1/4}}\right) \geq \delta^{(\mathrm{penEfr}(q_n))}_{n,\widehat{p}_\lambda} \geq \frac{-2}{n\widehat{p}_\lambda} - e^{-q\widehat{p}_\lambda} \ . \tag{6.42}$$

(2) Rademacher (p):

$$\frac{2}{1-p} \times (\kappa_4 - 1) \wedge \left(\frac{\kappa_3}{(np\widehat{p}_\lambda)^{1/4}} \right) \geq \delta_{n,\widehat{p}_\lambda}^{(\text{penRad}(p))} \geq \frac{-2e^{-pn\widehat{p}_\lambda}}{1-p} \tag{6.43}$$

$$\left(1 + 3 \times 10^{-4}\right) \wedge \left(\frac{\kappa_3 \times 2^{1/4}}{(n\widehat{p}_\lambda)^{1/4}} \right) \geq \delta_{n,\widehat{p}_\lambda}^{(\text{penRad}(1/2))} \geq -\mathbb{1}_{n\widehat{p}_\lambda \leq 2} \ . \tag{6.44}$$

(3) Poisson (μ):

$$1 \wedge \frac{2\left(1+e^{-3}\right)}{\left(\mu n\widehat{p}_\lambda - 2\right)_+} \geq \delta_{n,\widehat{p}_\lambda}^{(\text{penPoi}(\mu))} \geq \frac{-2}{n\widehat{p}_\lambda} - \left(e^{-\mu n\widehat{p}_\lambda} \wedge \mathbb{1}_{\mu n\widehat{p}_\lambda < 1.61} \right) \ . \tag{6.45}$$

(4) Random hold-out (q_n): let $\kappa_5(\epsilon) = 0.9 + 1.4 \times \epsilon^{-2}$.

$$\text{If} \qquad n\widehat{p}_\lambda \geq 2 \quad \text{and} \quad \frac{n}{q} \leq (1-\epsilon)\frac{2r}{2+\sqrt{3(r+1)\ln(r)}} \qquad \text{with } \epsilon \in (0;1)$$

$$\text{Then,} \qquad \kappa_5(\epsilon)\frac{n^2}{q_n(n-q_n)}\sqrt{\frac{\ln(n\widehat{p}_\lambda)}{n\widehat{p}_\lambda}} \geq \delta_{n,\widehat{p}_\lambda}^{(\text{penRho}(q_n))} \geq \frac{-ne^{-q\widehat{p}_\lambda}}{n-q} \ . \tag{6.46}$$

(5) Leave-one-out:

$$\frac{\mathbb{1}_{n\widehat{p}_\lambda \geq 2}}{n\widehat{p}_\lambda - 1} \geq \delta_{n,\widehat{p}_\lambda}^{(\text{penLoo})} \geq -\mathbb{1}_{n\widehat{p}_\lambda = 1} \ . \tag{6.47}$$

Notice that the lower bound in (6.46) does not require the assumption above. □

A byproduct of the proof of Prop. 6.5 (combined with Lemma 5.6 in Sect. 5.7.2), is the following:

LEMMA 6.6. *Assume that W is a weight vector among Efr, Rad, Poi, Rho and Loo. Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda \in \Lambda_m}$, $p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$ and $\text{pen}(m)$ be defined by (6.39) with $C = C_{W,\infty}$ (cf. Tab. 6.1). Then, if $\min_{\lambda \in \Lambda_m}\{n\widehat{p}_\lambda\} \geq 3$,*

$$\mathbb{E}^{\Lambda_m}\left[\text{pen}(m)\right] \geq \frac{5}{4}\mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \ . \tag{6.48}$$

REMARK 6.8. For another exchangeable resampling scheme, we would define $C_{W,\infty}$ such that it satisfies:

- (6.48) with a constant $\kappa > 1$ instead of $5/4$
- $\lim_{n\widehat{p}_\lambda \to 0} \delta_{n,\widehat{p}_\lambda}^{(\text{penW})} = 0$ as in Prop. 6.5.

In Tab. 6.1, $C_{W,\infty}$ always satisfies the second condition, but not necessarily the first one. One can then change condition (6.48) by replacing the threshold 3 by T, and use the same value T in the definition of $\widehat{\mathcal{M}}_n$ in algorithm 6.2.

In the proof of Prop. 6.5, we use the following lemma.

LEMMA 6.7. *Let $n \in \mathbb{N}$ and $\widehat{p}_\lambda \in [0,1]$ such that $n\widehat{p}_\lambda \in \{1,\ldots,n\}$. Then, for every $q \in \mathbb{N}\backslash\{0\}$ (and $q \leq n$ for the Rho case), $p \in (0;1]$ and $\mu > 0$,*

$$R_{1,\text{Efr}(q)} = \frac{n}{q} e^+_{\mathcal{B}(q,\widehat{p}_\lambda)}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right) \qquad R_{2,\text{Efr}(q)} = \frac{n}{q}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right) \tag{6.49}$$

$$R_{1,\text{Rad}(p)} = \frac{1}{p} e^+_{\mathcal{B}(n\widehat{p}_\lambda,p)} - 1 \qquad R_{2,\text{Rad}(p)} = \frac{1}{p} - 1 \tag{6.50}$$

$$R_{1,\text{Poi}(\mu)} = \frac{1}{\mu} e^+_{\mathcal{P}(n\widehat{p}_\lambda\mu)}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right) \qquad R_{2,\text{Poi}(\mu)} = \frac{1}{\mu}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right) \tag{6.51}$$

$$R_{1,\mathrm{Rho}(q)} = \frac{n}{q} e^+_{\mathcal{H}(n,n\widehat{p}_\lambda,q)} - 1 \qquad R_{2,\mathrm{Rho}(q)} = \frac{n}{q} - 1 \tag{6.52}$$

$$R_{1,\mathrm{Loo}} = \frac{n\widehat{p}_\lambda}{n(n\widehat{p}_\lambda - 1)} \mathbb{1}_{n\widehat{p}_\lambda \geq 2} \qquad R_{2,\mathrm{Loo}} = \frac{1}{n-1} \tag{6.53}$$

where $\mathcal{B}$, $\mathcal{P}$ and $\mathcal{H}$ are respectively the Binomial, Poisson and Hypergeometric distributions, and $e^+_\mu = \mathbb{E}[Z]\,\mathbb{E}\left[Z^{-1} \,\middle|\, Z > 0\right]$ with $Z \sim \mu$.

PROOF OF LEMMA 6.7. Since the randomness of W is independent from the data, we can assume that the observations with $X_i \in I_\lambda$ are the $n\widehat{p}_\lambda$ first ones: $(X_1, Y_1), \ldots, (X_{n\widehat{p}_\lambda}, Y_{n\widehat{p}_\lambda})$. The random vector $(W_i)_{1 \leq i \leq n\widehat{p}_\lambda}$ is then exchangeable (since W is). By definition of $W_\lambda = (n\widehat{p}_\lambda)^{-1} \sum_{i=1}^{n\widehat{p}_\lambda} W_i$, we deduce

$$\forall i \in \{1, \ldots, n\widehat{p}_\lambda\}, \quad \mathbb{E}_W[W_i \mid W_\lambda] = W_\lambda \ . \tag{6.54}$$

Then, the quantity

$$R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) = R_V(W_\lambda) = \mathbb{E}\left[(W_i - W_\lambda)^2 \,\middle|\, W_\lambda\right]$$

appearing both in $R_{1,W}$ and $R_{2,W}$ is the variance of the weight W_i conditionally to W_λ.

Exchangeable subsampling weights. We call *subsampling weight* any resampling weight W such that $W_i \in \{0, \kappa\}$ a.s. for every i. Such weights can be written $W_i = \kappa \mathbb{1}_{i \in I}$ for some random $I \subset \{1, \ldots, n\}$. Rad and Rho are the two main examples of such weights, and they are both exchangeable. In their example 3.6.14, van der Vaart and Wellner [**vdVW96**] call this kind of weights "bootstrap without replacement". Using (6.54), we derive that

$$W_\lambda = \mathbb{E}_W[W_i \mid W_\lambda] = \kappa \mathbb{P}(W_i = \kappa \mid W_\lambda)$$

and thus

$$\mathcal{L}(W_i \mid W_\lambda) = \kappa \mathcal{B}(\kappa^{-1} W_\lambda) \qquad \text{and} \qquad R_V(W_\lambda) = W_\lambda(\kappa - W_\lambda) \ .$$

We then apply this result to Rad, for which $\kappa = p^{-1}$ and $\mathcal{L}(W_\lambda) = (n\widehat{p}_\lambda p)^{-1} \times \mathcal{B}(n\widehat{p}_\lambda, p)$ and deduce (6.50). In the Rho case, we have $\kappa = (n/q)$ and $\mathcal{L}(W_\lambda) = (q\widehat{p}_\lambda)^{-1} \mathcal{H}(n, n\widehat{p}_\lambda, q)$, so that (6.52) follows. The Loo is a particular case of Rho (with $q = n-1$), so that we only have to compute $e^+_{\mathcal{H}(n,n\widehat{p}_\lambda,n-1)}$. This is done with (6.26) in Lemma 6.2.

Efron (q). Efron weights may also be written

$$W_i = \frac{n}{q} \mathrm{Card}\{1 \leq j \leq q \text{ s.t. } U_j = i\} \tag{6.55}$$

with $(U_j)_{1 \leq j \leq q}$ a sequence of i.i.d. random variables, uniform in $\{1, \ldots, n\}$. From this, we deduce

$$\mathcal{L}(W_\lambda) = (q\widehat{p}_\lambda)^{-1} \mathcal{B}(q, \widehat{p}_\lambda) \qquad \text{and} \qquad \mathcal{L}(W_i \mid W_\lambda) = \frac{n}{q} \mathcal{B}\left(q\widehat{p}_\lambda W_\lambda, \frac{1}{n\widehat{p}_\lambda}\right) \ .$$

Thus,

$$R_V(W_\lambda) = \frac{n}{q} W_\lambda \left(1 - \frac{1}{n\widehat{p}_\lambda}\right)$$

and (6.49) follows.

Poisson (μ). It is easy to check that the weights defined by (6.55), with $q = N_n \sim \mathcal{P}(\mu n)$ independent from the $(U_j)_{j \geq 1}$, are actually Poisson (μ) weights. This is the classical poissonization trick. Moreover, conditionally to W_λ and $N_n = q$, the same reasoning as for Efron (q) (with a multiplicative constant μ^{-1} instead of n/q) leads to (6.51). □

6.8.7. Concentration inequalities.

Resampling penalties. According to (5.19) and (5.26), the ideal penalty is a U-statistics of order 2, conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda\in\Lambda_m)}$. From the asymptotic viewpoint, this is sufficient to show that resampling gives a consistent estimate of it ([**AG92**] treated the bootstrap case; [**HJ93**] extended it to general resampling weights). In our non-asymptotic framework, we need more accurate results, that really use the explicit computations (5.29).

PROPOSITION 6.8. *Let W be an exchangeable weight vector and* $\mathrm{pen}(m)$ *the corresponding Resampling Penalty defined by* (6.39). *Let $\gamma>0$ and $A_n\geq 2$. Assume that*

$$\forall q\geq 2,\qquad P_m^\ell(q)\leq a_\ell q^{\xi_\ell}\ . \tag{$\mathbf{A_{m},\ell}$}$$

Then, on an event of probability at least $1-Ln^{-\gamma}$,

$$\begin{aligned}\left|\mathrm{pen}(m)-\mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right|\mathbb{1}_{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq A_n}&\leq CL(a_\ell,\xi_\ell,\gamma)\\ \times\sup_{np\geq A_n}\left\{R_{1,W}(n,p)+R_{2,W}(n,p)\right\}&\frac{\ln(n)^{\xi_\ell+1}}{\sqrt{A_nD_m}}\mathbb{E}\left[p_2(m)\right]\end{aligned} \tag{6.56}$$

where $R_{1,W}$ and $R_{2,W}$ are defined by (6.8) *and* (6.9).

If moreover W satisfies the assumptions of the second part of Prop. 6.5 and $C_{W,\infty}$ is defined as in Tab. 6.1, then

$$\begin{aligned}\left|\mathrm{pen}(m)-\mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right|\mathbb{1}_{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq A_n}&\leq\frac{C}{C_{W,\infty}}L(a_\ell,\xi_\ell,\gamma)\\ &\times L(W)\times\frac{\ln(n)^{\xi_\ell+1}}{\sqrt{A_nD_m}}\mathbb{E}\left[p_2(m)\right]\ .\end{aligned} \tag{6.57}$$

with $L(\mathrm{Rad}(p))=L\times(1-p)^{-1}$ and $L(W)=L$ for the other resampling schemes.

REMARK 6.9. With the $A_n^{-1/2}$ factor, we obtain better bounds for resampling penalties than for ideal penalties. In our particular framework, this is due to the better concentration properties of $S_{\lambda,2}$ compared to $S_{\lambda,1}^2-S_{\lambda,2}$.

This phenomenon is classical with bootstrap and may be understood in the asymptotic viewpoint through Edgeworth expansions (Hall [**Hal92**]). In a non-asymptotic gaussian framework, [**ABR07**] (see Sect. 10.2.3) show the same property for resampling estimators, which concentrates at the rate n^{-1} instead of $n^{-1/2}$ (n being the amount of data). As A_n plays the role of n in our case, it is reasonable to believe that the gain $A_n^{-1/2}$ may not be improved without some more assumptions.

This stresses the fact that resampling penalties do not estimate the ideal penalties themselves but their expectations. Thus, our procedure cannot take into account the fact that $\mathrm{pen}_{\mathrm{id}}(m)$ may be far from its expectation.

PROOF OF PROP. 6.8. According to (6.39), $\mathrm{pen}(m)$ is a U-statistics of order 2 conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda)}$. Then, we use either Prop. 5.5 (in Sect. 5.6.3), with

$$a_\lambda=\frac{R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)}{n(n\widehat{p}_\lambda-1)}\qquad b_\lambda=\frac{-\left(R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)\right)}{n^2\widehat{p}_\lambda(n\widehat{p}_\lambda-1)}$$

or results from [**GLZ00**]. This proves, for all $q\geq 2$,

$$\begin{aligned}\left\|\mathrm{pen}(m)-\mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right\|_q^{(\Lambda_m)}&\leq L(a_\ell,\xi_\ell)D_m^{-1/2}A_n^{-1/2}\\ \times\sup_{np\geq A_n}\left\{R_{1,W}(n,p)+R_{2,W}(n,p)\right\}&q^{\xi_\ell+1}\mathbb{E}\left[p_2(m)\right]\ .\end{aligned}$$

We deduce conditional concentration inequalities with Lemma 8.10 (Sect. 8.6.2), taking $x=\gamma\ln(n)$. Since x is deterministic, this implies unconditional concentration inequalities.

The second statement follows from the proof of Prop. 6.5, where we can find non-asymptotic upper bounds on

$$2 + \delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} = C_{W,\infty} \times \left(R_{1,W}(n,\widehat{p}_\lambda) + R_{2,W}(n,\widehat{p}_\lambda) \right) \ .$$

□

Ideal penalty. We split the ideal penalty into three terms: $p_1(m)$, $p_2(m)$ and $\delta(m)$. Concentrations inequalities for the two first ones are proven in Sect. 5.7.4, we recall them

PROPOSITION 6.9 (Prop. 5.8, Sect. 5.7.4). *Let $\gamma > 0$. Assume that $\min_{\lambda\in\Lambda_m} \{np_\lambda\} \geq B_n \geq 1$ and*

$$\forall q \geq 2, \qquad P^\ell_m(q) \leq a_\ell q^{\xi_\ell} \ . \tag{$\mathbf{A_{m},\ell}$}$$

Then, on an event of probability at least $1 - Ln^{-\gamma}$,

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L(a_\ell,\xi_\ell,\gamma)\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{6.58}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L(a_\ell,\xi_\ell,\gamma)\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{6.59}$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L(a_\ell,\xi_\ell,\gamma) D_m^{-1/2}\ln(n)^{\xi_\ell+1}\mathbb{E}\left[p_2(m)\right] \ . \tag{6.60}$$

If we only have a lower bound $B_n > 0$, then, with probability at least $1 - Ln^{-\gamma}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n^{-1}\ln(n)} - L(a_\ell,\xi_\ell,\gamma)\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right]\right)\mathbb{E}\left[\widetilde{p_2}(m)\right] \ . \tag{6.61}$$

We now come to $\delta(m) = (P - P_n)\gamma(s) + \overline{\delta}(m)$. When the data is bounded, we use Prop. 3.3 in Sect. 3.3:

LEMMA 6.10 (Prop. 3.3, Sect. 3.3). *Assume that $\|Y\|_\infty \leq A < \infty$. Then for all $x \geq 0$, on an event of probability at least $1 - 2e^{-x}$:*

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s,s_m)}{\sqrt{D_m}} + \frac{20}{3}\frac{A^2}{Q^{(p)}_m}\frac{\mathbb{E}^{\Lambda_m}[p_2(m)]}{\sqrt{D_m}}x \tag{6.62}$$

$$\text{and} \quad \forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta l(s,s_m) + \left(\frac{4}{\eta} + \frac{8}{3}\right)\frac{A^2 x}{n} \ . \tag{6.63}$$

In the unbounded case, we need another concentration inequality for $\overline{\delta}(m)$. There are many strategies for this, since it is a sum of i.i.d. centered random variables. In our framework, the following result is sufficient.

LEMMA 6.11. *Assume that*

$$\forall q \geq 2, \qquad P^{g\epsilon}(q) \leq a_{g\epsilon} q^{\xi_{g\epsilon}} \tag{$\mathbf{A_{g},\epsilon}$}$$

$$\|\sigma(X)\|_\infty \leq \sigma_{\max} \tag{$\mathbf{A}\sigma_{\max}$}$$

$$\|s - s_m\|_\infty \leq c^g_{\Delta,m}\|s(X) - s_m(X)\|_2 \ . \tag{$\mathbf{A}\delta$}$$

Then, for every $x \geq 0$, there exists an event of probability at least $1 - e^{-x}$ on which

$$\left|\overline{\delta}(m)\right| \leq \frac{T_{\delta,m}(x)}{\sqrt{D_m}}\left[l(s,s_m) + \frac{\sigma^2_{\max}}{Q^{(p)}_m}\mathbb{E}[p_2(m)]\right] \tag{6.64}$$

with $T_{\delta,m}(x) \leq L\left(a_{g\epsilon}, \xi_{g\epsilon}, c^g_{\Delta,m}\right)x^{\xi_{g\epsilon}+1/2}$.

On the other hand, if

$$\forall q \geq 2, \qquad P^{g\epsilon}(q) \leq a_{g\epsilon} q^{\xi_{g\epsilon}} \tag{$\mathbf{A}_{\mathbf{g},\epsilon}$}$$

$$\|\sigma(X)\|_\infty \leq \sigma_{\max} \tag{$\mathbf{A}\sigma_{\max}$}$$

$$\|s\|_\infty \leq A \ , \tag{$\mathbf{A}\mathbf{s}_{\max}$}$$

then, for every $x \geq 0$, there exists an event of probability at least $1 - e^{-x}$ on which

$$\left|\overline{\delta}(m)\right| \leq L\left(a_{g\epsilon}, \xi_{g\epsilon}, A, \sigma_{\max}\right) x^{\xi_{g\epsilon}+1/2} \ . \tag{6.65}$$

PROOF OF LEMMA 6.11. From Lemma 8.18, we have

$$\begin{aligned}
\left\|\overline{\delta}(m)\right\|_q &\leq \frac{2\sqrt{\kappa}\sqrt{q}}{\sqrt{n}} \left\|F_m - \mathbb{E}[F_m]\right\|_q \\
\text{with} \quad F_m &= (Y - s_m(X))^2 - (Y - s(X))^2 \\
&= (s_m(X) - s(X))^2 - 2\epsilon\sigma(X)(s_m(X) - s(X)) \ .
\end{aligned}$$

Notice that $\epsilon\sigma(X)(s_m(X) - s(X))$ is centered conditionally to $X \in I_\lambda$ for all $\lambda \in \Lambda_m$. We thus have

$$\left\|\overline{\delta}(m)\right\|_q \leq \frac{2\sqrt{\kappa}\sqrt{q}}{\sqrt{n}} \left(\|s - s_m\|_\infty^2 + 2\sigma_{\max} \|s - s_m\|_\infty \|\epsilon\|_q \right) \ . \tag{6.66}$$

We now use assumptions ($\mathbf{A}_{\mathbf{g},\epsilon}$) and ($\mathbf{A}\delta$). Then, for all $q \geq 2$,

$$\begin{aligned}
\left\|\overline{\delta}(m)\right\|_q &\leq 2\sqrt{\kappa}\sqrt{q} \left((c^g_{\Delta,m})^2 l(s, s_m) + 2c^g_{\Delta,m} \sqrt{l(s, s_m)} P^{g\epsilon}_m(q) \sigma_{\max} \right) \frac{1}{\sqrt{n}} \\
&= S^{(1)}_{\delta,m}(q) \frac{l(s, s_m)}{\sqrt{n}} + S^{(2)}_{\delta,m}(q) \sqrt{\frac{l(s, s_m)\sigma^2_{\max}}{n}} \\
&\leq \left(\frac{1}{\sqrt{n}} S^{(1)}_{\delta,m}(q) + \theta S^{(2)}_{\delta,m}(q) \right) l(s, s_m) + \frac{S^{(2)}_{\delta,m}(q)}{4\theta} \frac{\sigma^2_{\max}}{n} \quad \text{for all} \quad \theta > 0
\end{aligned}$$

with

$$S^{(1)}_{\delta,m}(q) = 2\sqrt{\kappa}(c^g_{\Delta,m})^2\sqrt{q} \qquad S^{(2)}_{\delta,m}(q) = 4\sqrt{\kappa} c^g_{\Delta,m} \sqrt{q} P^{g\epsilon}_m(q) \ .$$

We take $\theta = D_m^{-1/2}$ and deduce the concentration inequality (6.64) with the classical link between moments and concentration (see for instance Lemma 8.10 in Sect. 6.1).

For the second statement, start back from (6.66) and remark that $\|s - s_m\|_\infty \leq 2A$. □

6.8.8. Technical lemmas.

Empirical and expected frequencies.

LEMMA 6.12. *Let $(p_\lambda)_{\lambda\in\Lambda_m}$ be non-negative real numbers of sum 1, $(n\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$ a multinomial vector of parameters $(n; (p_\lambda)_{\lambda\in\Lambda_m})$. Then, for all $\gamma > 0$,*

$$\min_{\lambda\in\Lambda_m} \{ n\widehat{p}_\lambda \} \geq \frac{\min_{\lambda\in\Lambda_m} \{ np_\lambda \}}{2} - 2(\gamma + 1)\ln(n) \tag{6.67}$$

with probability at least $1 - 2n^{-\gamma}$.

PROOF OF LEMMA 6.12. By Bernstein inequality ([**Mas07**], Prop. 2.9), for all $\lambda \in \Lambda_m$,

$$\mathbb{P}\left(n\widehat{p}_\lambda \geq (1-\theta)np_\lambda - \sqrt{2npx} - \frac{x}{3} \right) \geq 1 - e^{-x} \ .$$

Take $x = (\gamma + 1)\ln(n)$ above, and remark that $\sqrt{2npx} \leq \frac{np}{2} + x$. The union bound gives the result since $\mathrm{Card}(\Lambda_m) \leq n$. □

Bounds for $Q_m^{(p)}$.

LEMMA 6.13. *Recall that*

$$Q_m^{(p)} := \frac{1}{D_m} \sum_{\lambda \in \Lambda_m} \left[(\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d \right)^2 \right] \ .$$

(1) *If $\sigma(X) \geq \sigma_{\min} > 0$ ($\mathbf{An}$), then*

$$Q_m^{(p)} \geq \sigma_{\min}^2 > 0 \ .$$

(2) *If $\mathcal{X} \subset \mathbb{R}^k$, $\max_{\lambda \in \Lambda_m} \{ \operatorname{diam}(I_\lambda) \} \leq c_{\mathrm{r,u}}^d D_m^{-\alpha_d} \operatorname{diam}(X)$ ($\mathbf{Ar_u^d}$), $\max_{\lambda \in \Lambda_m} \{ \operatorname{Leb}(I_\lambda) \} \leq c_{\mathrm{r,u}} D_m^{-1} \operatorname{Leb}(\mathcal{X})$ ($\mathbf{Ar_u}$), and σ is piecewise K_σ-Lipschitz with at most J_σ jumps ($\mathbf{A}\sigma$), then*

$$Q_m^{(p)} \geq \frac{\|\sigma\|_{L^2(\mathrm{Leb})}^2}{2c_{\mathrm{r,u}}} - \frac{K_\sigma^2 \left(c_{\mathrm{r,u}}^d \right)^2 \operatorname{diam}(\mathcal{X})^2}{D_m^{2\alpha_d}} - \frac{J_\sigma \|\sigma(X)\|_\infty^2}{2D_m} \ .$$

(3) *We also have the upper bound:*

$$Q_m^{(p)} \leq \|\sigma(X)\|_\infty^2 + \|s\|_\infty^2 \ .$$

REMARK 6.10. In 3., $\|\sigma(X)\|_2 > 0$ and σ is piecewise Lipschitz so that $\|\sigma\|_{L^2(\mathrm{Leb})} > 0$. Thus, the lower bound for $Q_m^{(p)}$ is positive when D_m is large enough.

PROOF OF LEMMA 6.13. The first and last results are straightforward. For the second one, remark that for every $\lambda \in \Lambda_m$ such that σ does not jump on I_λ,

$$(\sigma_\lambda^r)^2 \geq \min_{x \in I_\lambda} \left\{ \sigma(X)^2 \right\} \geq \frac{1}{2 \operatorname{Leb}(I_\lambda)} \int_{I_\lambda} \sigma(x)^2 \operatorname{Leb}(dx) - \left(K_\sigma \operatorname{diam}(I_\lambda) \right)^2$$

since $(a-b)^2 = a^2 - 2ab + b^2 \geq \frac{a^2}{2} - b^2$ and σ is K_σ Lipschitz. There is at most J_σ other λ, for which

$$(\sigma_\lambda^r)^2 \geq 0 \geq \frac{1}{2 \operatorname{Leb}(I_\lambda)} \int_{I_\lambda} \sigma(x)^2 \operatorname{Leb}(dx) - \frac{\|\sigma(X)\|_\infty^2}{2} \ .$$

This implies

$$\sum_{\lambda \in \Lambda_m} (\sigma_\lambda^r)^2 \geq \frac{\operatorname{Leb}(\mathcal{X})}{2 \max_{\lambda \in \Lambda_m} \{ \operatorname{Leb}(I_\lambda) \}} \|\sigma\|_{L^2(\mathrm{Leb})}^2 - D_m K_\sigma^2 \max_{\lambda \in \Lambda_m} \left\{ \operatorname{diam}(I_\lambda)^2 \right\} - \frac{J_\sigma \|\sigma(X)\|_\infty^2}{2} \ .$$

□

Sufficient condition for ($\mathbf{A}\delta$).

LEMMA 6.14. *Assume that $\mathcal{X} \subset \mathbb{R}$ is bounded and:*

($\mathbf{Al}$) *$s : \mathcal{X} \mapsto \mathbb{R}$ is B-Lipschitz, piecewise C^1 and non-constant (*i.e. *$\pm s' \geq B_0 > 0$ on some interval $J \subset \mathcal{X}$ with $\operatorname{Leb}(J) \geq c_J \operatorname{Leb}(\mathcal{X})$, with $c_J > 0$).*

($\mathbf{Ar}_{\ell,\mathbf{u}}$) *Regularity of the partitions for* Leb*:*

$$\forall \lambda \in \Lambda_m, \quad c_{\mathrm{r},\ell} D_m^{-1} \operatorname{Leb}(\mathcal{X}) \leq \operatorname{Leb}(I_\lambda) \leq c_{\mathrm{r,u}} D_m^{-1} \operatorname{Leb}(\mathcal{X}) \ .$$

($\mathbf{Ad}_\ell$) *Density bounded from below: $\exists c_X^{\min} > 0$, $\forall I \subset \mathcal{X}$, $P(X \in I) \geq c_X^{\min} \operatorname{Leb}(I) \operatorname{Leb}(\mathcal{X})^{-1}$.*

Then,

$$\|s - s_m\|_\infty \leq c_{\Delta,m}^g \|s(X) - s_m(X)\|_2 \qquad (\mathbf{A}\delta)$$

holds with

$$c_{\Delta,m}^g = \left(\frac{c_{\mathrm{r,u}}}{c_{\mathrm{r},\ell}} \right)^{3/2} \frac{B\sqrt{24}}{B_0 \sqrt{c_X^{\min} c_J}} \qquad \text{if} \quad D_m \geq 4 c_{\mathrm{r,u}} c_J^{-1} \ .$$

PROOF OF LEMMA 6.14. From (**Al**) and the upper bound in ($\mathbf{Ar}_{\ell,\mathbf{u}}$),

$$\|s - s_m\|_\infty \leq B \max_{\lambda\in\Lambda_m} \{\operatorname{diam}(I_\lambda)\} = B \max_{\lambda\in\Lambda_m} \{\operatorname{Leb}(I_\lambda)\} \leq B c_{\mathrm{r,u}} \operatorname{Leb}(\mathcal{X}) D_m^{-1} \ . \tag{6.68}$$

For the lower bound, let Λ_m^J be the set of $\lambda \in \Lambda_m$ such that $I_\lambda \subset J$,

$$s_{\lambda,\mathrm{Leb}} = \operatorname{Leb}(I_\lambda)^{-1} \int_{I_\lambda} s(x) \operatorname{Leb}(dx)$$

and $\mu = \mathcal{L}(X)$. Then, using ($\mathbf{Ad}_\ell$),

$$\begin{aligned}\|s - s_m\|_{L^2(\mu)}^2 &\geq \sum_{\lambda\in\Lambda_m^J} \int_{I_\lambda} (s(x) - s_m(x))^2 \, c_X^{\min} \operatorname{Leb}(\mathcal{X})^{-1} \operatorname{Leb}(dx) \\ &\geq c_X^{\min} \operatorname{Leb}(\mathcal{X})^{-1} \sum_{\lambda\in\Lambda_m^J} \int_{I_\lambda} (s(x) - s_{\lambda,\mathrm{Leb}})^2 \operatorname{Leb}(dx) \ .\end{aligned}$$

For any $\lambda \in \Lambda_m^J$, since s is continuous on I_λ, there is some $x_\lambda \in I_\lambda$ such that $s_{\lambda,\mathrm{Leb}} = s(x_\lambda)$. By (**Al**), for every $x \in I_\lambda$,

$$(s(x) - s(x_\lambda))^2 \geq B_0^2 (x - x_\lambda)^2$$

so that

$$\begin{aligned}\|s - s_m\|_{L^2(\mu)}^2 &\geq c_X^{\min} \operatorname{Leb}(\mathcal{X})^{-1} \sum_{\lambda\in\Lambda_m^J} \frac{B_0^2 \operatorname{Leb}(I_\lambda)^3}{12} \geq \frac{c_X^{\min} B_0^2 c_{\mathrm{r},\ell}^3 \operatorname{Leb}(\mathcal{X})^2}{12 D_m^3} \operatorname{Card}\left(\Lambda_m^J\right) \\ &\geq \frac{c_X^{\min} B_0^2 c_{\mathrm{r},\ell}^3 \operatorname{Leb}(\mathcal{X})^2}{12 D_m^3} \left(\frac{c_J D_m}{c_{\mathrm{r,u}}} - 2\right)_+ \ .\end{aligned}$$

Combined with (6.68), this gives the result. □

6.8.9. Expectations of inverses. There is no tight general upper bound, but the following decomposition may be useful if Z satisfies some appropriate concentration inequality around its expectation and if there exists $c_Z > 0$ such that $\mathbb{P}(c_Z > Z > 0) = 0$:

$$\begin{aligned}\forall \alpha > 0, \quad e_Z^0 &= \mathbb{E}\left[Z^{-1} \mathbb{1}_{Z>0}\right] \mathbb{E}[Z] \\ &= \mathbb{E}\left[Z^{-1} \mathbb{1}_{\alpha\mathbb{E}[Z]>Z>0}\right] \mathbb{E}[Z] + \mathbb{E}\left[Z^{-1} \mathbb{1}_{Z\geq\alpha\mathbb{E}[Z]}\right] \mathbb{E}[Z] \\ &\leq \mathbb{P}\left(\alpha\mathbb{E}[Z] > Z > 0\right) \mathbb{E}[Z] c_Z^{-1} + \alpha^{-1} \ .\end{aligned} \tag{6.69}$$

Hypergeometric case.

PROOF OF LEMMA 6.2. Let $Z \sim \mathcal{H}(n, r, q)$. It has an expectation $\mathbb{E}[Z] = \frac{qr}{n}$.

General lower bound. We first use (6.20) and

$$\mathbb{P}(Z = 0) \leq \left(1 - \frac{r}{n}\right)^q \leq \exp\left(-\frac{qr}{n}\right) \ .$$

Moreover, if $r \geq n - q + 1$, $\mathbb{P}(Z > 0) = 1$.

A general upper bound. According to (6.19) and the lower bound for $\mathbb{P}(Z > 0)$ above, it is sufficient to upper bound $e^0_{\mathcal{H}(n,r,q)}$. We first prove the following general result, that holds for every $n \geq r, q \geq 1$:

$$e^+_{\mathcal{H}(n,r,q)} \leq \frac{\inf_{\frac{q}{n}>\beta\geq\frac{2}{r}} \left\{\frac{qr}{n} \exp\left[-\frac{2(\beta r-1)^2}{r+1}\right] + \frac{1}{1-\frac{n\beta}{q}}\right\}}{1 - \exp\left(-\frac{qr}{n}\right)} \tag{6.70}$$

The idea of the proof is to use (6.69) with $c_Z = 1$, $\mathbb{E}[Z] = qrn^{-1}$. For this, we need the following concentration result by Hush and Scovel [**HS05**]: for all $x \geq 2$,

$$\begin{aligned}&\mathbb{P}\left(\mathbb{E}(Z) - Z > x\right)\\ &< \exp\left(-2(x-1)^2\left[\left(\frac{1}{r+1}+\frac{1}{n-r+1}\right)\vee\left(\frac{1}{q+1}+\frac{1}{n-q+1}\right)\right]\right)\end{aligned}$$

Taking $\alpha = 1 - \frac{n\beta}{q}$ with $\frac{q}{n} > \beta \geq \frac{2}{r}$, we obtain

$$\begin{aligned}e^0_{\mathcal{H}(n,r,q)} &\leq \frac{qr}{n}\exp\left[-\frac{2\left(\beta r-1\right)^2(n+2)}{(r+1)(n-r+1)}\right] + \frac{1}{1-\frac{n\beta}{q}}\\ &\leq \frac{qr}{n}\exp\left[-\frac{2(\beta r-1)^2}{r+1}\right] + \frac{1}{1-\frac{n\beta}{q}}\ .\end{aligned}$$

As a consequence, (6.70) holds.

Back to (6.24). With the supplementary conditions on n, r and q, we can take $\beta = \frac{1+\sqrt{\frac{3}{4}\ln(r)(r+1)}}{r}$ in (6.70). Hence

$$e^0_{\mathcal{H}(n,r,q)} \leq \frac{1}{2\sqrt{r}} + \frac{1}{1-\frac{n}{q}\left(\frac{1+\sqrt{\frac{3}{4}\ln(r)(r+1)}}{r}\right)} \leq 1 + \frac{n}{q}K(\epsilon)\sqrt{\frac{\ln(r)}{r}}$$

$$\text{with}\quad K(\epsilon) = \frac{1}{2\sqrt{\ln(2)}} + \frac{1}{\epsilon^2}\left(\sqrt{\frac{\ln(3)}{3}}+\frac{3}{4}\right)\ .$$

We then deduce (6.24) with

$$\kappa_5(\epsilon) = 0.9 + 1.4\times\epsilon^{-2} \geq 1.02\times K(\epsilon) + 0.03$$

since $r \geq 2$ and

$$\begin{aligned}e^+_{\mathcal{H}(n,r,q)} &\leq \left(1 + C_n K(\epsilon)\sqrt{\frac{\ln(r)}{r}}\right)\left(1-\exp\left(-\frac{rq}{n}\right)\right)^{-1}\\ &\leq \left(1 + \frac{n}{q}K(\epsilon)\sqrt{\frac{\ln(r)}{r}}\right)\left(1-\exp\left[-(2+\sqrt{3}\ln(r)\sqrt{r+1})\right]\right)^{-1}\\ &\leq \left(1 + \frac{n}{q}K(\epsilon)\sqrt{\frac{\ln(r)}{r}}\right)\left(1+\frac{e^{-2}r^{-3}}{(1-e^{-2-3\ln(2)})^2}\right)\ .\end{aligned}$$

"Rho" case. We now assume that $q = \lfloor \frac{n}{2} \rfloor$ so that $\frac{n}{q} = 2 + \frac{1}{\lfloor \frac{n}{2} \rfloor} \leq 3$ and converges to 2 when n goes to infinity.

For $r \geq 6$, we can take $\beta = \frac{2}{r}$ in (6.70) and we obtain:

$$e^+_{\mathcal{H}(n,6,q)} \leq 9.68 \qquad e^+_{\mathcal{H}(n,7,q)} \leq 7.61 \qquad e^+_{\mathcal{H}(n,8,q)} \leq 7.46 \qquad e^+_{\mathcal{H}(n,9,q)} \leq 7.32$$

For $r \geq 10$, taking $\beta = \frac{1}{4} + \frac{1}{r}$ in (6.70), we derive

$$\sup_{r\geq 10} e^+_{\mathcal{H}(n,r,q)} \leq 7.49 \qquad \sup_{r\geq 26} e^+_{\mathcal{H}(n,r,q)} \leq 3\ .$$

Small values of r. must be treated appart. For $r = 1$, it is easy to compute $e^+_{\mathcal{H}(n,1,q)} = qn^{-1} \leq 1$. When $n = r$, we have

$$e^+_{\mathcal{H}(n,n,q)} = 1 = e^0_{\mathcal{H}(n,r,n)}.$$

Otherwise, using the fact that for all $n \geq r+1$, $\frac{n!}{(n-r)!} \geq \frac{(r+1)!}{(r+1)^r} n^r$,

$$e^0_{\mathcal{H}(n,r,q)} \leq \frac{r}{R} \frac{(r+1)^r}{(r+1)! R^r} \left(\sum_{k=1}^{r} \binom{r}{k} \frac{(R-1)^{r-k}}{k} \right)$$

with $R = \frac{n}{q} \in [1; +\infty)$.

For $r = 2$, this upper bound is lower than 1.6. If $\frac{n}{q} \leq 3$ (this holds in the "Rho" case),

$$e^+_{\mathcal{H}(n,3,q)} \leq 4.67 \qquad e^+_{\mathcal{H}(n,4,q)} \leq 8.15 \qquad e^+_{\mathcal{H}(n,5,q)} \leq 14.29$$

"Loo" case. We now have $q = n-1$. We first consider $r = 1$. The conditionning make Z deterministic and equal to 1, so that

$$e^+_{\mathcal{H}(n,1,n-1)} = \mathbb{E}[Z] = \frac{n-1}{n} = 1 - \frac{1}{n} \ .$$

Now, if $r \geq 2$, $Z > 0$ holds a.s. since it only take two values:

$$Z = r-1 \text{ with probability } \frac{r}{n} \quad \text{and} \quad Z = r \text{ with probability } \frac{n-r}{n} \ .$$

As a consequence,

$$e^+_{\mathcal{H}(n,r,n-1)} = \frac{(n-1)r}{n} \left(\frac{r}{(r-1)n} + \frac{n-r}{nr} \right) = 1 + \frac{1}{n} \left(\frac{(n-1)r}{n(r-1)} - 1 \right) \ .$$

The lower bound is straightforward since $n \geq r$.

"Lpo" case. As noticed in Lemma 6.7, we have

$$\forall r \geq p+1, \qquad e^+_{\mathcal{H}(n,r,n-p)} \geq 1 \ .$$

Moreover, when $r \geq p+1$, $\mathcal{H}(n,r,n-p)$ has its support in $\{r-p, \dots, r\}$ and thus

$$\begin{aligned} e^+_{\mathcal{H}(n,r,n-p)} &= \frac{(n-p)r}{n} \sum_{j=r-p}^{r} \frac{\binom{r}{j}\binom{n-r}{n-p-j}}{j\binom{n}{n-p}} \\ &= \frac{(n-p)r}{n} \sum_{k=(p+r-n)\vee 0}^{p} \frac{\binom{r}{k}\binom{n-r}{p-k}}{(r-k)\binom{n}{p}} \ . \end{aligned}$$

More precisely, the k-th term of the sum is equal to

$$\begin{aligned} &\frac{(n-p)r}{n} \frac{\binom{r}{k}\binom{n-r}{p-k}}{(r-k)\binom{n}{p}} \\ &= \left(\frac{r}{n}\right)^k \left(1 - \frac{r}{n}\right)^{p-k} \binom{p}{k} \frac{n-p}{n} \frac{r}{r-k} \frac{r(r-1)\cdots(r-k+1)}{r^k} \\ &\quad \times \frac{(n-r)\cdots(n-r-(p-k)+1)}{(n-r)^{p-k}} \frac{n^p}{n\cdots(n-p+1)} \\ &\leq \left(\frac{r}{n}\right)^k \left(1 - \frac{r}{n}\right)^{p-k} \binom{p}{k} \frac{r}{r-p} \frac{n^p}{n\cdots(n-p+1)} \ . \end{aligned}$$

As a consequence, we have

$$\begin{aligned} e^+_{\mathcal{H}(n,r,n-p)} &\leq \sum_{k=(p+r-n)\vee 0}^{p} \left(\frac{r}{n}\right)^k \left(1 - \frac{r}{n}\right)^{p-k} \binom{p}{k} \frac{r}{r-p} \frac{n^p}{n\cdots(n-p+1)} \\ &\leq \frac{rn^p}{(r-p)n\cdots(n-p+1)} \ . \end{aligned}$$

The result follows. □

REMARK 6.11 (Asymptotics). If for some $\alpha > 0$, $q_k r_k^{1/2-\alpha} n_k^{-1} \xrightarrow[k\to+\infty]{} +\infty$, then the asymptotic result (6.28) holds. The upper bound is obtained by taking

$$\beta = \frac{1 + \sqrt{(r+1)\ln\left(\frac{qr}{n}\right)}}{r}$$

in (6.70) (it's possible for r sufficiently large). The lower bound is straightforward.

A particular case is when $\sup_k n_k q_k^{-1} < +\infty$ and $n_k \geq r_k \to +\infty$.

Poisson case.

PROOF OF LEMMA 6.3. Let $Z \sim \mathcal{P}(\mu)$, and define $g : [0;\infty) \mapsto \mathbb{R}$ by $g(0) = 0$ and for every $\mu > 0$

$$g(\mu) := e^+_{\mathcal{P}(\mu)} = \mu \mathbb{E}\left[Z^{-1} \mid Z > 0\right] = \frac{\mu e^{-\mu}}{1-e^{-\mu}} \sum_{k=1}^{+\infty} \frac{\mu^k}{k \times k!} = \frac{\mu}{e^\mu - 1} \int_0^\mu \frac{e^x - 1}{x} dx \ .$$

The function g is continuous at 0 and has a first derivative $g'(0) = 1$. For every $x \geq 0$, we define

$$h(x) = \frac{e^x - 1}{x} \qquad H(x) = \int_0^x h(t)dt \qquad a(x) = \frac{h'(x)}{h(x)} = 1 - \frac{e^x - 1 - x}{x(e^x - 1)} \ .$$

where the last equality holds if $x > 0$, and $a(0) = 1/2$. Then, $g(u) = H(u)/h(u)$ satisfies the following ordinary differential equation:

$$g(0) = 0 \qquad \forall u \geq 0, \quad g'(u) = 1 - a(u)g(u) \ .$$

Since

$$\forall u \geq 0, \quad \frac{1}{2} \leq a(u) \leq 1 \qquad \text{and} \qquad \lim_{u\to+\infty} a(u) = 1 \ ,$$

g satisfies a differential inequation

$$1 - \frac{g}{2} \leq g' \leq 1 - g \qquad g(0) = 0 \ .$$

Then, for every $x \geq x_0 \geq 0$,

$$2\left[1 - e^{2(x_0-x)}\left(1 - \frac{g(x_0)}{2}\right)\right] \geq g(x) \geq 1 + (g(x_0) - 1)e^{x_0 - x} \ . \tag{6.71}$$

Lower bound. The general lower bound (6.20) gives

$$g(\mu) \geq \mathbb{P}(Z > 0) = 1 - e^{-\mu} \ .$$

We can do better: remark that if $g(x_0) \geq 1$, (6.71) shows that $g(x) \geq 1$ for every $x \geq x_0$. Since $g = H/h$ and for every $u \geq 0$,

$$H(u) \geq u + \frac{u^2}{4} + \frac{u^3}{18} \ ,$$

we deduce that

$$g(u) \geq \frac{u\left(u + \frac{u^2}{4} + \frac{u^3}{18}\right)}{e^u - 1} \ .$$

Then, $g(1.61) \geq 1$, so that $g(x) \geq 1$ for every $x \geq 1.61$.

Upper bound. Using (6.71) with $x_0 = 0$ gives

$$\forall x \geq 0, \quad g(x) \leq 2 - 2e^{-2x} \leq 2.$$

Moreover, for every $\epsilon \in (0;1)$, $1 - \epsilon \leq a(x) \leq 1$ as soon as $x \geq \epsilon^{-1}$. Then, on $[\epsilon^{-1};\infty)$, g satisfies the differential inequation

$$g' \geq 1 - (1-\epsilon)g \ .$$

Integrating this between ϵ^{-1} and $2\epsilon^{-1}$, we obtain that

$$g(2\epsilon^{-1}) \leq \frac{1}{1-\epsilon}\left[1 + \left(g(\epsilon^{-1})(1-\epsilon) - 1\right)\exp\left(-\epsilon^{-1}(1-\epsilon)^{-1}\right)\right] \ .$$

For every $x > 2$, $\epsilon = 2x^{-1} \in (0;1)$ so that

$$g(x) \leq 1 + \frac{2 + (x-4)\exp\left(-\frac{x^2}{2(x-2)}\right)}{x-2} \leq 1 + \frac{2(1+e^{-3})}{x-2} \ .$$

The result follows. □

CHAPTER 7

The classification case

RÉSUMÉ. Ce chapitre est consacré à l'étude des pénalités par rééchantillonnage définies aux Chap. 5 et 6 dans un cadre général, incluant la classification binaire et la régression bornée. Ces pénalités peuvent être vues comme une version localisée des pénalités bootstrap globales de Fromont [**Fro04**], en particulier des complexités de Rademacher. Elles sont plus simples à calculer et beaucoup plus faciles à calibrer que les complexités de Rademacher locales. Nous prouvons des résultats intermédiaires en vue d'obtenir une inégalité-oracle non-asymptotique pour ces pénalités. Nous décrivons ensuite le chemin restant à parcourir pour comprendre théoriquement cette procédure, et discutons la manière de l'appliquer en pratique et de la calibrer.

7.1. Introduction

In the two previous chapters, we focused on the regression framework. However, in many practical applications, the outcome Y takes only a finite number of values (often two; then, we take the convention $Y \in \mathcal{Y} = \{0,1\}$), and we would like to predict it on new data with as few errors as possible. For instance, in genomics, given a DNA sequence $X \in \{A,T,C,G\}^{\mathbb{N}}$, we would like to know whether it can be transcripted or not. In pattern recognition, we are given a $N \times N$ bitmap image $X \in [0,1]^{N^2}$ and we would like to know which character (*e.g.* a figure or a letter) it represents. This is the *classification* framework (called "binary classification" when $\mathcal{Y} = \{0,1\}$). Then, we do not aim at estimating the regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$, but only the Bayes predictor $s(x) = \mathbb{1}_{\eta(x) \geq 1/2}$. More precisely, we want to build a predictor $t : \mathcal{X} \mapsto \mathcal{Y} = \{0,1\}$ which minimizes the prediction loss (also called "0-1 loss")

$$P\gamma(t) := \mathbb{P}(t(X) \neq Y) \qquad \text{where} \quad \gamma(t,(x,y)) = \mathbb{1}_{t(x) \neq y} \ .$$

As in the regression setting, given a set of predictors S_m (a *model*), this can be done by empirical risk minimization:

$$\widehat{s}_m \in \arg\min_{t \in S_m} \{P_n\gamma(t)\} \qquad \text{with} \quad P_n\gamma(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{t(X_i) \neq Y_i} \ .$$

In general, several models can be considered, and the *model selection* problem occurs. A classical answer is the penalization method (also called Structural Risk Minimization), initially introduced by Vapnik [**Vap82, Vap98**]. Basically, it states that one should choose the model $\widehat{m} \in \mathcal{M}_n$ which minimizes the sum of the empirical risk and some complexity term (the "penalty"):

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m)\} \ .$$

There has been many papers on this topic since the works of Vapnik, usually based on upper bounds on the "ideal penalty"

$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\,\gamma(\widehat{s}_m) \quad . \tag{7.1}$$

We can distinguish between two main kinds of results. First, the global viewpoint consists in bounding $\mathrm{pen}_{\mathrm{id}}(m)$ by the supremum of $(P - P_n)\gamma(t)$ over $t \in S_m$. This leads to penalties which measure the complexity of the entire model S_m, called *global penalties* or *global complexities*. For instance, one can build penalties upon the VC-dimension (see *e.g.* Lugosi [**Lug02**]), Rademacher complexities (independently introduced by Koltchinskii [**Kol01**] and Bartlett, Boucheron and Lugosi [**BBL02**]) or bootstrap (global) penalties (Fromont [**Fro04**]). Notice that the VC-dimension has the drawback of being independent from the distribution P of the data, so that it is adapted to the worst case. On the other hand, Rademacher complexities and bootstrap (global) penalties are measure dependent and lead to sharper bounds.

A second approach has been used more recently. It relies on the fact that the empirical risk minimizer is likely to have a small loss, so that a global upper bound on $\mathrm{pen}_{\mathrm{id}}(m)$ is over-pessimistic. This is the *localization* idea. Using for instance a link between variance and loss (the "margin condition", first introduced by Mammen and Tsybakov [**MT99**]), one can obtain much smaller bounds on the loss of the empirical risk minimizer (*cf.* Tsybakov [**Tsy04**] and Massart and Nédélec [**MN06**]). Then, several *localized complexity measures* (often based on Rademacher processes and called *Local Rademacher Complexities*) have been proposed, *e.g.* by Lugosi and Wegkamp [**LW04**], Bartlett, Bousquet and Mendelson [**BBM05**] and Koltchinskii [**Kol06**].

However, to our knowledge, none of these penalties can be used in practice since they depend on unknown quantities, or because they involve absolute constants on which we only know large upper bounds (see Sect. 2.2.1, page 77). In addition, the computational cost of local Rademacher complexities is prohibitive in general. This is why the most widely used model selection procedures in classification are based upon data splitting, in particular cross-validation. We refer to Yang [**Yan07**] and the references therein for the use of cross-validation in classification. In addition, Massart [**Mas07**] (Sect. 8.5) has shown that a simple method like hold-out (which is a primitive version of cross-validation) is naturally adaptive to Tsybakov's margin condition. It is then designed to compete with local penalties.

In Chap. 5 and 6, we propose new penalties (resp. called V-fold and Resampling penalties) which are a localized version of Fromont's bootstrap (global) penalties (and Rademacher complexities, since these are a particular case of bootstrap penalties). Indeed, while the bootstrap penalties are estimating $\sup_{t\in S_m}(P-P_n)\gamma(t)$, V-fold and Resampling penalties aim at estimating the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m)$ itself.

In a nutshell, their construction is the following. According to (7.1), the ideal penalty is a function $F(P, P_n)$ of the true distribution P and the empirical distribution $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. Then, the *resampling heuristics* (introduced by Efron [**Efr79**]) states that we can mimic the pair (P, P_n) with the pair (P_n, P_n^W), where P_n^W is the empirical distribution of a n-sample with distribution P_n ("the resample"). $\mathbb{E}_W[\cdot]$ denoting the expectation w.r.t. the randomness of the resampling (which is independent from the sample), the ideal penalty should be close to

$$\mathrm{pen}(m) := \mathbb{E}_W\left[F\left(P_n, P_n^W\right)\right] = \mathbb{E}_W\left[\left(P_n - P_n^W\right)\gamma\left(\widehat{s}_m^W\right)\right] \qquad \text{with} \quad \widehat{s}_m^W \in \arg\min_{t\in S_m}\left\{P_n^W\gamma(t)\right\} \quad . \tag{7.2}$$

This heuristics has then been generalized to other resampling schemes with the exchangeable weighted bootstrap (Mason and Newton [**MN92**] and Præstgaard and Wellner [**PW93**]), where

$$P_n^W := \frac{1}{n}\sum_{i=1}^{n} W_i \delta_{(X_i,Y_i)} \qquad \text{with} \quad W \in \mathbb{R}^n \quad \text{an exchangeable weight vector,}$$

independent from the data.

In Chap. 5 and 6, we prove that the penalization procedure based on $\mathrm{pen}(m)$ proportional to the one of (7.2) satisfies oracle inequalities with constant almost one, in the least-square regression framework, when all the models are made of histograms. Of course, histograms were not our final goal, and we would like to investigate the properties of Resampling Penalties in the classification case.

The main result of this chapter is a concentration inequality for

$$\widehat{p}_2(m) := \mathbb{E}_W\left[P_n^W\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(\widehat{s}_m^W\right)\right)\right] \ , \tag{7.3}$$

which is the resampling estimate of

$$p_2(m) := P_n\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(\widehat{s}_m^W\right)\right) \ .$$

The keystone of the slope heuristics (see Birgé and Massart [**BM06c**] and Chap. 3) is that $p_2(m)$ is the "minimal penalty", whereas the ideal penalty is close to $2p_2(m)$. If this holds in the classification case, we would only have to prove that

$$\mathbb{E}\left[p_2(m)\right] \propto \mathbb{E}\left[\widehat{p}_2(m)\right]$$

in order to derive oracle inequalities for the Resampling Penalization procedure described in Sect. 7.4.

The rest of the chapter is organized as follows. We precise the framework and some notations in Sect. 7.2. Our main results are stated in Sect. 7.3, where we also explain what remains to be proven about Resampling Penalization in classification. Then, Sect. 7.4 tackles practical issues. The proofs are given in Sect. 7.5.

7.2. Framework

The framework of this chapter is the general statistical learning framework, as in Chap. 8 of Massart [**Mas07**] (see also Massart and Nédélec [**MN06**], Sect. 2). It is slightly more general than binary classification, and includes the bounded regression framework.

7.2.1. General framework. We observe i.i.d. variables $\xi_1, \ldots, \xi_n \in \mathcal{X} \times \mathcal{Y}$ with common distribution P, where $\mathcal{X} \times \mathcal{Y}$ is some measurable space and $\mathcal{Y} \subset [0,1]$ (in binary classification, $\mathcal{Y} = \{0,1\}$). Our goal is to build a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$ such that given a new data $\xi = (X,Y)$ (with Y unobserved), $t(X)$ is a good prediction for Y. The set of all predictors is denoted by $\mathcal{S}$.

The quality of a predictor t is measured by its *prediction loss* $P\gamma(t) := \mathbb{E}_{\xi\sim P}\left[\gamma(t,\xi)\right]$ where $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto [0,1]$ is some *contrast* function. A popular choice is $\gamma(t,(x,y)) = (t(x)-y)^2$, which coincides with the 0-1 contrast $\gamma(t,(x,y)) = \mathbb{1}_{t(x)\neq y}$ in binary classification. The best predictor is then the Bayes predictor s, which minimizes $P\gamma(t)$ over $\mathcal{S}$. Instead of the prediction loss, we often consider the *excess loss*

$$l(s,t) := P\gamma(t) - P\gamma(s) \ .$$

Defining the regression function as $\eta(x) = \mathbb{E}\left[Y \mid X = x\right]$ for every $x \in \mathcal{X}$, we have

$$\forall x \in \mathcal{X}, \qquad s(x) = \mathbb{1}_{\eta(x)\geq \frac{1}{2}}$$

in the binary classification case.

As described in Introduction, given a set of predictors S_m (a model), we define the empirical risk minimizer on S_m as

$$\widehat{s}_m \in \arg\min_{t\in S_m} \{ P_n\gamma(t) \} = \arg\min_{t\in S_m} \left\{ \frac{1}{n}\sum_{i=1}^{n} \gamma(t,\xi_i) \right\} \qquad \text{where} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i} \ .$$

Given a family of models $(S_m)_{m\in\mathcal{M}_n}$, the purpose of model selection is to choose a (data-dependent) $\widehat{m} \in \mathcal{M}_n$ such that $\widehat{s}_{\widehat{m}}$ is a good predictor, or at least as good as the best predictor among the family $(\widehat{s}_m)_{m\in\mathcal{M}_n}$.

We now assume that $\widehat{m}$ is chosen according to some penalization procedure:

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \{ P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \} \tag{7.4}$$

where the penalty is any function $\mathrm{pen} : \mathcal{M}_n \mapsto [0,\infty)$, possibly data-dependent. From this definition of $\widehat{m}$, we can prove that

$$l(s,\widehat{s}_{\widehat{m}}) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(\widehat{m}) \le \inf_{m\in\mathcal{M}_n} \{ l(s,\widehat{s}_m) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(m) \} \ , \tag{7.5}$$

where the ideal penalty $\mathrm{pen}_{\mathrm{id}}$ is defined by (7.1). We can thus obtain an oracle inequality by proving that pen is close to $\mathrm{pen}_{\mathrm{id}}$ for every model $m \in \mathcal{M}_n$. Obviously, (7.5) also holds with $\mathrm{pen}'_{\mathrm{id}} = \mathrm{pen}_{\mathrm{id}} + (P_n - P)\gamma(s)$ instead of $\mathrm{pen}_{\mathrm{id}}$. This alternative "ideal penalty" is more convenient for the proof (it has better concentration properties), and can be split into $\mathrm{pen}'_{\mathrm{id}} = p_1 + p_2 - \overline{\delta}$ with

$$p_1(m) := P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) \qquad p_2(m) := P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$$
$$\text{and} \qquad \overline{\delta}(m) := \left(P_n - P\right)\left(\gamma(s_m) - \gamma(s)\right) \ .$$

7.2.2. Main assumptions. Remember that the first assumption in this chapter is that the contrast γ takes its values in $[0,1]$. Then, we assume that there exists some pseudo-distance d on $\mathcal{S}$ (which may depend on P) such that

$$\forall t \in \mathcal{S}, \qquad \mathrm{var}_P\left[\gamma(t,\cdot) - \gamma(s,\cdot)\right] \le d^2(s,t) \ . \tag{7.6}$$

Let $\mathcal{C}_1$ be the set of nondecreasing and continuous functions $\psi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $x \mapsto \psi(x)/x$ is nonincreasing on $(0,+\infty)$ and $\psi(1) \ge 1$. Our third assumption is that there is a function $w \in \mathcal{C}_1$ such that for every $\epsilon > 0$,

$$\sup_{t\in S, l(s,t)\le\epsilon^2} d(s,t) \le w(\epsilon) \ . \tag{7.7}$$

Combined with (7.6), this *margin condition* generalizes the one introduced by Tsybakov. It links the variance of the process $\gamma(t,\cdot) - \gamma(s,\cdot)$ with its expectation (which is equal to the excess loss at t).

Finally, we need an assumption in order to control the sizes of the models S_m. This is the following: for every $m \in \mathcal{M}_n$, there is a function $\phi_m \in \mathcal{C}_1$ such that for every $u \in S_m$ and every $\sigma > 0$ such that $\phi_m(\sigma) \le \sqrt{n}\sigma^2$,

$$\sqrt{n}\mathbb{E}\left[\sup_{t\in S_m,\, d(u,t)\le\sigma} \{ \left(P_n - P\right)\left(\gamma(u) - \gamma(t)\right) \} \right] \le \phi_m(\sigma) \ . \tag{7.8}$$

Remark that when S_m is uncountable, measurability issues may occur in (7.8). We refer to Sect. 2.2 in [**MN06**] where a separability condition is introduced for this purpose.

We can then define $\varepsilon_{\star,m}$ (or $\varepsilon_{\star,m,(n)}$, when we want to emphasize the dependence on the sample size n) as the unique positive solution of the equation

$$\sqrt{n}\varepsilon_{\star,m}^2 = \phi_m(w(\varepsilon_{\star,m})) \ . \tag{7.9}$$

According to Massart and Nédélec [**MN06**], this quantity measures the quality of an estimator $\widehat{s}_m$. As we shall see in the following, it also appears in the remainder terms of some concentration inequalities.

7.2.3. Binary classification. The main example of this chapter is binary classification. We then have $\mathcal{Y} = \{0,1\}$ and consider the 0-1 contrast $\gamma(t,(x,y)) = \mathbb{1}_{t(x)\neq y}$, which takes its values in $[0,1]$. In (7.6), the pseudo-distance d can be chosen as the $L^2(\mathcal{L}(X))$ distance

$$d^2(t,s) = \mathbb{E}\left[\left(t(X)-s(X)\right)^2\right] \ ,$$

since for every $t \in \mathcal{S}$,

$$\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad |\gamma(t,(x,y)) - \gamma(s,(x,y))| = \left|\mathbb{1}_{t(x)\neq y} - \mathbb{1}_{s(x)\neq y}\right| \leq |t(x)-s(x)| \ .$$

We can derive (7.7) with $w(\epsilon) = h^{-1/(2\theta)}\epsilon^{1/\theta}$ from Tsybakov margin condition

$$l(s,t) \geq h\left(\mathbb{E}\left|s(X)-t(X)\right|\right)^\theta = hd(s,t)^{2\theta} \qquad \text{for some} \quad \theta \geq 1 \text{ and } h \leq 1 \ .$$

Noticing that the excess loss is equal to

$$l(s,t) = \mathbb{E}\left[\left|2\eta(X)-1\right|\left|s(X)-t(X)\right|\right] \ ,$$

so that (7.7) holds with $w(\epsilon) = h^{-1/2}\epsilon$ under the simpler margin condition

$$\mathbb{P}\left(\left|2\eta(X)-1\right| \geq h > 0\right) = 1 \ .$$

Then, we only have to compute some ϕ_m satisfying (7.8) for each model S_m. This is done under several assumptions in Sect. 2.4 in [**MN06**].

7.2.4. Bounded regression. In addition, bounded regression fits with the above general framework. We then have $\mathcal{Y} = [0,1]$ (up to some rescaling of the data) and choose the least-square contrast $\gamma(t,(x,y)) = (t(x)-y)^2$, so that the Bayes predictor s is the regression function η. Notice that γ has its values in $[0,1]$.

It is straightforward to compute

$$l(s,t) = \mathbb{E}\left[\left(t(X)-s(X)\right)^2\right] \qquad \text{and} \qquad d^2(s,t) = 2\mathbb{E}\left[\left(t(X)-s(X)\right)^2\right]$$

which satisfies (7.6). As a consequence, the margin condition (7.7) automatically holds in bounded regression with $w(\epsilon) = \sqrt{2}\epsilon$.

In the particular case of histograms (that we have considered in the previous chapters), we can show that (7.8) holds with

$$\phi_m(\sigma) = \sigma\left(1 + \frac{\sqrt{D_m}}{\sqrt{2}}\right). \tag{7.10}$$

This fact comes from (3.38) in Sect. 3.5.7, where it is completely proven. It is then straightforward to compute

$$\varepsilon_{\star,m} = \frac{\sqrt{2}+\sqrt{D_m}}{\sqrt{n}} \ . \tag{7.11}$$

See also Sect. 2.3 in [**MN06**] for computations in the "binary images" framework.

7.2.5. Resampling schemes. In introduction, we defined two resampling quantities by (7.2) and (7.3). They depend on some resampling weight vector $W \in \mathbb{R}^n$, assumed to be independent from the data.

In the following, we will often assume that W is a *subsampling weight vector*. This means that we can write $W_i = \kappa \mathbb{1}_{i \in I}$ for some random subset I of $\{1, \dots, n\}$, with $\kappa = n/card(I)$. Up to some multiplicative constant, this coincides with the "Bootstrap without replacement" defined by van der Vaart and Wellner [**vdVW96**] (see Example 3.6.14). On subsampling, see also the book from Politis, Romano and Wolf [**PRW99**]. In other words, the "resample" is a subsample of the entire data set. The main examples or subsampling weights are the following:

- *Random hold-out (q)*, $q \in \{1, \dots, n\}$: I is chosen uniformly among subsets of $\{1, \dots, n\}$ of size q, and $\kappa = nq^{-1}$. A classical choice is $q = n/2$.
- *Hold-out (q)*, $q \in \{1, \dots, n\}$: $I \subset \{1, \dots, n\}$ is deterministic with cardinality q, and $\kappa = nq^{-1}$.
- *V-fold cross-validation*, $V \in \{1, \dots, n\}$: let $(B_j)_{1 \leq j \leq V}$ be a partition of $\{1, \dots, n\}$, then $I = \{1, \dots, n\} \backslash B_J$ with $J \sim \mathcal{U}(\{1, \dots, n\})$ independent from the data, and $\kappa = n/(n - \mathrm{Card}(B_J))$. It is classical to assume that the partition is regular, and then $\kappa = V/(V-1)$.

We have already defined these weights (with $\kappa = V/(V-1)$ even if the partition is not regular in the VFCV case) in Chap. 5 and 6, where we explain the links between V-fold penalties (when the partition is regular) and the classical V-fold cross-validation procedure. Notice that the small difference between the above VFCV weights and the ones of Sect. 5.3.1 is small if the partition is "almost regular". We can then expect the resulting V-fold penalties to have the same behaviour in practice.

The particular shape of subsampling weights allows us to write the resampling procedure in a different way. The resampling empirical distribution P_n^W is equal to

$$P_n^W = \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{\xi_i} = \frac{\kappa}{n} \sum_{i \in I} \delta_{\xi_i} = \frac{1}{\mathrm{Card}(I)} \sum_{i \in I} \delta_{\xi_i} =: P_n^{(I)}$$

the empirical distribution of the subsample $(\xi_i)_{i \in I}$. Moreover, since I is chosen independently from the sample, *the subsample* $(\xi_i)_{i \in I}$ *is an i.i.d. sample of size* $\mathrm{Card}(I)$ *with common distribution* P. This is the key property that we will use to prove our concentration inequality for $\widehat{p}_2$.

7.3. Main results

7.3.1. A recipe of oracle inequalities. As noticed in Sect. 7.2.1, it is sufficient to prove that pen is close to $\mathrm{pen}'_{\mathrm{id}} = p_1 + p_2 - \overline{\delta}$ in order to derive some oracle inequality for the procedure defined by (7.4). Following the proof of Thm. 3.1 in Sect. 3.3 (or equivalently Thm. 6.1 in Sect. 6 or Thm. 5.1 in Sect. 5), we can describe a recipe for proving oracle inequalities for penalization procedures:

(1) Concentration inequality for $\overline{\delta}(m)$ around its expectation.
(2) Concentration inequality for $p_2(m)$ around its expectation.
(3) Concentration inequality for $\mathrm{pen}(m)$ around its expectation.
(4) $\mathbb{E}[\mathrm{pen}(m)] \approx \mathbb{E}[2p_2(m)]$.
(5) With large probability, $p_2(m)$ is close to $p_1(m)$.
(6) All the remainder terms are negligible in front of $l(s, s_m) + \mathbb{E}[p_2(m)]$.

In this chapter, we prove the three first steps for a resampling penalty $\mathrm{pen}(m) = 2C_W \widehat{p}_2(m)$ with subsampling weights (the constant C_W should only depend on the resampling scheme $\mathcal{L}(W)$, so that step 4 holds true). This is the object of the next subsection.

7.3.2. Concentration inequalities. It is not new that steps 1 and 2 can be solved in such a general framework. First, we only need Bernstein's inequality to derive concentration properties for $\overline{\delta}(m)$. We recall below this classical result, which is proven for instance in Sect. 3.5.6.

PROPOSITION 7.1 (Prop. 3.3, Sect. 3.3). *Assume that γ has its values in $[0,1]$. Then for all $x \geq 0$, on an event of probability at least $1 - 2e^{-x}$:*

$$\forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta l(s, s_m) + \left(\frac{1}{\eta} + \frac{2}{3}\right)\frac{x}{n} \ . \tag{7.12}$$

The second step is quite harder to solve, but all the work has been done recently by Boucheron and Massart [**BM04**] (with Thm. 2.2 in a preliminary version).

PROPOSITION 7.2. *Let $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto [0,1]$ be a contrast function, $(\xi_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ some i.i.d. data with common distribution P, and $(S_m)_{m \in \mathcal{M}_n}$ be a family of models. Make all the assumptions of Sect. 7.2.2,* i.e. (7.6), (7.7) *and* (7.8). *As in* (7.9), *define $\varepsilon_{\star,m}$ the unique positive solution of the equation*

$$\sqrt{n}\varepsilon_{\star,m}^2 = \phi_m(w(\varepsilon_{\star,m})) \ .$$

Let $p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$.

Then, there is a constant $C > 0$ such that, for every $x \geq 0$, there exists a set of probability at least $1 - e^{1-x}$ on which

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq \frac{C}{\sqrt{n}}\left[\sqrt{2ex}w\left(\sqrt{l(s,s_m) \vee \varepsilon_{\star,m}^2}\right) + 2exw\left(\frac{w(\varepsilon_{\star,m})}{\sqrt{n}\varepsilon_{\star,m}}\right)\right] \ . \tag{7.13}$$

The third step is the main result of this chapter. In a few words, its proof is based upon Prop. 7.2 and the key remark that $(\xi_i)_{i \in I}$ is an i.i.d. sample of size $\mathrm{Card}(I)$ with common distribution P (see the end of Sect. 7.2.5).

THEOREM 7.1. *Let $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto [0,1]$ be a contrast function, $(\xi_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ some i.i.d. data with common distribution P, and $(S_m)_{m \in \mathcal{M}_n}$ be a family of models. Make all the assumptions of Sect. 7.2.2,* i.e. (7.6), (7.7) *and* (7.8). *Moreover, we assume that for every $m \in \mathcal{M}_n$, ϕ_m is independent from n (in the sense that* (7.8) *holds for any sample size $q \leq n$). As in* (7.9), *define $\varepsilon_{\star,m}$ the unique positive solution of the equation*

$$\sqrt{n}\varepsilon_{\star,m}^2 = \phi_m(w(\varepsilon_{\star,m})) \ .$$

Let $\widehat{p}_2(m)$ be defined by (7.3), *with some subsampling weights W. Let $B > 0$ such that*

$$B \geq \frac{n}{\mathrm{Card}\left\{1 \leq i \leq n \text{ s.t. } W_i \neq 0\right\}} \qquad a.s. \ .$$

Then, there is a constant $C > 0$ such that for every $x \geq 0$, there exists an event of probability at least $1 - e^{1-x}$ on which

$$|\widehat{p}_2(m) - \mathbb{E}\left[\widehat{p}_2(m)\right]| \leq \frac{C}{\sqrt{n}}\left[B\sqrt{2ex}w\left(\sqrt{l(s,s_m) \vee \varepsilon_{\star,m}^2}\right) + 2exw\left(\frac{w(\varepsilon_{\star,m})}{\sqrt{n}\varepsilon_{\star,m}}\right)\right] \ . \tag{7.14}$$

REMARK 7.1.

- The assumption that ϕ_m is independent from n is rather mild, because of the scaling $\sqrt{n}$. Indeed, computations made by Massart and Nédélec ([**MN06**], Sect. 2.4) show that ϕ_m can be bounded independently from n, with the universal metric entropy (*e.g.* when

S_m is a VC-class) or the $L^1(\mathcal{L}(X))$ entropy with bracketing. The resulting bounds for $\varepsilon_{\star,m}$ are sharp since they are equal (up to some logarithmic factor) to the minimax lower bounds.
Notice that this assumption is also satisfied in regression when S_m is an histogram (see Sect. 7.2.4) or in the binary image case when S_m is a finite dimensional vector space (see [**MN06**], Sect. 2.3).

- The constant B appearing in the upper bound is smaller than 2 for V-fold cross-validation (with partitions such that $\max_j \mathrm{Card}(B_j) \leq n/2$), hold-out ($q$) and Random hold-out ($q$) if $q \geq n/2$ (*e.g.* Leave-one-out).
Thm. 7.1 thus shows that $\widehat{p}_2$ concentrates almost as well as p_2. However, our remainder term can not be used for choosing a resampling scheme, since we showed in Sect. 6.8.7 that $\widehat{p}_2$ concentrates better than p_2 when the weights are exchangeable.
- A similar result with Rademacher weights (*i.e.* $2W_i$ i.i.d. binomial with parameter 1/2) may be proven, up to some small additional work. These are indeed almost subsampling weights (the only difference is that $I := \{1 \leq i \leq n \text{ s.t. } W_i \neq 0\}$ has a random cardinality). Another technical issue is that $\mathrm{Card}(I)$ can be very small (and even equal to zero) with a positive probability, so that B defined in Thm. 7.1 would be infinite.

In order to evaluate the goodness of these concentration inequalities, let us consider the framework of least-square regression on histogram models. Remember that we assume $Y \in [0,1]$ a.s. (this can be done with any bounded histogram by translating and rescaling Y). According to Sect. 7.2.4, $w(\epsilon) = \epsilon\sqrt{2}$ and $\varepsilon_{\star,m} \leq L\sqrt{D_m/n}$ for some absolute constant[1] L. As a consequence, the remainder terms in (7.13) and (7.14) can be bounded as follows:

$$\begin{aligned}\frac{C}{\sqrt{n}}\left[\sqrt{2ex}w\left(\sqrt{l(s,s_m)\vee\varepsilon_{\star,m}^2}\right)+2exw\left(\frac{w(\varepsilon_{\star,m})}{\sqrt{n}\varepsilon_{\star,m}}\right)\right] &\leq L\sqrt{\frac{xl(s,s_m)}{n}}+L\frac{x+\sqrt{xD_m}}{n}\\ &\leq L\theta\left(l(s,s_m)+\frac{D_m}{n}\right)+L\left(1+\theta^{-1}\right)\frac{x}{n}\end{aligned} \tag{7.15}$$

for every $\theta > 0$.

Assume moreover that $\mathbb{E}[p_2(m)] \geq Q_m^{(p)} D_m n^{-1}$ for some $Q_m^{(p)} > 0$ (*cf.* Lemma 6.13 in Sect. 6.8.8). Then, taking $\theta = D_m^{-1/2}$ in (7.15), this remainder term is smaller than

$$\frac{l(s,s_m)}{\sqrt{D_m}}+\frac{1+x}{Q_m^{(p)}\sqrt{D_m}}\mathbb{E}[p_2(m)] \quad ,$$

which is negligible in front of $l(s,s_m)+\mathbb{E}[p_2(m)]$ if $D_m \geq \ln(n)^3$ and $x \leq L\ln(n)$.

Under the same assumptions, taking $\eta = D_m^{-1/2}$ in Prop. 7.1 gives a similar remainder term. We thus recover part of the results from Chap. 5 and 6 in the bounded case, under different assumptions on the resampling weights. This shows that step 6 (in the recipe of Sect. 7.3.1) is likely to be satisfied with these remainder terms.

7.3.3. Program for further research. In Sect. 7.3.1, we described a recipe for proving oracle inequalities, partially solved with the concentration inequalities stated in Sect. 7.3.2. We give here some comments on the remaining points.

[1] In the following, L denotes some absolute constant, possibly different from a line to another, or even within the same line

Asymptotically, for every fixed model m, steps 3 and 4 (with pen $= 2\widehat{p}_2$) are consequences of Thm. 3.6.13 of van der Vaart and Wellner [**vdVW96**]. This is not sufficient here, even for an asymptotic oracle inequality, because the collection of models $\mathcal{M}_n$ is allowed to depend on n. Notice that it is sufficient to consider exchangeable weights at step 4 because of Lemma 8.4 in Sect. 8.4.1. In the case of subsampling weights, it is thus sufficient to consider Random hold-out (q) weights for every $q \in \{1, \dots, n\}$.

The fifth step is the keystone of the slope heuristics of Birgé and Massart (*cf.* Chap. 3, Birgé and Massart [**BM06c**] and [**Mas07**], Sect. 8.5.2). To our knowledge, this points remains an open problem in the general case (and in particular for binary classification). A major difficulty here would be to prove lower bounds on $p_1(m) = P\gamma(\widehat{s}_m) - P\gamma(s_m)$ with large probability (up to the bias term, which is deterministic, $p_1(m)$ is equal to the excess loss $l(s, s_m)$).

The last step is also an open question, even for the concentration results stated in Sect. 7.3.2. If $\mathbb{E}[p_2(m)] \gg \ln(n)n^{-1}$, one can choose η in Prop. 7.1 so that step 6 holds for $\overline{\delta}$. For the deviations of $p_2(m)$ and $\widehat{p}_2(m)$, the remainder terms can always be chosen negligible in front of $\varepsilon_{\star,m}$. It seems to be a sharp complexity measure, at least in binary classification (Massart and Nédélec [**MN06**]), but we do not know if these remainder terms are actually small enough. Moreover, in the case of $\widehat{p}_2(m)$, remember that the remainders will be multiplied by some constant C_W like $2\widehat{p}_2$. If this multiplicative term is large (and this may occur for Random hold-out (q) when $q \sim n$, according to computations of Chap. 6), Thm. 7.1 may not be sufficient for step 6 to hold with pen $= 2C_W\widehat{p}_2$.

Finally, we propose an alternative to the recipe of Sect. 7.3.1 without step 5. Instead, assume that we can prove a moment inequality for p_1 similar to the one for p_2 (see [**BM04**], or the proof of Prop. 7.2). Then, the proof of Thm. 7.1 can be adapted in order to prove a concentration inequality for

$$\widehat{p}_1(m) := \mathbb{E}_W\left[P_n\left(\gamma\left(\widehat{s}_m^W\right) - \gamma\left(\widehat{s}_m\right)\right)\right]$$

very similar to (7.14). Thus, the remaining open problems are the following:

- $(2')$ a moment inequality for p_1 similar to the one of Boucheron and Massart [**BM04**] for p_2.
- $(4')$ a comparison of ideal and resampling penalties in expectations:

$$\mathbb{E}[p_1(m)] \approx C_{W,1}\mathbb{E}[\widehat{p}_1(m)] \qquad \mathbb{E}[p_2(m)] \approx C_{W,2}\mathbb{E}[\widehat{p}_2(m)]$$

  for some $C_{W,1}$ and $C_{W,2}$ depending only on the resampling scheme $\mathcal{L}(W)$.
- $(6')$ show that all the remainders are negligible in front of $l(s, \widehat{s}_m) = l(s, s_m) + p_1(m) \approx l(s, s_m) + \mathbb{E}[p_1(m)]$.

and the resampling penalty would be pen $= C_{W,1}\widehat{p}_1 + C_{W,2}\widehat{p}_2$. According to asymptotic results [**vdVW96**] and histogram computations (Sect. 6.8.6), it is likely that $C_{W,1} \approx C_{W,2} \approx C_{W,\infty}$.

Remark that the two suggested penalties $C_{W,\infty}(\widehat{p}_1 + \widehat{p}_2)$ and $2C_{W,\infty}\widehat{p}_2$ are quite close as soon as $p_1 \approx p_2$. Moreover, in the maximum-log-likelihood framework, Shibata [**Shi97**] showed that those two approaches (with bootstrap weights) are asymptotically equivalent. See Remark 6.3 in Sect. 6.3.3 for more details. However, according to our computations and simulation studies in the histogram regression case, p_1 is slightly larger than p_2, and similarly $\widehat{p}_1$ is slightly larger than $\widehat{p}_2$ (differences which disappear when n goes to infinity). Then, it is possible that $2C_{W,\infty}\widehat{p}_2$ has to be slightly enlarged in order to avoid underpenalization, at least from the non-asymptotic viewpoint.

7.4. Practical application

The (partial) results of Sect. 7.3 lead to the following algorithm.

ALGORITHM 7.1 (Resampling penalization).

(1) Choose a resampling scheme, *i.e.* the law $\mathcal{L}(W)$ of a weight vector W.
(2) Choose a constant $C \geq C_{W,\infty} \approx \left(n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(W_i-1\right)^2\right)^{-1}$.
(3) Compute the following resampling penalty for each $m \in \mathcal{M}_n$:

$$\mathrm{pen}(m) = 2C\widehat{p}_2(m) = 2C\mathbb{E}_W\left[P_n^W\left(\gamma\left(\widehat{s}_m\right)-\gamma\left(\widehat{s}_m^W\right)\right)\right] \quad . \tag{7.16}$$

(4) Minimize the penalized empirical criterion to choose $\widehat{m}$ and thus $\widehat{s}_{\widehat{m}}$:

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n}\left\{P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}(m)\right\} \quad .$$

According to Sect. 7.3.3 above, an alternative penalty shape could be

$$\mathrm{pen}(m) = C\left(\widehat{p}_1(m) + \widehat{p}_2(m)\right) \quad ,$$

as in algorithm 6.1 of Sect. 6.2.

Notice that algorithm 7.1 is a plug-in method, so that it should not be used when $\mathrm{Card}(\mathcal{M}_n)$ is too large (*i.e.* larger than any power of n). We refer to Sect. 6.6.2 for an answer to this problem.

7.4.1. Tuning parameters. There are two parameters in algorithm 7.1: the resampling scheme and the constant C.

Choice of C. First, notice that there has to be a constant C in front of the penalty, even if it does not appear in Efron's resampling heuristics. Indeed, as show asymptotically by van der Vaart and Wellner [**vdVW96**] (in their Thm. 3.6.13), the variance of the weights has to be taken into account for the resampling estimates to be unbiased. The approximation

$$C_{W,\infty} \approx \left(n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(W_i-1\right)^2\right)^{-1}$$

comes from their asymptotic result.

However, this asymptotic value for C may not be the right one when the sample size n is small, or even for large n because $\mathcal{M}_n$ may depend on n. We suggest two ways of computing $C_{W,\infty}$ non-asymptotically:

- use the constant $C_{W,\infty}$ computed in the histogram regression framework (see also Tab. 6.1 in Sect. 6.3.3 and Prop. 6.5 in Sect. 6.8.6; for non-exchangeable weights, see Lemma 8.4 in Sect. 8.4.1).
- use Birgé and Massart "slope heuristics" (for instance algorithm 3.1 in Sect. 3.4).

In both cases, we do not have enough theoretical evidence to prove that any of these methods work in classification (most of our proofs being restricted to histogram regression). We can only notice that the first one gives results very similar to the "asymptotic" calibration for most weights (*i.e.* all the classical ones except leave-one-out).

A second problem (which is also a major interest of our penalization method compared to V-fold cross-validation, see Sect. 5.5) is that an unbiased penalty is not necessarily optimal in a non-asymptotic framework. It is often better to overpenalize a little, so that the fluctuations of $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ does not make it negative with a significant probability (*cf.* Sect. 2.4.1 and 6.6.1).

The problem of overpenalization thus remains to estimate accurately these fluctuations, then decide a confidence level α, and finally choose C such that for every $m \in \mathcal{M}_n$, $\mathbb{P}(\mathrm{pen}(m) < \mathrm{pen}_{\mathrm{id}}(m)) \leq \alpha$. This needs at least a variance estimate of $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$.

Using the fact that our penalty is computed by resampling, we also propose in Sect. 6.6.1 two methods using a conditional $(1-\alpha)$ quantile instead of an expectation in (7.16). The main

advantage of this idea is that it does not need more computations[2], since we already have to consider several weight vectors in order to compute an expectation w.r.t. W.

Of course, there remains the question of choosing α, and we have no clear idea about this. We believe that this choice should anyway depend on each particular problem (there is no universally optimal choice for α), but simulation studies and theoretical results would be helpful for practical users. Further remarks on overpenalization and related problems are given in Sect. 11.3.3.

Choice of a resampling scheme. As for choosing C, we have no theoretical evidence in the classification framework about the choice of $\mathcal{L}(W)$. Indeed, the upper bound (7.14) in Thm. 7.1 is not sharp, since it does not show that resampling estimates (with exchangeable weights) concentrate better than the original quantities (p_2 in this case). This phenomenon is well-known asymptotically (see *e.g.* Hall [**Hal92**]), and appears in some non-asymptotic results (Prop. 6.8 in Sect. 6.8.7 and Prop. 10.4 in Sect. 10.2.3).

Even if our results are limited to subsampling weights, it is very likely that exchangeable weights (like Efron and Rademacher weights, see Sect. 6.3.3) are at least as efficient as subsampling ones.

In view of the results from Chap. 5 and 6 in the histogram case, we can sum up this question as follows. On the one hand, exchangeable weights are efficient but need a long computation time (at least n different weight vectors). Then, a classical method is to make a Monte-Carlo approximation (but then, do not use Leave-one-out weights). On the other hand, non-exchangeable weights such as hold-out and V-fold cross-validation ones need far less computation time. The choice of V is then quite similar the choice of the number of Monte-Carlo samples with the previous method: the larger V, the more accurate pen.

In both cases, the problem is to find the optimal trade-off between complexity (the number of resampling weight vectors to consider) and accuracy (measured by the fluctuations of the penalty around its expectation). See the simulation study of Sect. 5.4 (and the discussion of Sect. 5.5) for further clues on this question.

7.4.2. Comparison with other penalties. By construction, Resampling penalties are smaller than Fromont's bootstrap (global) penalties [**Fro04**]. Indeed,

$$\mathbb{E}_W\left[\left(P_n - P_n^W\right)\left(\gamma\left(\widehat{s}_m^W\right)\right)\right] \leq \mathbb{E}_W\left[\sup_{t\in S_m}\left(P_n - P_n^W\right)\left(\gamma\left(t\right)\right)\right] \ . \tag{7.17}$$

In particular, when W is a Rademacher resampling weight vector, this upper bound is equal to the well-known (global) Rademacher complexities (see [**Fro07**], Sect. 2.2):

$$\frac{1}{n}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n}\epsilon_i \mathbb{1}_{t(X_i)\neq Y_i}\,\middle|\,\xi_{1\ldots n}\right] \qquad \text{where } (\epsilon_i)_{1\leq i\leq n} \text{ are independent Rademacher variables.}$$

As a consequence, all the upper bounds on Rademacher complexities, for instance when the models are VC-classes, are still upper bounds on Resampling Penalties. This shows that algorithm 7.1 is not overpenalizing from the global viewpoint.

Moreover, inequality (7.17) highlights the reason why we can consider resampling penalties as "local" penalties, as compared to the global Rademacher complexities. Since local Rademacher complexities are resampling estimates of a "localized" upper bound on $\mathrm{pen}_{\mathrm{id}}(m)$, Resampling Penalties are likely to be even smaller (while being more natural, and easier to compute). It is then reasonable to think that algorithm 7.1 is efficient from the localized viewpoint (*i.e.* when the unknown distribution P has some good properties, such as Tsybakov's margin condition).

[2]whereas using (V-fold) cross-validation for tuning C would generally have a prohibitive computational cost.

7.5. Proofs

7.5.1. Proof of Prop. 7.2. According to Thm. 2.2 in [**BM04**], there exists some absolute constant C such that for every real number $q \geq 2$ one has

$$\|p_2(m) - \mathbb{E}[p_2(m)]\|_q \leq \frac{C}{\sqrt{n}} \left[\sqrt{q} w \left(\sqrt{l(s, s_m) \vee \varepsilon_{\star,m}^2} \right) + q w \left(\frac{w(\varepsilon_{\star,m})}{\sqrt{n}\varepsilon_{\star,m}} \right) \right] \quad . \tag{7.18}$$

With Markov inequality, it is classical to derive concentration inequalities from such moment inequalities. See for instance Lemma 8.10 in Sect. 8.6.2 for a complete proof. The result follows. □

7.5.2. Proof of Thm. 7.1.

General resampling weights. We first state a simple inequality (7.19) that is valid for any resampling scheme (only assumed to be independent from the sample $\xi_{1\dots n}$). It is quite similar to (5.55) (in Sect. 5.7.4). Let $\|\cdot\|_q$ be the q-th moment, and for every $W \in \mathbb{R}^n$ and $\xi_1, \dots, \xi_n \in \mathcal{X} \times \mathcal{Y}$,

$$F(W; \xi_{1\dots n}) := P_n^W \left(\gamma\left(\widehat{s}_m\right) - \gamma\left(\widehat{s}_m^W\right) \right) \quad .$$

By Jensen inequality, for every $q \geq 1$,

$$\begin{aligned} \|\widehat{p}_2(m) - \mathbb{E}\left[\widehat{p}_2(m)\right]\|_q^q &= \mathbb{E}\left(\left| \mathbb{E}_W\left[F(W; \xi_{1\dots n})\right] - \mathbb{E}\left[F(W; \xi_{1\dots n})\right] \right|^q \right) \\ &\leq \mathbb{E}\left(\left| F(W; \xi_{1\dots n}) - \mathbb{E}\left[F(W; \xi_{1\dots n}) \mid W\right] \right|^q \right) \\ &\leq \sup_{W_0 \in \mathrm{supp}(W)} \left\{ \mathbb{E}\left(\left| F(W_0; \xi_{1\dots n}) - \mathbb{E}\left[F(W_0; \xi_{1\dots n})\right] \right|^q \right) \right\} \\ &= \sup_{W_0 \in \mathrm{supp}(W)} \left\{ \left\| F(W_0; \xi_{1\dots n}) - \mathbb{E}\left[F(W_0; \xi_{1\dots n})\right] \right\|_q^q \right\} \end{aligned} \tag{7.19}$$

As a consequence, it is sufficient to prove moment inequalities for $\widehat{p}_2(m) - \mathbb{E}\left[\widehat{p}_2(m)\right]$ when W is deterministic (equal to some $W_0 \in \mathbb{R}^n$. The result follows only by taking a supremum over the support of W.

Remark that (7.19) would be valid for any measurable function $F : \mathbb{R}^n \times (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathbb{R}$. It can for instance be applied to $\widehat{p}_1(m) = P_n\left(\gamma\left(\widehat{s}_m^W\right) - \gamma\left(\widehat{s}_m\right)\right)$.

Subsampling weights. We now assume that W is a subsampling weight vector, *i.e.* $W_i = \kappa \mathbb{1}_{i \in I}$ for some random $I \subset \{1, \dots, n\}$ and $\kappa = n/\operatorname{Card}(I)$. According to (7.19), we only have to consider deterministic subsampling weights, *i.e.* such that I is deterministic.

Then, $F(W; \xi_{1\dots n})$ is equal to $p_2(m; (\xi_i)_{i \in I})$, *i.e.* $p_2(m)$ computed with the subsample $(\xi_i)_{i \in I}$ instead of the whole sample $\xi_{1\dots n}$. Since $(\xi_i)_{i \in I}$ has the distribution of an i.i.d. sample of size $\operatorname{Card}(I)$ with common distribution P, we can apply (7.18) from the proof of Prop. 7.2 (which comes from Boucheron and Massart [**BM04**]):

$$\|F(W; \xi_{1\dots n}) - \mathbb{E}\left[F(W; \xi_{1\dots n})\right]\|_q \leq \frac{C}{\sqrt{n}} \left[\sqrt{q} w \left(\sqrt{l(s, s_m) \vee \varepsilon_{\star,m,(\operatorname{Card}(I))}^2} \right) + q w \left(\frac{w(\varepsilon_{\star,m,(\operatorname{Card}(I))})}{\sqrt{n}\varepsilon_{\star,m,(\operatorname{Card}(I))}} \right) \right] \quad .$$

Since ϕ_m can also be used with a sample size $\operatorname{Card}(I) \leq n$, we can use (7.21) from Lemma 7.3 (in Sect. 7.5.3):

$$\varepsilon_{\star,m,(n)} \leq \varepsilon_{\star,m,(\operatorname{Card}(I))} \leq \left(\frac{n}{\operatorname{Card}(I)} \right)^{1/2} \varepsilon_{\star,m,(n)} \quad .$$

Using that $w \in \mathcal{C}_1$, we easily derive

$$\|F(W; \xi_{1\dots n}) - \mathbb{E}\left[F(W; \xi_{1\dots n})\right]\|_q \leq \frac{C}{\sqrt{n}} \left[\sqrt{q} \frac{n}{\operatorname{Card}(I)} w \left(\sqrt{l(s, s_m) \vee \varepsilon_{\star,m,(n)}^2} \right) + q w \left(\frac{w(\varepsilon_{\star,m,(n)})}{\sqrt{n}\varepsilon_{\star,m,(n)}} \right) \right] \quad .$$

Since $n/\operatorname{Card}(I) \leq B$ a.s., the result follows (see *e.g.* Lemma 8.10 in Sect. 8.6.2 for the link between moment and concentration inequalities). □

7.5.3. $\varepsilon_{\star,m}$ as a function of n and m. In this section, we show how $\varepsilon_{\star,m}$ depends on the sample size n and the models m. Since the function $\phi_m ow$ belongs to the class $\mathcal{C}_1$, this is a consequence of the following lemma.

LEMMA 7.3. *For any $n \in \mathbb{N}$ and $f \in \mathcal{C}_1$,* i.e. *$f : \mathbb{R}^+ \mapsto \mathbb{R}^+$ nondecreasing such that $x \mapsto f(x)/x$ is nonincreasing, define $\epsilon(n, f)$ the unique positive solution of the equation $\sqrt{n}\epsilon^2 = f(\epsilon)$.*

Then, for any $\lambda > 0$,

$$\left(\sqrt{\lambda} \wedge \lambda\right) \epsilon(n, f) \leq \epsilon(n, \lambda f) \leq \left(\sqrt{\lambda} \vee \lambda\right) \epsilon(n, f) \ . \tag{7.20}$$

In addition, for any $p \in \{1, \ldots, n\}$,

$$\left(\frac{n}{p}\right)^{1/4} \epsilon(n, f) \leq \epsilon(p, f) \leq \sqrt{\frac{n}{p}} \epsilon(n, f) \ . \tag{7.21}$$

In particular, $n \mapsto \varepsilon_{\star,m,(n)}$ is decreasing.

PROOF OF LEMMA 7.3. We first prove (7.20). Assume for instance that $\lambda \geq 1$ (the case $\lambda < 1$ is a consequence of it). The mapping $x \mapsto \frac{f(x)}{x^2}$ is nonincreasing on $(0, \infty)$. It is equal to $\sqrt{n}$ when $x = \epsilon(n, f)$ and $\sqrt{n}/\lambda$ when $x = \epsilon(n, \lambda f)$. This shows $\epsilon(n, f) \leq \epsilon(n, \lambda f)$.

Now, we use that $x \mapsto \frac{f(x)}{x}$ is nonincreasing:

$$\frac{\sqrt{n}}{\lambda} \epsilon(n, \lambda f) = \frac{f(\epsilon(n, \lambda f)}{\epsilon(n, \lambda f)} \leq \frac{f(\epsilon(n, f)}{\epsilon(n, f)} = \sqrt{n}\epsilon(n, f)$$

that is the right-hand inequality. Then, f being increasing,

$$\sqrt{n}\epsilon(n, f)^2 = f(\epsilon(n, f)) \leq f(\epsilon(n, \lambda f)) = \frac{\sqrt{n}}{\lambda} \epsilon(n, \lambda f)^2$$

and the left-hand inequality follows.

We derive (7.21) noticing that

$$\epsilon(p, f) = \epsilon\left(n, \sqrt{\frac{n}{p}} f\right) \ .$$

□

CHAPTER 8

Appendix on resampling penalties

RÉSUMÉ. Nous avons réuni dans cet appendice des commentaires ou résultats additionnels à ceux des trois chapitres précédents, ainsi que des lemmes ou preuves techniques que nous y utilisons. Les notations utilisées ici sont celles de ces trois chapitres.

This chapter is dependent from Chap. 5 to 9, and should not be read apart from them, except for some tools that may be used in other frameworks. Throughout this chapter, we use several notations defined in Chap. 5 and 6.

It is organized as follows. Sect. 8.1 and 8.2 provide some comments and suggestions about the way of using RP, and the reason why it may work in general. Then, we give a few more results about Resampling Penalization in Sect. 8.3 and 8.4. Sect. 8.5 to 8.9 contain several probabilistic tools and proofs of probabilistic results used in the previous chapters. Finally, we state and prove some tools that belongs to approximation theory in Sect. 8.10.

8.1. Uniqueness and existence of $\widehat{s}_m^W$

In Sect. 5.3.1 and 6.2, we defined general penalization algorithms based on the resampling heuristics (algorithms 5.1 and 6.1). In the case of least-square regression on histograms, we have to modify these general algorithms because

$$\widehat{s}_m\left(P_n^W\right) = \arg\min_{t\in S_m}\left\{P_n^W\gamma\left(t,\cdot\right)\right\}$$

is not uniquely defined for a.e. weight vector W. This leads to algorithms 5.2 and 6.2. In this section, we try to give a more general answer to this issue.

8.1.1. Existence. In the histogram case, there always exists a minimizer of $P_n^W\gamma\left(t,\cdot\right)$ in S_m. If such an existence issue happened, we may replace $\widehat{s}_m\left(P_n^W\right)$ by a ρ-minimizer of $P_n^W\gamma\left(t,\cdot\right)$ in S_m, for instance with $\rho = n^{-2}$.

8.1.2. Uniqueness: histogram case. In the histogram case, $\widehat{s}_m\left(P_n^W\right)$ is well-defined if and only if for every $\lambda\in\Lambda_m$, $\widehat{p}_\lambda^W > 0$, *i.e.* $W_\lambda = (n\widehat{p}_\lambda)^{-1}\sum_{X_i\in I_\lambda} W_i > 0$. This has a positive probability for any resampling scheme among Efron (q) $(q\geq 1)$, Rademacher(p) $(p\in(0,1))$, Poisson(μ) $(\mu>0)$ and Random hold-out (q) $(q\leq n-n\widehat{p}_\lambda)$. We thus have to find a way to define

$$\mathrm{pen}(m) = \mathbb{E}_W\left[P_n\gamma\left(\widehat{s}_m\left(P_n^W\right)\right) - P_n^W\gamma\left(\widehat{s}_m\left(P_n^W\right)\right)\right] \ .$$

Replacing $P_n\gamma(\widehat{s}_m(P_n^W)) - P_n^W\gamma(\widehat{s}_m(P_n^W))$ by its supremum over all possible values of $\widehat{s}_m(P_n^W)$ would lead to infinite penalties when the models are not bounded, and very large ones in the bounded case. For the same reason, choosing a fixed arbitrary value β_λ^0 for $\widehat{s}_m(P_n^W)$ on I_λ such that $P_n^W(I_\lambda) = 0$ is not relevant. A good alternative to this may be β_λ^0 equal to $\mathbb{E}_{P_n^W}(Y)$ or $\mathbb{E}_{P_n^W}\left(Y \mid d(X,I_\lambda)\leq h\right)$. However, a wise choice of h may be quite a difficult problem.

Another natural answer to this problem may be "consider only weight vectors such that no problem arises". However, if $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}$ is not very large, this may be a considerable constraint and the weight vector W is no longer exchangeable under this conditionning. Our proposal here is to adapt this idea under the light of the following crucial remark. Like the ideal penalty, $P_n\gamma(\widehat{s}_m(P_n^W)) - P_n^W\gamma(\widehat{s}_m(P_n^W))$ is the sum over $\lambda\in\Lambda_m$ of terms that only depend on what happens on I_λ (see (5.7) in Sect. 5.3.1). Thus, the penalty may also be written as

$$\mathrm{pen}(m,P_n) = \mathbb{E}_W\left[P_n\gamma(\widehat{s}_m(P_n^W)) - P_n^W\gamma(\widehat{s}_m(P_n^W))\right] = \sum_{\lambda\in\Lambda_m}\mathbb{E}_W\left[\mathrm{pen}_{\lambda,W}(m,P_n)\right] \ .$$

Now, each term $\mathrm{pen}_{\lambda,W}(m,P_n)$ is well-defined if and only if $\sum_{X_i\in I_\lambda}W_i>0$. Computing each expectation separately and conditionally to $W_\lambda>0$, the following candidate penalty is at least well-defined:

$$\mathrm{pen}(m,P_n) = \sum_{\lambda\in\Lambda_m}\mathbb{E}_W\left[\mathrm{pen}_{\lambda,W}(m,P_n)\,\Big|\,\sum_{X_i\in I_\lambda}W_i>0\right] \tag{8.1}$$

and such "local conditionnings" are quite mild constraints on the weights. This is the strategy of algorithm 6.2. Notice that (8.1) leads to a penalty slightly larger than the previous proposed one when $\beta_\lambda^0 = \mathbb{E}_{P_n}[Y\mid X\in I_\lambda]$. Indeed, this remains to posing $\mathrm{pen}_{\lambda,W}(m,P_n)=0$ for each W such that $W_\lambda=0$, which may lead to underpenalization if $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}$ is small. However, the results of Sect. 6.4 are valid for both penalties.

8.1.3. Uniqueness: general case. We now propose another natural modification of the penalty, that is modify the law of the weight vector W. When the weights do not have to be exchangeable (as in the V-fold case), this leads to algorithm 5.2. When we would like to keep the exchangeability of the weights, we propose the following. Let $\epsilon>0$ and $P_n^{W(0)},\dots,P_n^{W(k)},\dots$ be independent copies of the resampling empirical distribution. Then, replace P_n^W in the definition of the penalty by

$$P_n^{W\epsilon} = \frac{1}{1-\epsilon}\sum_{k\geq 0}\epsilon^k P_n^{W(k)} = \frac{1}{n}\sum_{i=1}^n\left(\frac{1}{1-\epsilon}\sum_{k=0}^\infty\epsilon^k W_i^{(k)}\right)\delta_{(X_i,Y_i)} \ .$$

Since each W_i is square-integrable, this is well defined for every $\epsilon\in[0;1)$ since for all $i\in\{1,\dots,n\}$, $W_i^\epsilon := \sum_k\epsilon^k W_i^{(k)}$ converges a.s. Moreover, for every $\epsilon>0$, $\mathbb{P}(W_i^\epsilon>0)>0$ so that $W_i^\epsilon>0$ a.s.

This ensures that $\mathrm{supp}\,P_n^{W\epsilon} = \mathrm{supp}\,P_n$. Then, in the histogram case, the penalty

$$\mathrm{pen}_\epsilon(m,P_n) = \mathbb{E}_W\left[P_n\gamma(\widehat{s}_m(P_n^{W\epsilon})) - P_n\gamma(\widehat{s}_m(P_n))\right]$$

is well-defined if and only if $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}>0$. Our suggest (which may be applied in more general cases than histograms) is to take

$$\mathrm{pen}(m,P_n) = \limsup_{\epsilon\to 0}\mathrm{pen}_\epsilon(m,P_n) \ . \tag{8.2}$$

As shown in the lemma below, this is the same as definition (8.1) in the histogram case.

LEMMA 8.1. *If S_m is an histogram model, the definitions* (8.1) *and* (8.2) *coincide.*

PROOF. As S_m contains all piecewise constant functions on some fixed partition $(I_\lambda)_{\lambda\in\Lambda_m}$, pen_ϵ may be written as the sum over $\lambda\in\Lambda_m$ of terms that only depend on the restrictions of P_n

and $P_n^{W\epsilon}$ to I_λ:

$$\mathrm{pen}_\epsilon(m,P_n) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\mathrm{pen}_{\epsilon,\lambda,W}(m,P_n)\right]$$

$$\text{with}\quad \mathrm{pen}_{\epsilon,\lambda,W}(m,P_n) = \left(\widehat{p}_\lambda + \widehat{p}_\lambda^{W\epsilon}\right)\left(\widehat{\beta}_\lambda - \widehat{\beta}_\lambda^{W\epsilon}\right)^2 \ .$$

We can thus treat each term $\mathrm{pen}_{\epsilon,\lambda,W}(m,P_n)$ separately.

Conditionally to W, the value $\widehat{\beta}_\lambda^{W\epsilon}$ of $\widehat{s}_m(P_n^{W\epsilon})$ on I_λ is the mean (weighted by the ϵ^k) of the $\widehat{\beta}_\lambda^{W(k)}$ such that $P_n^{W(k)}(I_\lambda) > 0$. Thus, conditionally to the set $K_\lambda = \{k\in\mathbb{N} \text{ s.t. } P_n^{W(k)}(I_\lambda) > 0\}$, the restriction to I_λ of $P_n^{W\epsilon}$ and $Z_\epsilon^{-1}\sum_{k\in K_\lambda}\epsilon^k P_n^{W(k)}$ are equal (where $Z_\epsilon = \sum_{k\in K_\lambda}\epsilon^k$). When ϵ goes to 0, the first term is dominating the other ones, which are of order at most ϵ a.s. Since $\mathrm{pen}_{\epsilon,\lambda,W}(m,P_n)$ is a bounded continuous function of $P_n^{W\epsilon}$, we obtain that conditionally to K_λ,

$$\left(\widehat{p}_\lambda + \widehat{p}_\lambda^{W\epsilon}\right)\left(\widehat{\beta}_\lambda - \widehat{\beta}_\lambda^{W\epsilon}\right)^2 \rightharpoonup_{\epsilon\to 0} \left(\widehat{p}_\lambda + \widehat{p}_\lambda^{W(k_0)}\right)\left(\widehat{\beta}_\lambda - \widehat{\beta}_\lambda^{W(k_0)}\right)^2 \qquad \text{with } k_0 = \min K_\lambda.$$

Since the restriction of $P_n^{W(\min K_\lambda)}$ to I_λ has the same law as P_n^W conditionally to $P_n^W(X\in I_\lambda) > 0$, restricted to I_λ , we thus have

$$\lim_{\epsilon\to 0}\mathbb{E}_W\left[\mathrm{pen}_{\epsilon,\lambda,W}(m,P_n)\right] = \mathbb{E}_W\left[\mathrm{pen}_{\lambda,W}(m,P_n)\,\middle|\, P_n^W(X\in I_\lambda) > 0\right]$$

and the result follows. □

8.2. Resampling and structural constraints on the penalties

One of the first consequences of the resampling heuristics is that the Resampling Penalties have the same structural properties as the ideal penalty.

For instance, in the histogram case, with the notations of Sect. 5.7, we have

$$\mathrm{pen}_{\mathrm{id}}(m) := (P-P_n)\gamma\left(\widehat{s}_m\right) = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left[\frac{\left(\sum_{X_i\in I_\lambda}\left(Y_i-\beta_\lambda\right)\right)^2}{n\widehat{p}_\lambda}\left(1+\frac{p_\lambda}{\widehat{p}_\lambda}\right)\right]$$

according to (5.19) and (5.26) in Sect. 5.7.2. Then, $((X_i,Y_i)_{1\le i\le n},P)\mapsto \mathrm{pen}_{\mathrm{id}}(m)$ satisfies the following:

- It is the sum of $\mathrm{Card}(\Lambda_m)$ terms describing depending on what happens on I_λ:

$$\mathrm{pen}_{\mathrm{id}} = \sum_{\lambda\in\Lambda_m} F_\lambda\left(\widehat{p}_\lambda,(Y_i)_{X_i\in I_\lambda},\beta_\lambda,p_\lambda\right) \ . \tag{8.3}$$

- Each term is exchangeable: for every permutation σ of $\{i \text{ s.t. } X_i\in I_\lambda\}$,

$$F_\lambda\left(\widehat{p}_\lambda,(Y_i)_{X_i\in I_\lambda},\beta_\lambda,p_\lambda\right) \equiv F_\lambda\left(\widehat{p}_\lambda,(Y_{\sigma(i)})_{X_i\in I_\lambda},\beta_\lambda,p_\lambda\right) \ . \tag{8.4}$$

- Each term is translation invariant: for every $c\in\mathbb{R}$ and $\lambda\in\Lambda_m$,

$$F_\lambda\left(\widehat{p}_\lambda,(Y_i)_{X_i\in I_\lambda},\beta_\lambda,p_\lambda\right) \equiv F_\lambda\left(\widehat{p}_\lambda,(Y_i+c)_{X_i\in I_\lambda},\beta_\lambda+c,p_\lambda\right) \ . \tag{8.5}$$

- Each term is a polynomial in $(Y_i)_{X_i\in I_\lambda}$, homogeneous of order two:

$$F_\lambda\left(\widehat{p}_\lambda,(Y_i)_{X_i\in I_\lambda},\beta_\lambda,p_\lambda\right) = \sum_{X_i\in I_\lambda, X_j\in I_\lambda} a_{i,j}(\widehat{p}_\lambda,p_\lambda)Y_iY_j \ . \tag{8.6}$$

The last point becomes clear by remarking that one can translate Y so that $\beta_\lambda = 0$ according to the previous point.

Let us stress here on the fact that translation invariance and homogeneity of order 2 are necessary conditions for any good penalty in the least-square regression framework. Otherwise,

the selected model $\widehat{m}$ can be changed by replacing Y by $\lambda(Y+c)$, whereas the oracle model is unchanged, and the excess loss of every model is multiplied by λ^2. Then, making a particular choice for c and λ (*e.g.* depending on n), one can prove that $\widehat{m}$ is necessary can not satisfy a general non-asymptotic oracle inequality with constant almost one (and moreover, the constant may go to infinity when c and λ go to infinity). Condition (8.3) is only a consequence of the particular structure of the histogram models. We can expect that other models lead to other particular structures for reasonable penalties. Finally, the exchangeability condition (8.4) is the quite natural consequence of the exchangeability of the data.

We now come back to the histogram regression case. The resampling heuristics automatically ensures that $\mathrm{pen}(m)$ satisfies properties (8.3), (8.4), (8.5) and (8.6). According to the lemma below, this is a sufficient condition for $\mathrm{pen}(m)$ to be of the form

$$\sum_{\lambda\in\Lambda_m} a_\lambda(\widehat{p}_\lambda)\left(\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)^2-\frac{1}{n\widehat{p}_\lambda-1}\sum_{X_i\in I_\lambda,\,X_j\in I_\lambda,i\neq j}(Y_i-\beta_\lambda)(Y_j-\beta_\lambda)\right) .$$

As a consequence, up to the constants $R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)$, Resampling Penalties are the only reasonable penalties in the least-square histogram regression framework. We can conjecture that this structural preservation is a key property of RP in general frameworks, which may be sufficient to ensure its efficiency (up to the choice of the constant C in front of the penalty).

LEMMA 8.2. *Let $r\in\mathbb{N}\setminus\{0\}$ and $f:(Y_i)_{1\le i\le r}\mapsto\mathbb{R}$ satisfying*

(8.4) *f is exchangeable,* i.e. *for every permutation σ of $\{1,\dots,r\}$ and $Y\in\mathbb{R}$, $f\left((Y_{\sigma(i)})_{1\le i\le r}\right)=f(Y)$.*
(8.5) *for every $c\in\mathbb{R}$ and $Y\in\mathbb{R}^r$, $f\left((Y_i+c)_{1\le i\le r}\right)=f(Y)$.*
(8.6) *f is a polynomial and for every $\lambda\in\mathbb{R}$ and $Y\in\mathbb{R}^r$, $f\left((\lambda Y_i)_{1\le i\le r}\right)=\lambda^2 f(Y)$.*

Then, there exists $\alpha\in\mathbb{R}$ such that for every $Y\in\mathbb{R}^r$,

$$f(Y)=a\left(\sum_{i=1}^r Y_i^2-\frac{1}{r}\left(\sum_{1\le i\le r}Y_i\right)^2\right) .$$

PROOF OF LEMMA 8.2. According to (8.6), there exists $(a_{i,j})_{1\le i,j\le r}\in R^{r^2}$ such that for every $Y\in\mathbb{R}^r$,

$$f(Y)=\sum_{1\le i,j\le r}a_{i,j}Y_iY_j .$$

From (8.4), $a_{i,j}$ can only depend on $\mathbb{1}_{i=j}$, *i.e.* for every $Y\in\mathbb{R}^r$,

$$f(Y)=\alpha\sum_{i=1}^r Y_i^2+2\beta\sum_{1\le i<j\le r}Y_iY_j=(\alpha-\beta)\sum_{i=1}^r Y_i^2+\beta\left(\sum_{i=1}^r Y_i\right)^2$$

with $a_{1,1}=\alpha$ and $a_{1,2}=\beta$. Using now (8.5), we have for every $c\in\mathbb{R}$ and $Y\in\mathbb{R}^r$,

$$f(Y+c)=f(Y)+(\alpha+(r-1)\beta)\left(2c\sum_i Y_i+rc^2\right) .$$

We then must have $\alpha+(r-1)\beta=0$, and the result follows. □

8.3. Other assumption sets for oracle inequalities for RP

In this section, we consider alternative assumption sets for the results of Sect. 6.4 about Resampling Penalization. Since Thm. 6.1 relies on a general result (Lemma 6.4), giving alternative assumptions for Thm. 6.1 remains to give sufficient conditions for (**Bg**) or (**Ug**).

8.3.1. Bounded case. We already suggested one way of removing (**An**) in Sect. 6.4. It is actually possible to replace it in the assumptions of Thm. 6.1 by

(1) ($\mathbf{A}_{\text{gauss}}$) the noise is sub-gaussian
and ($\mathbf{Ar}_{\mathbf{u}}^{\mathbf{X}}$) the partition is "upper-regular" for $\mathcal{L}(X)$, *i.e.* $D_m \max_{\lambda\in\Lambda_m} p_\lambda \leq c_{\text{r,u}}^X$.

(2) $X \subset \mathbb{R}^k$, ($\mathbf{A}_{\text{gauss}}$) the noise is sub-gaussian,
($\mathbf{Ar}_{\mathbf{u}}$) the partition is "upper-regular" for Leb: $D_m \max_{\lambda\in\Lambda_m} \text{Leb}(I_\lambda) \leq c_{\text{r,u}} \text{Leb}(\mathcal{X})$
and the density of X w.r.t. Leb is bounded from above:

$$\forall I \subset \mathcal{X}, \quad \mathbb{P}(X \in I) \leq c_X^{\max} \frac{\text{Leb}(I)}{\text{Leb}(\mathcal{X})} \ . \tag{$\mathbf{Ad}_{\mathbf{u}}$}$$

Proof. Following the proof given in Sect. 6.8.3, we only have give a lower bound on

$$Q_m^{(p)} := \frac{1}{D_m} \sum_{\lambda\in\Lambda_m} \left[(\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2 \right] \ .$$

In the first case, we have

$$\begin{aligned} Q_m^{(p)} &\geq \frac{1}{D_m} \sum_{\lambda\in\Lambda_m} (\sigma_\lambda^r)^2 = \sum_{\lambda\in\Lambda_m} \frac{p_\lambda (\sigma_\lambda^r)^2}{D_m p_\lambda} \\ &\geq \sum_{\lambda\in\Lambda_m} \frac{p_\lambda (\sigma_\lambda^r)^2}{\max_{\lambda\in\Lambda_m} \{D_m p_\lambda\}} \geq \frac{\|\sigma(X)\|_2^2}{c_{\text{r,u}}^X} \ . \end{aligned}$$

The second case is a consequence of the first one since

$$\max_{\lambda\in\Lambda_m} \{p_\lambda\} \leq c_X^{\max} \max_{\lambda\in\Lambda_m} \left\{ \frac{\text{Leb}(I_\lambda)}{\text{Leb}(\mathcal{X})} \right\} \leq c_X^{\max} c_{\text{r,u}} D_m^{-1} \ .$$

□

Moreover, in all the assumption sets above and in those of Sect. 6.4, the sub-gaussian assumption on the noise ($\mathbf{A}_{\text{gauss}}$) can be replaced by a general moment inequality:

($\mathbf{A}\epsilon$) Pointwise moment inequality for the noise: there exists P^{pt} growing as some power of q such that

$$\forall q \geq 2, \forall x \in \mathcal{X}, \quad \mathbb{E}\left[|\epsilon|^q \mid X = x \right]^{1/q} \leq P^{pt}(q)\sigma(x) \ .$$

For instance, when $P^{pt}(q) \leq cq$ for every $q \geq 2$ for some constant c, this means that ϵ is *sub-poissonian*.

8.3.2. Unbounded case. In Sect. 6.4, we also give a set of assumptions for Thm. 6.1 in the unbounded case. One can actually remove both (**Ab**) and the lower bound on the noise (**An**) from the assumptions of Thm. 6.1, at the price of adding

($\mathbf{A}_{\text{gauss}}$) The noise is sub-gaussian: there exists $c_{\text{gauss}} > 0$ such that

$$\forall q \geq 2, \forall x \in \mathcal{X}, \quad \mathbb{E}\left[|\epsilon|^q \mid X = x \right]^{1/q} \leq c_{\text{gauss}} \sqrt{q} \sigma(x) \ .$$

($\mathbf{A}\delta$) Global moment assumption for the bias: there is a constant $c^g_{\Delta,m} > 0$ such that, for every $m \in \mathcal{M}_n$ of dimension $D_m \geq D_0$,

$$\|s - s_m\|_\infty \leq c^g_{\Delta,m} \|s(X) - s_m(X)\|_2$$

($\mathbf{A}\sigma_{\max}$) Noise-level bounded from above: $\sigma^2(X) \leq \sigma^2_{\max} < +\infty$ a.s.

($\mathbf{As}_{\max}$) Bound on the target function: $\|s\|_\infty \leq A$.

and one among the following

(1) ($\mathbf{Ar}^{\mathbf{X}}_{\mathbf{u}}$) the partition is "upper-regular" for $\mathcal{L}(X)$, *i.e.* $D_m \max_{\lambda\in\Lambda_m} p_\lambda \leq c^X_{\mathrm{r,u}}$.

(2) $X \subset \mathbb{R}^k$,
($\mathbf{Ar_u}$) the partition is "upper-regular" for Leb: $D_m \max_{\lambda\in\Lambda_m} \mathrm{Leb}(I_\lambda) \leq c_{\mathrm{r,u}} \mathrm{Leb}(\mathcal{X})$
and the density of X w.r.t. Leb is bounded from above:

$$\forall I \subset \mathcal{X}, \quad \mathbb{P}(X \in I) \leq c^{\max}_X \frac{\mathrm{Leb}(I)}{\mathrm{Leb}(\mathcal{X})} \quad . \qquad (\mathbf{Ad_u})$$

(3) $X \subset \mathbb{R}^k$ is bounded, equipped with $\|\cdot\|_\infty$,
($\mathbf{Ar}^{\mathbf{d}}_{\mathbf{u}}$) the partition is "upper-regular": $\max_{\lambda\in\Lambda_m} \{\mathrm{diam}(I_\lambda)\} \leq c^d_{\mathrm{r,u}} D_m^{-\alpha_d} \mathrm{diam}(X)$
($\mathbf{Ar_u}$) the partition is "upper-regular" for Leb: $\max_{\lambda\in\Lambda_m} \{\mathrm{Leb}(I_\lambda)\} \leq c_{\mathrm{r,u}} D_m^{-1} \mathrm{Leb}(X)$
and ($\mathbf{A}\sigma$) σ is piecewise K_σ-Lipschitz with at most J_σ jumps.

PROOF. In Sect. 6.8.3, we made this proof with ($\mathbf{An}$) as last additional assumption. It is actually only used to give a lower bound on $Q^{(p)}_m$. The two first cases thus follows from the proof given in Sect. 8.3.1 above. The last one follows from Lemma 6.13. □

As in the bounded case, the sub-gaussian assumption on the noise ($\mathbf{A}_{\mathrm{gauss}}$) can be replaced everywhere by the more general moment assumption ($\mathbf{A}\epsilon$).

Sufficient conditions for ($\mathbf{A}\delta$) can be derived either from Lemma 6.14 or from Lemma 8.3 below.

8.3.3. Sufficient condition for ($\mathbf{A}\delta$). In Sect. 6.4, we give a sufficient condition for ($\mathbf{A}\delta$)that relies on the regularity of s, a lower bound on the density of X w.r.t. Leb and the regularity of the partition (Lemma 6.14). We state below a lemma which gives a more accurate estimation of the constant $c^g_{\Delta,m}$ when $\mathcal{X} \subset \mathbb{R}$ is bounded.

LEMMA 8.3. *Let s be a C_{Lip}-Lipschitz fuction on $\mathcal{X} \subset \mathbb{R}$ and μ a probability measure on $\mathcal{X}$. We assume that μ and* Leb *are mutually absolutely continuous. Let $\left((I_\lambda)_{\lambda\in\Lambda_{m_k}}\right)_{k\in\mathbb{N}}$ be a sequence of partitions of $\mathcal{X}$. We assume that their sizes D_{m_k} are going to infinity and*

$$\max_{\lambda\in\Lambda_m} \{\mathrm{diam}\,(I_\lambda)\} \leq c^d_{\mathrm{r,u}} D_m^{-1} \mathrm{diam}(\mathcal{X}) \quad . \qquad (\mathbf{Ar}^{\mathbf{d}}_{\mathbf{u}})$$

Then, there exists a constant $c^g_{\Delta,m}$ (depending on s, μ and $c^d_{\mathrm{r,u}}$) such that for every $k \in \mathbb{N}$,

$$\|s - s_m\|_\infty \leq c^g_{\Delta,m} \|s - s_m\|_{L^2(\mu)} \quad .$$

PROOF. If s is constant, the result is obvious. Otherwise, both $\|s - s_m\|_\infty$ and $\|s - s_m\|_2$ are positive.

Since s is Lipschitz, with constant C_{Lip}, we have

$$\|s - s_m\|_\infty \leq C_{Lip} \max_{\lambda\in\Lambda_m} \{\mathrm{diam}\,(I_\lambda)\} \leq D_m^{-1} \mathrm{diam}(\mathcal{X}) c^d_{\mathrm{r,u}} C_{Lip} \quad .$$

On the other hand, $\|s - s_{m_k}\|^2_{L^2(\mu)}$ is equivalent to $\frac{\|s'\|^2_{L^2(\mu)}}{12 D^2_{m_k}}$ as long as the Riemann sums of s' are converging. The result follows. □

REMARK 8.1. If one assumes more regularity conditions on s, then the difference between $\|s - s_{m_k}\|^2_{L^2(\mu)}$ and its limit when $k \to \infty$ can be controlled. Then, the constant $c^g_{\Delta,m}$ only depends on s through $\|s'\|^2_{L^2(\mu)}$ and these regularity conditions, at least for $k \geq k_0$ for some k_0 depending on the same conditions.

8.4. Resampling Penalties with general weights

In Chap. 6, we focus on RP with exchangeable weights, whereas Chap. 5 studies a particular kind of non-exchangeable weights. In this section, we state results that are valid for any resampling weight vector. As in the two aforementioned chapters, we can use them to derive oracle inequalities for RP with general weights, under some mild conditions on the weights. See the proofs of Thm. 5.1 and 6.1 for more details.

8.4.1. Expectations. When the resampling weights are exchangeable, we are able to compute explicitly the Resampling Penalties pen(m) in the histogram case (Lemma 5.7). With Prop. 5.2, we extended this result in expectation to general weights. However, this last result seems very particular to histograms, and may not always be easy to handle. With Lemma 8.4 below, we state it in a more general form, which will be sometimes easier to use.

General result.

LEMMA 8.4. *Let W be a resampling weight vector, $(X_i, Y_i)_{1\leq i\leq n}$ some exchangeable data,*

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)} \qquad \text{and} \qquad P_n^W = \frac{1}{n}\sum_{i=1}^n W_i \delta_{(X_i,Y_i)} \ .$$

We define

$$F_{\mathcal{L}(W)}((X_i,Y_i)_{1\leq i\leq n}) = \mathbb{E}\left[G(P_n, P_n^W) \,\middle|\, (X_i,Y_i)_{1\leq i\leq n}\right]$$

for some measurable function G, and

$$W^\sigma = \left(W_{\sigma(1)}, \dots, W_{\sigma(n)}\right) \qquad \text{with } \sigma \text{ uniform in } \Sigma_n = \Sigma\left(\{1,\dots,n\}\right)$$

and independent from W and $(X_i,Y_i)_{1\leq i\leq n}$.

Then, $(W_i^\sigma)_{1\leq i\leq n}$ is exchangeable and

$$\mathbb{E}\left[F_{\mathcal{L}(W)}(P_n)\right] = \mathbb{E}\left[F_{\mathcal{L}(W^\sigma)}(P_n)\right] \ . \tag{8.7}$$

REMARK 8.2. In the histogram case, (8.7) also holds true conditionally to $(\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$.

PROOF OF LEMMA 8.4. Noticing that the data is exchangeable, we have

$$\begin{aligned}\mathbb{E}\left[F_{\mathcal{L}(W)}\left((X_i,Y_i)_{1\leq i\leq n}\right)\right] &= \frac{1}{n!}\sum_{\tau\in\Sigma_n}\mathbb{E}\left[F_{\mathcal{L}(W)}\left((X_{\tau(i)},Y_{\tau(i)})_{1\leq i\leq n}\right)\right] \\ &= \frac{1}{n!}\sum_{\tau\in\Sigma_n}\mathbb{E}\left[G(P_n, P_n^{W,\tau})\right]\end{aligned}$$

$$\text{with} \quad P_n^{W,\tau} = \frac{1}{n}\sum_{i=1}^n W_i \delta_{(X_{\tau(i)},Y_{\tau(i)})} = \frac{1}{n}\sum_{j=1}^n W_{\tau^{-1}(j)}\delta_{(X_j,Y_j)} = P_n^{W o \tau^{-1}}$$

and $W o\tau^{-1} = (W_{\tau^{-1}(j)})_{1\leq j\leq n}$. By definition of $W^\sigma \overset{(d)}{=} W o\tau^{-1}$ (if τ has a uniform law in Σ_n), the result follows. □

V-fold case. In algorithm 5.2 in Sect. 5.3.1, we modify the general V-fold weights so that they fit well to the partitions $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$. It is also possible to fix *a priori* the blocks $(B_j)_{1\leq j\leq V}$ (*e.g.* approximatively of the same sizes), and then compute the Resampling Penalties as in (6.5).

According to Lemma 8.4 above, we have to determine the law of the exchangeable weight W^σ. If the blocks all have the same size, W^σ is the Rho $(n-n/V)$ scheme. If the blocks have different sizes, the law of W^σ is some kind of generalized Rho scheme, with a sample size that is not fixed. Its law can be described as follows:

(1) Pick up a sample size Q at random according to the distribution
$$V^{-1}\sum_{j=1}^{V}\delta_{n-\mathrm{Card}(B_j)}\ .$$
(2) Independently from Q, choose a subset I of $\{1,\dots,n\}$ uniformly among those of size Q. The weights are then equal to $W_i=\frac{V}{V-1}\mathbb{1}_{i\in I}$.

The constant $V/(V-1)$ apart, one can see this resampling scheme as a generalized Rho, with a random sample size.

We obtain the constants R_{1,W^σ} and R_{2,W^σ} in the general VFCV case:
$$\begin{aligned}R_{1,W^\sigma}(n,\widehat{p}_\lambda)&=\frac{1}{V}\sum_{j=1}^{V}\frac{V(n-\mathrm{Card}(B_j))}{(V-1)n}\left(\frac{n}{n-\mathrm{Card}(B_j)}e^{+}_{\mathcal{H}(n,n\widehat{p}_\lambda,n-\mathrm{Card}(B_j))}-1\right)\\&=\left(\frac{1}{V-1}\sum_{j=1}^{V}e^{+}_{\mathcal{H}(n,n\widehat{p}_\lambda,n-\mathrm{Card}(B_j))}\right)-1 \qquad (8.8)\\ R_{2,W^\sigma}(n,\widehat{p}_\lambda)&=\frac{1}{V}\sum_{j=1}^{V}\left(\frac{n}{n-\mathrm{Card}(B_j)}-1\right) \qquad (8.9)\end{aligned}$$

Indeed, conditionally to Q, the quantity $(W_i-W_\lambda)^2W_\lambda^{-\alpha}$ (for $\alpha\in\{1,2\}$) has already been computed in the Rho(Q) case (up to the multiplicative constant $\frac{VQ}{(V-1)n}$ when $\alpha=1$). Then, we only have to integrate w.r.t. Q.

We now assume that $n^{-1}\max_j\mathrm{Card}(B_j)\leq c_B<1$ and $np_\lambda\geq B_n$ and that the partition is quasi-regular:
$$\sup_j\left\{\left|\frac{\mathrm{Card}(B_j)}{n}-\frac{1}{V}\right|\right\}\leq\epsilon_n^{reg}\xrightarrow[n\to\infty]{}0\ . \qquad (\mathbf{A_{reg,VF}})$$

Then,
$$R_{1,W^\sigma}(n,\widehat{p}_\lambda)=R_{2,W^\sigma}(n,\widehat{p}_\lambda)\left(1+\delta^{(penV)}_{n,\widehat{p}_\lambda}\right)\quad\text{with}\quad\delta^{(penV)}_{n,\widehat{p}_\lambda}\xrightarrow[n\widehat{p}_\lambda\to\infty]{}0\ .$$

Explicit bounds can be derived from Lemma 6.2.

8.4.2. Concentration inequality. We have proved concentration inequalities in the histogram case for Resampling Penalties when the weights are exchangeable (Prop. 6.8 in Sect. 6.8.7) and for V-fold weights (Prop. 5.10 in Sect. 5.7.4). In this section, we extend this last result to more general non-exchangeable weights.

PROPOSITION 8.5. *Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Let* pen($m$) *be defined by*
$$\mathrm{pen}(m)=C\left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}^W_\lambda-\widehat{\beta}_\lambda\right)^2\,\middle|\,W_\lambda>0\right]+\mathbb{E}_W\left[\widehat{p}^W_\lambda\left(\widehat{\beta}^W_\lambda-\widehat{\beta}_\lambda\right)^2\right]\right)\ . \qquad (8.10)$$

and $W\in[0;\infty)$ a random vector such that, conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$:

- *W is independent from $(X_i, Y_i)_{1\leq i\leq n}$.*
- *$\max_i W_i - \min_i W_i \leq M_W < \infty$ a.s.*
- *$\min_{\lambda\in\Lambda_m} W_\lambda \geq m_W > 0$ a.s.*

The constants M_W and m_W may depend on $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$. Let $\gamma > 0$ and assume that

$$\forall q \geq 2, \qquad P_m^\ell(q) \leq a_\ell q^{\xi_\ell} \ . \tag{$\mathbf{A_{m},\ell}$}$$

Then, there exists an event of probability at least $1 - n^{-\gamma}$ such that

$$\begin{aligned} &\left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \mathbb{1}_{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq 2} \\ &\leq C M_W^2 \left(m_W^{-1} + m_W^{-2}\right) L(a_\ell, \xi_\ell, \gamma) D_m^{-1/2} \ln(n)^{\xi_\ell+1} \mathbb{E}\left[p_2(m)\right] \ . \end{aligned} \tag{8.11}$$

PROOF OF PROP. 8.5. In the proof of Prop. 5.10 (*cf.* Sect. 5.7.4), we have shown that for general weights W, for every $q \geq 2$,

$$\begin{aligned} &\left\|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right\|_q^{(\Lambda_m)} \\ &\leq \sup_W \left\{ Lq \sqrt{\sum_{\lambda\in\Lambda_m} \left(\frac{1+W_\lambda}{n^2\widehat{p}_\lambda W_\lambda^2}\right)^2 m_{2q,\lambda}^4 \left(\sum_{i=1}^{r_\lambda} \left(W_{(i,\lambda)} - W_\lambda\right)^2\right)^2} \right\} \\ &\leq \sup_W \left\{ \frac{Lq}{n} \left(\max_i W_i - \min_i W_i\right)^2 \left(\frac{1}{\min_{\lambda\in\Lambda_m}\left\{W_\lambda\right\}} + \frac{1}{\min_{\lambda\in\Lambda_m}\left\{W_\lambda^2\right\}}\right) \sqrt{\sum_{\lambda\in\Lambda_m} m_{2q,\lambda}^4} \right\} \end{aligned}$$

where the supremum holds over the support of W.

The result follows with $(\mathbf{A_{m},\ell})$ and the classical link between moment and concentration inequalities (Lemma 8.10 in Sect. 8.6.2). □

8.5. Useful concentration inequalities

We recall in this section some basic concentration inequalities that we often use in this thesis. Complete proofs (and many other concentration inequalities) are to be found in Massart's Saint-Flour lecture notes [**Mas07**].

PROPOSITION 8.6 (Hoeffding's inequality ([**Mas07**], Lemma 2.6 and Prop. 2.7)). *If $Y \in [a; b]$ a.s., then for every $\mu \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left(\mu(Y - \mathbb{E}[Y])\right)\right] \leq \exp\left(\frac{\mu^2(b-a)^2}{8}\right) \ .$$

Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes values in $[a_i, b_i]$ almost surely for all $i \leq n$. Then for any positive x, we have

$$\mathbb{P}\left(\sum_{i=1}^n \left(X_i - \mathbb{E}\left[X_i\right]\right)\right) \leq \exp\left(-\frac{x^2}{2\sum_{i=1}^n (b_i - a_i)^2}\right) \ .$$

PROPOSITION 8.7 (McDiarmid's inequality ([**Mas07**], Thm. 5.1)). *Let $X_1 \in \mathcal{X}_1, \ldots, X_n \in \mathcal{X}_n$ be some independent random variables, $\mathcal{X}^n = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and $\zeta\mathcal{X}^n \mapsto \mathbb{R}$ be some measurable functional satisfying for some positive constants $(c_i)_{1\leq i\leq n}$, the bounded difference condition*

$$\forall x \in \mathcal{X}^n, \forall y \in \mathcal{X}^n, \forall i \in \{1, \ldots, n\}, \qquad |\zeta(x_1, \ldots, x_i, \ldots, x_n) - \zeta(x_1, \ldots, x_i, \ldots, x_n)| \leq c_i \ .$$

Then, the random variable $Z = \zeta(X_1, \ldots, X_n)$ satisfies, for every $x \geq 0$,

$$\mathbb{P}\left(Z - \mathbb{E}\left[Z\right] \geq x\right) \leq \exp\left(-\frac{2x^2}{\sum_{i=1}^n c_i^2}\right)$$

and similarly

$$\mathbb{P}\left(\mathbb{E}\left[Z\right]-Z\geq x\right)\leq\exp\left(-\frac{2x^2}{\sum_{i=1}^n c_i^2}\right)\ .$$

This result is due to McDiarmid [**McD89**]. It is also known as the "bounded difference inequality".

PROPOSITION 8.8 (Bernstein's inequality ([**Mas07**], Prop. 2.9)). *Let $X_1,\ldots,X_n$ be independent random variables. Assume that there exists some positive numbers v and c such that $X_i\leq 3c$ almost surely and $\sum_{i=1}^n\mathbb{E}\left[X_i^2\right]\leq v$. Then for every positive x,*

$$\mathbb{P}\left(\sum_{i=1}^n\left(X_i-\mathbb{E}\left[X_i\right]\right)\geq\sqrt{2vx}+cx\right)\leq\exp(-x)\ .$$

For instance, if $X\sim\mathcal{B}(n,p)$, Prop. 8.8 with $c=\frac{1}{3}$ and $v=np(1-p)$ gives that for every positive x,

$$\mathbb{P}\left(X\geq\mathbb{E}(X)+\sqrt{2np(1-p)x}+\frac{x}{3}\right)\leq e^{-x}$$
$$\mathbb{P}\left(X\leq\mathbb{E}(X)-\sqrt{2np(1-p)x}-\frac{x}{3}\right)\leq e^{-x}\ .$$

THEOREM 8.1 (Gaussian concentration theorem ([**Mas07**], Thm. 3.4)). *Let $X_1,\ldots,X_n$ be independent random variables with distribution $\mathcal{N}(0,1)$, and $\zeta:\mathbb{R}^n\mapsto\mathbb{R}$ be some Lipschitz function ($\mathbb{R}^n$ being equipped with the canonical Euclidean norm) with Lipschitz constant $L\geq 0$. Then, the random variable $Z=\zeta(X_1,\ldots,X_n)$ satisfies, for every $x\geq 0$,*

$$\mathbb{P}\left(|Z-M|\geq x\right)\leq 2\exp\left(-\frac{x^2}{2L^2}\right)$$
$$\text{and}\qquad\mathbb{P}\left(Z\geq M+x\right)\leq\exp\left(-\frac{x^2}{2L^2}\right)$$

where M denotes either the mean or the median of Z.

When M is the mean, this result is due to Cirel'son, Ibragimov and Sudakov [**CIS76**] (and the upper bound on the probabilities can be replaced by $4\overline{\Phi}(x/L)$, resp. $4\overline{\Phi}(x/L)$, where $\overline{\Phi}$ is the standard Gaussian upper tail function). See Massart [**Mas07**], Sect. 3.2 for further references and results.

8.6. Moments, Exponential moments and Concentration

In this section, we recall several classical results giving links between moments, exponential moments and concentration inequalities.

8.6.1. Exponential moments ⇒ concentration. This link relies on Markov inequality.

LEMMA 8.9. *Let Z be a random variable such that*

$$\forall\mu>0,\mathbb{E}\left[e^{\mu Z}\right]\leq\exp\left(\frac{a}{4}\mu^2+b\mu+c\right),$$

then, for all $x\in\mathbb{R}$,

$$\mathbb{P}(Z\geq x)\leq\inf_{\mu>0}\left\{\exp\left(\frac{a}{4}\mu^2+(b-x)\mu+c\right)\right\}\leq\exp\left(c-\frac{(b-x)_-^2}{a}\right).$$

PROOF. We apply Markov inequality to $e^{\mu Z}$. If $b-x\geq 0$, the infimum is achieved with $\mu=0$. Otherwise, it is achieved for $\mu=\frac{2(x-b)}{a}$. □

8.6.2. Moments ⇒ concentration. This link relies on the same idea, with $x \mapsto x^q$ instead of $x \mapsto \exp(\lambda x)$.

LEMMA 8.10. *Let $\lambda_1, \dots, \lambda_N \geq 0$, $\mu_1, \dots, \mu_N > 0$ and X be a random variable such that for all $q \geq q_0 > 0$,*

$$\|X\|_q \leq \sum_{i=1}^{N} \lambda_i q^{\mu_i} \ .$$

Then for every $y \geq 0$,

$$\mathbb{P}\left(|X| \geq \sum_{i=1}^{N} \left[\lambda_i \left(\frac{ey}{\min_j \mu_j}\right)^{\mu_i}\right]\right) \leq e^{q_0 \min_j\{\mu_j\}} e^{-y} \ .$$

PROOF. First notice that for every $t \geq 0$ and $q \geq q_0$,

$$\mathbb{P}\left(|X| \geq t\right) \leq \mathbb{P}\left(|X|^q \geq t^q\right) \leq \frac{\|X\|_q^q}{t^q} \leq \left(\frac{\sum_{i=1}^{N} \lambda_i q^{\mu_i}}{t}\right)^q \ .$$

We take

$$t = \min_i \left\{N\lambda_i \left(\frac{ye}{\min_j \mu_j}\right)^{\mu_i}\right\} \leq \sum_{i=1}^{N} \left[\lambda_i \left(\frac{ye}{\min_j \mu_j}\right)^{\mu_i}\right] \ ,$$

and if

$$q = \widetilde{q}(t) = e^{-1} \min_{1 \leq i \leq N} \left\{\left(\frac{t}{N\lambda_i}\right)^{1/\mu_i}\right\}$$

is larger than q_0, the result follows. Otherwise, $e^{q_0 \min_j\{\mu_j\}} e^{-y} \geq 1$ so that the result is obvious. □

REMARK 8.3. If there is a constant term in the bound on the moments, one can upper bound λ_0 by $\lambda_0 q_0^{-\mu_1} q^{\mu_1}$ for instance. The same trick can be used if $\min_j \mu_j$ is too small.

8.6.3. Moments *vs.* Exponential moments. In general, moment inequalities give better concentration results than exponential moments. The first reason for this is simply that a random variable may have q-th moments for $1 \leq q \leq q_1 \leq \infty$ but not exponential moments. As remarked by Chatterjee [**Cha04**] (Sect. 3.8), there is also a second reason: if one uses an optimized Markov inequality to derive concentration bounds, moments always give a better result than exponential moments. This is shown by Lemma 8.11 below, with $Z = (X - \mathbb{E}[X])_+$ or $Z = (X - \mathbb{E}[X])_-$.

LEMMA 8.11. *Let Z be a non-negative random variable such that $\mathbb{E}\left[\exp(\lambda_0 Z)\right] < \infty$ for some $\lambda_0 > 0$. Then, for every $x \geq 0$,*

$$\inf_{q \in \mathbb{N}} \frac{\mathbb{E}\left(Z^q\right)}{x^q} \leq \inf_{0 \leq \lambda \leq \lambda_0} \frac{\mathbb{E}\left(e^{\lambda Z}\right)}{e^{\lambda x}} \ .$$

PROOF. Let $\lambda \in [0; \lambda_0]$. Then,

$$\mathbb{E}\left(e^{\lambda Z}\right) = \sum_{q=0}^{+\infty} \frac{\lambda^q}{q!} \mathbb{E}(Z^q) \geq \sum_{q=0}^{+\infty} \frac{(\lambda x)^q}{q!} \inf_{q \in \mathbb{N}} \frac{\mathbb{E}\left(Z^q\right)}{x^q} = e^{\lambda x} \inf_{q \in \mathbb{N}} \frac{\mathbb{E}\left(Z^q\right)}{x^q} \ .$$

□

8.6.4. Concentration ⇒ moments. It is sometimes useful to go back to a moment inequality from a concentration result. Basically, this is an integration.

PROPOSITION 8.12. *Let Y be a random variable such that*

$$\mathbb{P}\left(Y \geq ay\right) \leq b \exp\left(-y^\alpha\right)$$

with $a > 0$, $b > 0$, $\alpha > 0$. Then, for every $q \geq \alpha \vee 1$,

$$\|Y\|_q \leq ae \left[1 \vee \left(\frac{eb}{\alpha^{3/2}}\right)\right] \left(\frac{q}{\alpha e}\right)^{1/\alpha} . \tag{8.12}$$

PROOF. This relies on the following integration by parts formula:

$$\mathbb{E}(|X|) = \int_0^\infty \mathbb{P}(|X| \geq x)dx.$$

We apply it to $|Y/a|^q$ and deduce (up to a change of variable)

$$\mathbb{E}(|Y|^q) \leq a^q \int_0^\infty b \exp\left(-x^{\alpha/q}\right) dx = a^q \frac{bq}{\alpha} \int_0^\infty t^{q/\alpha - 1} e^{-t} dt = a^q \frac{bq}{\alpha} \Gamma\left(\frac{q}{\alpha}\right) .$$

The result follows by Lemma 8.13 (with $\beta = q\alpha^{-1}$) and the following inequalities: for every $q \geq 1$,

$$q^{3/(2q)} \leq e^{3/(2e)} \leq e \qquad \text{and} \qquad \left(\frac{eb}{\alpha^{3/2}}\right)^{1/q} \leq 1 \vee \left(\frac{eb}{\alpha^{3/2}}\right) .$$

□

LEMMA 8.13. *For every $\beta \geq 1$,*

$$\Gamma(\beta) := \int_0^\infty t^{\beta-1} e^{-t} dt \leq e \left(\frac{\beta}{e}\right)^\beta \sqrt{\beta} .$$

PROOF. It's true for $\beta \in [1; 2]$ since the right-hand side is non-decreasing and $\Gamma \leq 1$ on this interval. We deduce the result by induction since the right-hand side grows faster than the left-hand side:

$$\frac{\Gamma(\beta+1)}{\Gamma(\beta)} = \beta \leq e^{-1} \left(\frac{\beta+1}{\beta}\right)^{\beta+3/2} \beta .$$

Indeed, for every $\beta \geq 1$,

$$\left(\frac{\beta+1}{\beta}\right)^{\beta+3/2} \geq e \qquad \text{since} \qquad g : x \mapsto \left(x + \frac{3}{2}\right) \ln\left(1 + \frac{1}{x}\right)$$

decreases on $[1; +\infty)$ and goes to 1 at infinity. □

8.7. Expectation of inverses: symmetric case

When $p = 1/2$, we can improve the result of Lemma 5.3 in Sect. 5.6.1. This relies on the following general argument.

Let X be a nonnegative symmetric random variable, *i.e.* such that $\mathcal{L}(X) = \mathcal{L}(2 - X)$. Recall the definition

$$e^+_{\mathcal{L}(X)} := e^+_X := \mathbb{E}\left[\frac{1}{X} \,\middle|\, X > 0\right] .$$

Define $p_0 = \mathbb{P}(X = 0) = \mathbb{P}(X = 2)$. Then,

$$\begin{aligned} e^+_X &= \frac{\mathbb{P}(X = 2 \,|\, X > 0)}{2} + \mathbb{E}\left[\frac{1}{X} \,\middle|\, 0 < X < 2\right] \frac{\mathbb{P}(0 < X < 2)}{\mathbb{P}(X > 0)} \\ &= \frac{p_0}{2(1-p_0)} + \frac{1-2p_0}{1-p_0} \mathbb{E}\left[\frac{1}{2}\left(\frac{1}{X} + \frac{1}{2-X}\right) \,\middle|\, 0 < X < 2\right] \\ &= \frac{p_0}{2(1-p_0)} + \frac{1-2p_0}{1-p_0} \left(1 + \mathbb{E}\left[\frac{(X-1)^2}{X(2-X)} \,\middle|\, 0 < X < 2\right]\right) . \end{aligned} \tag{8.13}$$

LEMMA 8.14. *Let $n \in \mathbb{N} \setminus \{0\}$. Then,*

$$2 + 3 \times 10^{-4} \geq e^+_{\mathcal{B}(n, \frac{1}{2})} \geq \mathbb{1}_{n \geq 3} . \tag{8.14}$$

PROOF. Let $Z \sim \mathcal{B}\left(n, \frac{1}{2}\right)$. Since $X = 2n^{-1}Z$ is nonnegative and symmetric, we can apply the above result with $p_0 = \mathbb{P}(Z = 0) = 2^{-n}$.

Lower bound. We here assume that $n \geq 3$. For every $x \in (0; 2)$, define

$$T(x) = \frac{(x-1)^2}{x(2-x)} \ . \qquad \text{In particular,} \qquad T\left(\frac{2}{n}\right) = \frac{(n-2)^2}{4(n-1)}$$

so that

$$\begin{aligned}(1-2p_0)\mathbb{E}\left[T(X) \,|\, 0 < X < 2\right] &\geq \mathbb{P}\left(X = \frac{2}{n} \text{ or } X = 2 - \frac{2}{n}\right) T\left(\frac{2}{n}\right) \\ &\geq \frac{2n}{2^n}\frac{(n-2)^2}{4(n-1)} = \frac{n(n-2)^2}{2^{n+1}(n-1)}\end{aligned}$$

since $2/n \neq 2 - 2/n$ for $n \geq 3$.

Putting this inequality into (8.13) we obtain:

$$\begin{aligned}e^+_{\mathcal{B}\left(n,\frac{1}{2}\right)} &\geq \frac{1}{1-2^{-n}}\left(2^{-n-1} + 1 - 2^{1-n} + \frac{n(n-2)^2}{2^{n+1}(n-1)}\right) \\ &\geq \frac{1}{1-2^{-n}}\left(2^{-n-1} + 1 - 2^{1-n} + 2^{-n-1}\right) = 1 \ .\end{aligned}$$

Small values of n. We easily compute the following values:

n	1	2	3	4	5	6	7	8
$e^+_{\mathcal{B}\left(n,\frac{1}{2}\right)}$	$\frac{1}{2}$	$\frac{5}{6}$	$\frac{29}{28}$	$\frac{103}{90}$	$\frac{887}{744}$	$\frac{1517}{1260}$	$\frac{18239}{15240}$	$\leq 1{,}18$

Then, for $n = 1, 2$, $e^+_{\mathcal{B}\left(n,\frac{1}{2}\right)} \leq 1$ and for $3 \leq n \leq 8$,

$$1 \leq e^+_{\mathcal{B}\left(n,\frac{1}{2}\right)} \leq e^+_{\mathcal{B}\left(6,\frac{1}{2}\right)} \leq 1{,}21 \ .$$

Upper bound. The computations above give the upper bound for $n \leq 8$. For $n \geq 9$, we use (6.19) and (5.60). □

8.8. Concentration of inverses of multinomials: proofs

In this section, we prove Lemma 5.4 which is stated in Sect. 5.6.2. We will make use of the following constant:

$$h_+ = \frac{3\left(\sqrt{5} - \sqrt{3}\right)^2}{2} \approx 0.38 \ .$$

PROOF OF LEMMA 5.4. As the coordinates of X are not independent, we cannot apply here classical tools of concentration in product-spaces. The point here is that $(X_\lambda)_{\lambda \in \Lambda_m}$ are negatively associated [**JDP83**], so decreasing functions of the X_λ as $(a_\lambda f_1(X_\lambda))_{\lambda \in \Lambda_m}$ are still negatively associated. In such situations, one can still apply Cramér-Chernoff method in order to obtain concentration for $Z_{m,T}$ [**DR98**].

Lower deviations. For all $\mu \geq 0$,

$$\begin{aligned}\mathbb{E}\left[e^{\mu(\mathbb{E}[Z_{m,1}]-Z_{m,1})}\right] &= \mathbb{E}\left[\prod_{\lambda\in\Lambda_m} \exp\left(\mu a_\lambda \left(\mathbb{E}[f_1(X_\lambda)] - f_1(X_\lambda)\right)\right)\right] \\ &\leq \prod_{\lambda\in\Lambda_m} \mathbb{E}\left[\exp\left(\mu a_\lambda \left(\mathbb{E}[f_1(X_\lambda)] - f_1(X_\lambda)\right)\right)\right]\end{aligned}$$

because of the negative association property.

These exponential moments are bounded by Lemma 8.15 since $X_\lambda \sim \mathcal{B}(n, p_\lambda)$. Thus, for all $\mu \geq 0$,

$$\begin{aligned}\mathbb{E}\left[e^{\mu(\mathbb{E}[Z_m]-Z_m)}\right] &\leq \prod_{\lambda\in\Lambda_m}\left[\exp\left(\frac{c_1' a_\lambda^2 \mu^2}{(np_\lambda)^2} + \frac{a_\lambda \mu}{c_1 \times (np_\lambda)}\varphi(c_1 \times np_\lambda)\right) + \exp(-c_1 \times np_\lambda)\right] \\ &\leq \prod_{\lambda\in\Lambda_m}\left[\exp\left(\frac{c_1' a_\lambda^2 \mu^2}{(np_\lambda)^2} + \frac{a_\lambda \mu}{c_1 \times (np_\lambda)}\varphi(c_1 \times np_\lambda)\right)(1 + \exp(-c_1 \times np_\lambda))\right] \\ &= \exp\left(\frac{A_m}{4}\mu^2 + B_m\mu + C_m\right)\end{aligned}$$

with

$$\begin{aligned}A_m &= 4c_1' \sum_{\lambda\in\Lambda_m} \frac{a_\lambda^2}{(np_\lambda)^2} \\ B_m &= \frac{1}{c_1}\sum_{\lambda\in\Lambda_m} \frac{a_\lambda \varphi(c_1 \times np_\lambda)}{np_\lambda} \\ C_m &= \sum_{\lambda\in\Lambda_m} \ln\left(1 + \exp(-c_1 \times np_\lambda)\right)\end{aligned}$$

The result follows from obvious upper bounds on B_m and C_m and Lemma 8.9.

Upper deviations. We use the negative assocation of the $f_T(X_\lambda)$ and bounds on exponential moments given by Lemma 8.16 with $a_\lambda\mu$ instead of μ. Thus, as soon as

$$0 \leq \mu \leq \min_{\lambda\in\Lambda_m}\left\{\frac{(1-\eta)h_+ np_\lambda T^{-1}}{a_\lambda}\right\} \ ,$$

$$\begin{aligned}\mathbb{E}\left[e^{\mu(Z_{m,T}-\mathbb{E}[Z_{m,T}])}\right] &\leq \exp\left[\sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda^2\mu^2\beta_\eta^2\zeta_0^2}{2(np_\lambda)^2} + \frac{a_\lambda\mu}{np_\lambda\gamma_\eta}\varphi\left(np_\lambda\gamma_\eta\right) + \exp(-np_\lambda\delta_\eta)\right)\right] \\ &= \exp\left(\frac{A_m'\mu^2}{4} + B_m'\mu + C_m'\right) \qquad (8.15)\end{aligned}$$

with

$$\begin{aligned}A_m' &= 2\beta_\eta^2\zeta_0^2 \sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda}\right)^2 \\ B_m' &= \sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda\gamma_\eta}\varphi\left(np_\lambda\gamma_\eta\right)\right) \leq \frac{\varphi\left(B_n\gamma_\eta\right)}{\gamma_\eta}\sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda}\right) \\ C_m' &= \sum_{\lambda\in\Lambda_m}\exp(-np_\lambda\delta_\eta) \leq D_m\exp(-B_n\delta_\eta) \ .\end{aligned}$$

We can then derive a deviation inequality for $Z_{m,T}$ if and only if

$$\frac{2(y-B'_m)_+}{A'_m} \leq \min_{\lambda\in\Lambda_m}\left\{\frac{(1-\eta)h_+np_\lambda T^{-1}}{a_\lambda}\right\} \quad \text{with} \quad y = B'_m + \sqrt{A'_m(C'_m+x)}$$

i.e. if

$$x \leq A'_m\left(\frac{(1-\eta)h_+}{2}\right)^2 (nT^{-1})^2\left(\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{a_\lambda}\right\}\right)^2 - C'_m \ .$$

This condition always holds when

$$\zeta_0^2 = 1 \vee \sqrt{\frac{(x+C'_m)\,4T^2}{2\beta_\eta^2\left(\sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda}\right)^2\right)(1-\eta)^2h_+^2n^2\min_{\lambda\in\Lambda_m}\left\{\left(\frac{p_\lambda}{a_\lambda}\right)^2\right\}}} \ .$$

Finally, we take $\eta = \frac{1}{2}$ so that

$$c_2 \leq \gamma_\eta \quad c_3 \geq \sqrt{2}\beta_\eta \quad c_4 \leq \delta_\eta \quad c_5 \geq \frac{2}{(1-\eta)h_+}$$

□

REMARK 8.4. Taking $\zeta_0 = 1$, (5.17) only holds when

$$0 \leq x \leq A'_m\left(\frac{(1-\eta)h_+}{2}\right)^2 (nT^{-1})^2\left(\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{a_\lambda}\right\}\right)^2 - C'_m \ .$$

Otherwise, we have, on a set of probability at least $1-e^{-x}$,

$$Z_m \leq \mathbb{E}(Z_m) + B'_m + \frac{C'_m+x}{\mu_0} + \frac{A'_m\mu_0}{4} \leq \mathbb{E}(Z_m) + B'_m + 2\frac{C'_m+x}{\mu_0}$$

$$\text{with} \quad \mu_0 = (1-\eta)h_+nT^{-1}\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{a_\lambda}\right\} \ .$$

Indeed, when y is too large, we take

$$\mu = \mu_0 = (1-\eta)h_+nT^{-1}\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{a_\lambda}\right\}$$

instead and thus obtain

$$\mathbb{P}\left(Z_m - \mathbb{E}(Z_m) \geq y\right) \leq \exp\left(\frac{A'_m\mu_0^2}{4} - (B'_m-y)\mu_0 + C'_m\right)$$

which can be rewritten as

$$\mathbb{P}\left(Z_m - \mathbb{E}(Z_m) \geq B'_m + \frac{C'_m+x}{\mu_0} + \frac{A'_m\mu_0}{4}\right) \leq e^{-x} \ .$$

In the proof above, we used the two following lemmas.

LEMMA 8.15 (Exponential moments for the inverse of a binomial). *Let $X \sim \mathcal{B}(n,p)$ a binomial random variable with parameters $n \in \mathbb{N}\backslash\{0\}$ and $p \in (0;1]$.*

Let $c_1 = 0.183$ and $c'_1 = 4.5$. Then, for all $\mu > 0$,

$$\mathbb{E}\left[\exp(\mu(\mathbb{E}[f_1(X)] - f_1(X)))\right] \leq \exp\left(\frac{c'_1\mu^2}{(np)^2} + \frac{\mu}{c_1\times(np)}\varphi(c_1\times np)\right) + \exp(-c_1\times np)$$
$$\leq \exp\left(\frac{c'_1\mu^2}{(np)^2} + \frac{\mu}{c_1e\times(np)}\right) + 1 \ .$$

LEMMA 8.16 (Exponential moments for the inverse of a binomial). *Let $X \sim \mathcal{B}(n,p)$ a binomial random variable with parameters $n \in \mathbb{N}\backslash\{0\}$ and $p \in (0;1]$; $T \in (0;1]$. For all $\eta \in (0;1)$, there*

exists $\beta_\eta > 1$, $\gamma_\eta, \delta_\eta > 0$, *such that for all* $\zeta_0 \geq 1$ *and* $0 \leq \mu \leq (1-\eta)h_+ npT^{-1}$,

$$\mathbb{E}\left[e^{\mu(f(X)-\mathbb{E}[f(X)])}\right] \leq \exp\left(\frac{\mu^2\beta_\eta^2\zeta_0^2}{2(np)^2} + \frac{\mu}{np\gamma_\eta}\varphi\left(np\gamma_\eta\right) + \exp(-np\delta_\eta)\right) \quad . \tag{8.16}$$

PROOF OF LEMMA 8.15. Let $Z = f_1(X)$ and $\psi^-(\mu) = \mathbb{E}[e^{\mu(\mathbb{E}(Z)-Z)}]$.

Split of ψ^-. Let $\alpha \geq 0$, and $Z_\alpha = f_1(X) \wedge \alpha$. We have

$$\begin{aligned}\psi^-(\mu) &= \mathbb{E}\left[\exp\left(\mu(\mathbb{E}[Z]-Z)\right)\mathbb{1}_{Z\leq\alpha}\right] + \mathbb{E}\left[\exp\left(\mu(\mathbb{E}[Z]-Z)\right)\mathbb{1}_{Z>\alpha}\right] \\ &\leq \exp\left(\mu(\mathbb{E}[Z]-\mathbb{E}[Z_\alpha])\right)\mathbb{E}\left[\exp\left(\mu(\mathbb{E}[Z_\alpha]-Z_\alpha)\right)\right] + \exp\left(\mu(\mathbb{E}[Z]-\alpha)\right)\mathbb{P}(Z>\alpha) \quad .\end{aligned}$$

Since $Z \in [0;1]$,

$$\mathbb{E}[Z] = \mathbb{E}[Z_\alpha \mathbb{1}_{Z\leq\alpha}] + \mathbb{E}[Z\mathbb{1}_{Z\geq\alpha}] \leq \mathbb{E}[Z_\alpha] + \mathbb{P}(Z \geq \alpha) \quad ,$$

and by Hoeffding's lemma ([**Mas07**], Lemma 2.6), we obtain

$$\psi^-(\mu) \leq \inf_{\alpha\geq\mathbb{E}[Z]}\left\{\exp\left(\frac{\mu^2\alpha^2}{2} + \mu\mathbb{P}(Z\geq\alpha)\right) + \mathbb{P}(Z\geq\alpha)\right\} \quad .$$

The idea of the proof is to take $\alpha = \beta(np)^{-1} \geq \mathbb{E}[Z]$. By Lemma 5.3 in Sect. 5.6.1, this holds as soon as

$$\beta \geq \sup_{np\geq B_n}\left\{e^0_{\mathcal{B}(n,p)} + (1-p)^n\right\} \quad .$$

Deviation bound. For all $\alpha > 0$,

$$\mathbb{P}(Z \geq \alpha) = \mathbb{P}\left(X \leq \frac{1}{\alpha}\right)$$

and Bernstein inequality ([**Mas07**], Prop. 2.9) applied to X gives for all $x \geq 0$

$$\mathbb{P}\left(X \leq np - \sqrt{2np(1-p)x} - \frac{x}{3}\right) \leq e^{-x} \quad .$$

For all $\beta \in (1;+\infty)$ we define

$$L(\beta) = \frac{9}{2}\left(\sqrt{1+\frac{2}{3}(1-\beta^{-1})} - 1\right)^2 \in (0; h_+) \tag{8.17}$$

so that

$$\mathbb{P}\left(Z \geq \beta(np)^{-1}\right) \leq e^{-npL(\beta)} =: e^{-x_\beta} \quad .$$

Notice that L is increasing on $(1;+\infty)$. If $0 < c_1 \leq L(\beta)$ and $c_1' = \frac{\beta^2}{2}$ for some β large enough, we obtain

$$\begin{aligned}\psi^-(\mu) &\leq \exp\left(\frac{c_1'\mu^2}{(np)^2} + \frac{\mu}{c_1\times(np)}\varphi(c_1\times np)\right) + \exp(-c_1\times np) \\ &\leq \exp\left(\frac{c_1'\mu^2}{(np)^2} + \frac{\mu}{c_1 e\times(np)}\right) + 1\end{aligned}$$

because φ is maximal for $x = 1$.

For instance, we can take $\beta = 3$ (using (5.60)), so

$$L(3) = \frac{(\sqrt{13}-3)^2}{2} \geq 0.183 = c_1 \quad \text{and} \quad c_1' = \frac{3^2}{2} = \frac{9}{2} \quad .$$

□

REMARK 8.5. In our proof, we obtain a more general upper bound which allows to choose β depending on μ, n and p.

PROOF OF LEMMA 8.16. We start as in the proof of lemma 8.15: let $Z = f_T(X)$, $Z_\alpha = f_T(X) \wedge \alpha$ and $\psi^+(\mu) = \mathbb{E}[e^{\mu(Z-\mathbb{E}(Z)}]$ for all $\mu \geq 0$. We split ψ^+, take $\alpha = \beta(np)^{-1}$ with $\beta > 1$ and deduce

$$\begin{aligned}\psi^+(\mu) &\leq \exp\left(\frac{\mu^2\alpha^2}{2} + \mu\mathbb{P}(Z \geq \alpha)\right) + \mathbb{E}\left[e^{\mu(Z-\mathbb{E}[Z])}\mathbb{1}_{Z>\alpha}\right] \\ &\leq \exp\left(\frac{\mu^2\beta^2}{2(np)^2} + \mu\exp(-npL(\beta))\right) + \mathbb{E}\left[e^{\mu(Z-\mathbb{E}[Z])}\mathbb{1}_{Z>\frac{\beta}{np}}\right]\end{aligned}$$

with $L(\beta)$ defined by (8.17).

In order to control the remaining term, we use that $Z \leq T$:

$$\mathbb{E}\left[e^{\mu(Z-\mathbb{E}[Z])}\mathbb{1}_{Z>\frac{\beta}{np}}\right] \leq e^{T\mu}\mathbb{P}\left(Z > \frac{\beta}{np}\right) \leq e^{T\mu - npL(\beta)} \ .$$

For all $\eta \in (0;1)$, there exists β_η such that $L(\beta_\eta) \geq \left(1-\frac{\eta}{2}\right)h_+$. Hence, for all $0 \leq \mu \leq (1-\eta)h_+ npT^{-1}$,

$$\begin{aligned}\psi^+(\mu) &\leq \exp\left(\frac{\mu^2\beta_\eta^2}{2(np)^2} + \mu\exp(-npL(\beta_\eta))\right) + e^{-\frac{np\eta h_+}{2}} \\ &\leq \exp\left(\frac{\mu^2\beta_\eta^2\zeta_0^2}{2(np)^2} + \mu\exp\left(-np\left(1-\frac{\eta}{2}\right)h_+\right) + \exp\left(-\frac{np\eta h_+}{2}\right)\right) \ .\end{aligned}$$

The result follows with

$$\beta_\eta^{-1} = 1 - \sqrt{h_+(2-\eta)} - \frac{h_+\left(1-\frac{\eta}{2}\right)}{3} \in (0;1) \qquad \gamma_\eta = \left(1-\frac{\eta}{2}\right)h_+ \qquad \delta_\eta = \frac{\eta h_+}{2} \ .$$

□

8.9. Moment inequalities for some U-statistics

In this section, we prove Prop. 5.5 which is stated in Sect. 5.6.3.

PROOF OF PROP. 5.5. We split $Z - \mathbb{E}[Z]$ into two parts:

$$Z - \mathbb{E}[Z] = \sum_{\lambda\in\Lambda_m}(a_\lambda + b_\lambda)(S_{\lambda,2} - \mathbb{E}[S_{\lambda,2}]) + \sum_{\lambda\in\Lambda_m} b_\lambda(S_{\lambda,1}^2 - S_{\lambda,2}) = Z_1 + Z_2.$$

Moments of Z_1. It is a sum of centered independent random variables. According to Lemma 8.18,

$$\begin{aligned}\|Z_1\|_q &\leq 2\sqrt{\kappa}\sqrt{q}\sqrt{\sum_{\lambda\in\Lambda_m}\sum_{1\leq i\leq r_\lambda}(a_\lambda+b_\lambda)^2\left\|\xi_{\lambda,i}^2 - m_{2,\lambda,i}^2\right\|_q^2} \\ &= 4\sqrt{\kappa}\sqrt{q}\sqrt{\sum_{\lambda\in\Lambda_m}\left((a_\lambda+b_\lambda)^2\sum_{i=1}^{r_\lambda} m_{2q,\lambda,i}^4\right)}\end{aligned}$$

since

$$\left\|\xi_{\lambda,i}^2 - m_{2,\lambda,i}^2\right\|_q \leq 2\left\|\xi_{\lambda,i}^2\right\|_q = 2m_{2q,\lambda,i}^2 \ .$$

Moments of Z_2. It is a degenerated U-statistic of order 2:

$$S_{\lambda,1}^2 - S_{\lambda,2} = \sum_{1\leq i\neq j\leq r_\lambda}\xi_{\lambda,i}\xi_{\lambda,j} \ ,$$

we control its moments with Lemma 8.17:

$$\begin{aligned}\|Z_2\|_q &\le 2\sqrt{\kappa}\sqrt{q}\sqrt{\|V_2\|_{q/2}}\\ V_2 &= \sum_{\lambda\in\Lambda_m}\sum_{i=1}^{r_\lambda}\left(Z_2 - \mathbb{E}(Z_2 \mid (\xi_{\lambda',j})_{(\lambda',j)\neq(\lambda,i)})\right)^2\\ &= 4\sum_{\lambda\in\Lambda_m} b_\lambda^2 \sum_{i=1}^{r_\lambda}\left(S_{\lambda,1}-\xi_{\lambda,i}\right)^2\xi_{\lambda,i}^2\end{aligned}$$

The triangular inequality gives

$$\begin{aligned}\|V_2\|_{q/2} &\le 4\sum_{\lambda\in\Lambda_m} b_\lambda^2\sum_{i=1}^{r_\lambda}\left\|\xi_{\lambda,i}^2(S_{\lambda,1}-\xi_{\lambda,i})^2\right\|_{q/2}\\ &\le 2\sum_{\lambda\in\Lambda_m} b_\lambda^2\sum_{i=1}^{r_\lambda}\left\|\eta_{\lambda,i}\xi_{\lambda,i}^4+\eta_{\lambda,i}^{-1}(S_{\lambda,1}-\xi_{\lambda,i})^4\right\|_{q/2}\\ &\le 2\sum_{\lambda\in\Lambda_m} b_\lambda^2\sum_{i=1}^{r_\lambda}\left(\eta_{\lambda,i}m_{2q,\lambda,i}^4+\eta_{\lambda,i}^{-1}\left\|S_{\lambda,1}-\xi_{\lambda,i}\right\|_{2q}^4\right)\\ &\le 2\sum_{\lambda\in\Lambda_m} b_\lambda^2\sum_{i=1}^{r_\lambda}\left(\eta_{\lambda,i}m_{2q,\lambda,i}^4+\eta_{\lambda,i}^{-1}2^6\kappa^2q^2\left(\sum_{1\le j\le r_\lambda,\, j\neq i} m_{2q,\lambda,j}^2\right)^2\right)\ .\end{aligned}$$

for any $\eta_{\lambda,i}>0$, the last inequality coming from Lemma 8.18. Taking

$$\eta_{\lambda,i}=\frac{8q\kappa\sum_{1\le j\le r_\lambda,\, j\neq i} m_{2q,\lambda,j}^2}{m_{2q,\lambda,i}^2}\ ,$$

we obtain

$$\|V_2\|_{q/2}\le 32\kappa q\sum_{\lambda\in\Lambda_m}\left(b_\lambda^2\sum_{1\le i\neq j\le r_\lambda} m_{2q,\lambda,i}^2m_{2q,\lambda,j}^2\right)\ .$$

Hence,

$$\|Z_2\|_q\le 8\sqrt{2}\kappa q\sqrt{\sum_{\lambda\in\Lambda_m}\left(b_\lambda^2\sum_{1\le i\neq j\le r_\lambda} m_{2q,\lambda,i}^2m_{2q,\lambda,j}^2\right)}\ .$$

□

In the proof above, we need the following two corollaries of Thm. 2 in [**BBLM05**].

LEMMA 8.17. *Let $(X_1,\dots,X_n)$ be n independent random variables, f a measurable function $\mathbb{R}^n\mapsto\mathbb{R}$ and*

$$Z=f(X_1,\dots,X_n)\ .$$

Then, there exists $\kappa\le 1.271$ such that for every $q\ge 2$,

$$\|Z-\mathbb{E}[Z]\|_q\le 2\sqrt{\kappa}\sqrt{q\left\|\sum_{i=1}^n\left(Z-\mathbb{E}\left[Z\mid (X_j)_{j\neq i}\right]\right)^2\right\|_{q/2}}\ . \tag{8.18}$$

LEMMA 8.18. *Let $(X_1, \dots, X_n)$ be n independent random variables admitting q-th moments for some $q \geq 2$: $m_{i,q} = \mathbb{E}[|X_i|^q]^{1/q}$. Let $S = \sum_{i=1}^s X_i$. Then,*

$$\|S\|_q \leq 2\sqrt{\kappa}\sqrt{q}\sqrt{\sum_{i=1}^s m_{i,q}^2}.$$

PROOF OF LEMMA 8.18. Apply Lemma 8.17 to S:

$$\begin{aligned}\|S - \mathbb{E}[S]\|_q &\leq 2\sqrt{\kappa}\sqrt{q\left\|\sum_{i=1}^s \mathbb{E}\left[(S - \mathbb{E}[S \mid X_i])^2 \mid X_{1\dots n}\right]\right\|_{q/2}} \\ &= 2\sqrt{\kappa}\sqrt{q\left\|\sum_{i=1}^s X_i^2\right\|_{q/2}} \leq 2\sqrt{\kappa}\sqrt{q\sum_{i=1}^s \|X_i\|_q^2} \ .\end{aligned}$$

□

8.10. Approximation properties of histograms

In Chapt. 6, Thm. 6.1, we need the following assumption:

(**Ap**) Polynomial decreasing of the bias: there exists $\beta_1 \geq \beta_2 > 0$ and $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ such that

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \leq \|s - s_m\|_{L^2(\mu)} \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \ .$$

where $\mu = \mathcal{L}(X)$ and s_m is the $L^2(\mu)$ projection onto some histogram model $(S_m)_{m\in\mathcal{M}}$. It is somehow unintuitive, since it assumes that s is not too well approximated by histograms. For instance, it excludes the case of constant functions, which are both α-hölderian (for any α) and histogram functions. Lemma 8.19 below shows that it is the only excluded function among the hölderian ones. On approximation theory, we refer to the book of DeVore and Lorentz [**DL93**].

8.10.1. Results. Let $(\mathcal{X}, d)$ be a metric space. For every $\alpha \in (0; 1]$, $\delta, \epsilon, R > 0$, we define $\mathcal{H}_{\delta,\epsilon}(\alpha, R)$ the set of α-hölderian functions f on $\mathcal{X}$, *i.e.*:

$$\forall x, y \in [0; 1], \qquad |s(x) - s(y)| \leq R d(x, y)^\alpha$$

such that there exists $x_1, x_2 \in \mathcal{X}$ such that

$$d(x_1, x_2) \leq \delta \qquad \text{and} \qquad |s(x_1) - s(x_2)| \geq \epsilon \ .$$

When $\mathcal{X}$ is bounded, we also define

$$\mathcal{H}_\epsilon(\alpha, R) := \mathcal{H}_{\mathrm{diam}(\mathcal{X}),\epsilon}(\alpha, R) \ .$$

Regular histograms in $[0; 1]$. We first investigate the simplest case, where $(\mathcal{X}, d) = ([0; 1], \|\cdot\|_\infty)$ and regular histograms.

LEMMA 8.19. *Let $\alpha \in (0; 1]$, $\delta, \epsilon, R > 0$ and $s \in \mathcal{H}_{\delta,\epsilon}(\alpha, R)$.*

For every $D \in \mathbb{N}$, denote by s_D the $L^2(\mathrm{Leb})$ projection of s on the space of regular histograms with D pieces. Then, there exists a constant

$$C_1 = L(\alpha) R^{-\alpha^{-1}} \epsilon^{2+\alpha^{-1}} |x_1 - x_2|^{-1-\alpha^{-1}} > 0$$

$$\text{and} \qquad C_2 = R^2 \qquad \beta_1 = 1 + \frac{1}{\alpha} \qquad \beta_2 = 2\alpha$$

such that for all $D > 0$,

$$\frac{C_1}{D^{\beta_1}} \leq \|s - s_D\|_{L^2(\mathrm{Leb})}^2 \leq \frac{C_2}{D^{\beta_2}} \ . \tag{8.19}$$

REMARK 8.6. The upper bound holds with any probability measure μ on $\mathcal{X}$ instead of Leb, since

$$l(s, s_m) \leq \|s - s_m\|_\infty^2 \leq R^2 D_m^{-2\alpha} \ .$$

If $(\mathbf{Ad}_\ell)$ holds, then

$$l(s, s_m) \geq c_{\min}^X \operatorname{Leb}(\mathcal{X})^{-1} \|s - s_m\|_{L^2(\mathrm{Leb})}^2 \geq c \|s - s_{D_m}\|_{L^2(\mathrm{Leb})}^2$$

and thus the lower bound is still valid.

REMARK 8.7. The lower bound in (8.19) cannot be improved, as shown in Sect. 8.10.1: for every $\alpha, R, \delta, \epsilon > 0$, there exists C_1' such that for every D,

$$\inf_{s \in \mathcal{H}_{\delta,\epsilon}(\alpha,R)} \left\{ \|s - s_D\|_{L^2(\mathrm{Leb})}^2 \right\} \leq \frac{C_1'}{D^{1+\alpha^{-1}}} \ .$$

Regular histograms in $\mathbb{R}^k$. We now generalize the previous result to subsets of $\mathbb{R}^k$. For the sake of simplicity, we assume that $\mathcal{X}$ is a ball of $(\mathbb{R}^k, \|\cdot\|_\infty)$. Otherwise, if $\overset{o}{\mathcal{X}}$ is connex and non-empty, any non-constant continuous function s on $\mathcal{X}$ is non-constant on some ball $B(s) \subset \mathcal{X}$. Then, we can apply Lemma 8.20 on $\mathcal{B}(s)$ in order to derive $(\mathbf{Ap})$. The constants $\delta, \epsilon > 0$ have to take into account the restriction $x_1, x_2 \in B(s) \subset \mathcal{X}$ in the definition of $\mathcal{H}_{\delta,\epsilon}(\alpha, R)$. When $\mathcal{X}$ is a ball, this condition is automatically satisfied.

LEMMA 8.20. *Let $\mathcal{X}$ be a non-empty closed ball of $(\mathbb{R}^k, \|\cdot\|_\infty)$ and $s \in \mathcal{H}_{\delta,\epsilon}(\alpha, R)$. Let $D \in \mathbb{N} \backslash \{0\}$ and consider the "regular" partition (I_λ) of $\mathcal{X}$ of pace D^{-1},* i.e. *the collection of non-empty intersections between $\mathcal{X}$ and the family $\left(\prod_{i=1}^k [\frac{j_i}{D}; \frac{j_i+1}{D}) \right)_{j_1,\ldots,j_k \in \mathbb{Z}}$. Let s_D be the piecewise constant function, defined on each piece I_λ of this partition by*

$$s_D \equiv \frac{1}{\operatorname{Leb}(I_\lambda)} \int_{I_\lambda} s(t) dt \ .$$

Then,

$$\begin{aligned} \int_{\mathcal{X}} (s(t) - s_D)^2 \, dt \geq L_{k,\alpha} \epsilon^{2+k\alpha^{-1}} \delta^{-1-k\alpha^{-1}} R^{1-k(1+\alpha^{-1})} \\ \times \left(D \vee \delta^{-1} \right)^{-1-k\alpha^{-1}+(k-1)\alpha} \ . \end{aligned} \tag{8.20}$$

REMARK 8.8.

(1) The number of pieces in the partition is not D but (approximatively, depending on the shape of $\mathcal{X}$) $\operatorname{Leb}(\mathcal{X}) D^k$. Then, if X has a lower bounded density w.r.t. Leb on $\mathcal{X}$, under the assumptions of Lemma 8.20, $(\mathbf{Ap})$ is satisfied with $\beta_1 = k^{-1} + \alpha^{-1} - (k-1)k^{-1}\alpha$.

(2) The following upper bound on the bias is straightforward:

$$\frac{1}{\operatorname{Leb}(\mathcal{X})} \int_{\mathcal{X}} (s(t) - s_D)^2 \, dt \leq \|s - s_D\|_\infty^2 \leq R^2 D^{-2\alpha} \ .$$

(3) When $\mathcal{X}$ is not a ball of $\mathbb{R}^k$, we can use a general argument assuming only that for every $x_1, x_2 \in \overset{o}{\mathcal{X}}$, there is a path from x_1 to x_2 that has an η-enlargement in $\mathcal{X}$ for some $\eta > 0$. We then obtain (8.21) instead of (8.20), but this still implies $(\mathbf{Ap})$.

Optimality of the lower bound in $[0;1]$. When $\mathcal{X}$ is a non-empty compact interval of $\mathbb{R}$, the exponent $1 + \alpha^{-1}$ in Lemma 8.19 is unimprovable in the following sense. Without any loss of generality, we assume that $\mathcal{X} = [0;1]$.

LEMMA 8.21. *Let $\mathcal{X} = [0;1]$, $R > 0$, $\alpha \in (0;1]$, $1 \geq \delta \geq (1+\eta)D^{-1}$ (for some $\eta > 0$) and $L(\alpha)R\lfloor D\delta \rfloor D^{-1} \geq \epsilon > 0$.*

$$\inf_{s \in \mathcal{H}_{\delta,\epsilon}(\alpha,R)} \left\{ \int_{\mathcal{X}} \left(s(t) - s_D(t) \right)^2 dt \right\} \leq L(\alpha,\eta) R^{-\alpha^{-1}} \epsilon^{2+\alpha^{-1}} \delta^{-1-\alpha^{-1}} D^{-1-\alpha^{-1}} \ .$$

REMARK 8.9. If $D \geq 2\delta^{-1}$, one can replace η by 1 and this upper bound is (up to some factor $L(\alpha)$) the same as the lower bound in Lemma 8.19.

Thus, the exponent $\beta_1 = 1 + \alpha^{-1}$ cannot be improved as long as we look for a uniform bound on $\mathcal{H}_{\delta,\epsilon}(\alpha, R)$. However, this does not mean that there exists a function $s \in \mathcal{H}(\alpha, R)$ approximated by regular histograms at the rate $D^{-1-\alpha^{-1}}$. To our knowledge, this question remains unsolved. Some references about this problem (and the equivalent one when the knots of the partition are no longer fixed) may be found in Burchard and Hale [**BH75**]. See also the book of DeVore and Lorentz [**DL93**], in particular Chap. 12.

8.10.2. Proofs.

Regular histograms in $[0;1]$.

PROOF OF LEMMA 8.19. The upper bound directly follows from

$$l(s, s_D) \leq \|s - s_D\|_\infty^2 \leq R^2 D^{-2\alpha}$$

so that

$$\beta_2 = 2\alpha \qquad C_2 = R^2 \ .$$

The lower bound needs some more work. As $s \in \mathcal{H}_{\delta,\epsilon}(\alpha, R)$, there exists $x_1 < x_2$ in $[0;1]$ such that $|s(x_2) - s(x_1)| \geq \epsilon > 0$. The interval $[x_1, x_2]$ intersects $N(D) \leq 2 + D|x_2 - x_1| \leq 2 + D\delta$ intervals I_λ of the regular histogram of size D (denote by $\Lambda_D(x_1, x_2)$ this set). On each of the $(I_\lambda)_{\lambda \in \Lambda_D(x_1,x_2)}$, the variation

$$\mathrm{varia}_{I_\lambda}(s) = \sup_{I_\lambda} s - \inf_{I_\lambda} s$$

of s may be at most $RD^{-\alpha}$ since $s \in \mathcal{H}(\alpha, R)$, and the sum of the $N(D)$ variations $\mathrm{varia}_{I_\lambda}(s)$ is larger or equal to ϵ.

From lemma 8.22, we have

$$\begin{aligned}
\|s - s_D\|_2^2 &\geq \sum_{\lambda \in \Lambda_D(x_1,x_2)} \int_{I_\lambda} \left(s(t) - s_D(t) \right)^2 dt \\
&\geq \sum_{\lambda \in \Lambda_D(x_1,x_2)} R^{-\alpha^{-1}} 2^{-4-2\alpha^{-1}} \, \mathrm{varia}_{I_\lambda}(s)^{2+\alpha^{-1}} \\
&\geq R^{-\alpha^{-1}} 2^{-4-2\alpha^{-1}} \left(\sum_{\lambda \in \Lambda_D(x_1,x_2)} \mathrm{varia}_{I_\lambda}(s) \right)^{2+\alpha^{-1}} N(D)^{-1-\alpha^{-1}} \\
&\geq R^{-\alpha^{-1}} 2^{-4-2\alpha^{-1}} \epsilon^{2+\alpha^{-1}} N(D)^{-1-\alpha^{-1}}
\end{aligned}$$

where we used Hölder inequality in the last but one line, since $2 + \alpha^{-1} \geq 1$.

When $D \geq (2\delta)^{-1}$, we deduce that

$$\|s - s_D\|_2^2 \geq L(\alpha) R^{-\alpha^{-1}} \epsilon^{2+\alpha^{-1}} \delta^{-1-\alpha^{-1}} D^{-1-\alpha^{-1}} \ .$$

Since for every $k, D \in \mathbb{N} \setminus \{0\}$, $\|s - s_D\|_2 \geq \|s - s_{kD}\|_2$, we get for every $D \geq 1$ that

$$\|s - s_D\|_2^2 \geq L(\alpha) R^{-\alpha^{-1}} \epsilon^{2+\alpha^{-1}} \delta^{-1-\alpha^{-1}} \left(D \vee \delta^{-1} \right)^{-1-\alpha^{-1}} \ .$$

□

LEMMA 8.22. *Let s be an (α, R) hölderian function on a compact interval I. Denote by $\mathrm{varia}_I(s) := \sup_I s - \inf_I s$ the variation of s on I and s_I the mean of s on I. Then,*

$$\int_I (s(x) - s_I)^2 \, dx \geq \frac{\mathrm{varia}_I(s)^{2+\frac{1}{\alpha}}}{R^{\alpha^{-1}} 2^{4+2\alpha^{-1}}} \ .$$

PROOF OF LEMMA 8.22. Let $T = \mathrm{varia}_I(s)$. One among $\sup_I s$ and $\inf_I s$ must be at distance at least $\frac{T}{2}$ of s_I. By symmetry, we can assume that it is the supremum. As s is continuous, $\sup_I s = s(x_0)$ for some $x_0 \in I$. As $s \in \mathcal{H}(\alpha, R)$, there is an interval J around x_0, of size

$$\mathrm{Leb}(J) \geq \left(\frac{\sup_I s - s_I}{2R} \right)^{\alpha^{-1}} \geq \left(\frac{T}{4R} \right)^{\alpha^{-1}}$$

and on which

$$s \geq \frac{s_I + \sup_I s}{2} \geq \frac{T}{4} \ .$$

As a consequence,

$$\int_I (s(x) - s_I)^2 \, dx \geq \int_J (s(x) - s_I)^2 \, dx \geq \left(\frac{T}{4R} \right)^{\alpha^{-1}} \frac{T^2}{16} \ .$$

□

Regular histograms in $\mathbb{R}^k$.

PROOF OF LEMMA 8.20. A first simple reasoning leads to a large exponent in the lower bound, that is sufficient to derive (**Ap**) and may be easier to generalize to other metric spaces. In the case of $\mathbb{R}^k$, we get a more accurate result.

General argument. Let $x_1, x_2 \in \mathcal{X}$ that comes from $s \in \mathcal{H}_{\delta,\epsilon}(\alpha, R)$ (defined at the beginning of Sect. 8.10.1), and

$$A = \mathcal{B}_\infty \left(\frac{x_1 + x_2}{2}, \frac{\|x_1 - x_2\|_\infty}{2} \right) \ .$$

Since $\mathcal{X}$ is a ball, $A \subset \mathcal{X}$. There exists a path from x_1 to x_2 in A that crosses

$$N(D) \leq 2 + D \, \|x_1 - x_2\|_\infty \leq 2 + D\delta$$

pieces of the regular partition. Denote by $J_1, \ldots, J_{N(D)}$ those pieces.

Along this path, s is varying at least of ϵ, so that

$$\sum_{i=1}^{N(D)} \mathrm{varia}_{J_i}(s) \geq \epsilon \ .$$

From (8.24), we have

$$\int_A (s - s_D)^2 \geq \sum_{i=1}^{N(D)} \int_{J_i} (s - s_D)^2 \geq \frac{\sum_{i=1}^{N(D)} (\mathrm{varia}_{J_i}(s))^{2+k\alpha^{-1}}}{144 \times 4^{k\alpha^{-1}} R^{k\alpha^{-1}}} \ .$$

Since $2 + k\alpha^{-1} > 1$, we can apply Hölder inequality to deduce

$$\begin{aligned} \sum_{i=1}^{N(D)} (\mathrm{varia}_{J_i}(s))^{2+k\alpha^{-1}} &\geq \left(\sum_{i=1}^{N(D)} \mathrm{varia}_{J_i}(s) \right)^{2+k\alpha^{-1}} N(D)^{-1-k\alpha^{-1}} \\ &\geq \epsilon^{2+k\alpha^{-1}} (2 + D\delta)^{-1-k\alpha^{-1}} \ . \end{aligned}$$

Thus, for every $D \geq (2\delta)^{-1}$,

$$\int_A (s - s_D)^2 \geq L(k,\alpha)\epsilon^{2+k\alpha^{-1}}\delta^{-1-k\alpha^{-1}}R^{-k\alpha^{-1}}D^{-1-k\alpha^{-1}} \quad . \tag{8.21}$$

Improved lower bound. In the previous argument, we only used the existence of one path from x_1 to x_2 in $A \subset \mathcal{X}$ that crosses

$$N(D) \leq 2k + D\left\|x_1 - x_2\right\|_1 \propto kD\delta$$

pieces of the regular partition. We did not actually need that the whole ball A is in $\mathcal{X}$. Using this fact, we can find $\propto D^{k-1}$ such paths, "sufficiently" distinct, so that we obtain a better lower bound. We will show that it is possible with at most $N_c(D) \leq \frac{2}{3}\epsilon R^{-1}D^{\alpha}$ points in each path that may be shared with the union of the other paths.

Consider the spheres $S_1 = S(x_1, \frac{D^{\alpha-1}}{3R})$ and $S_2 = S(x_2, \frac{D^{\alpha-1}}{3R})$ for the norm $\|\cdot\|_1$. Denote also by O_1 (resp. O_2) the orthant of $\mathbb{R}^k$ that contains x_2 (resp. x_1), taking the origin at x_1 (resp. x_2). The idea of the proof is to build monotonic paths from x_1 to x_2 (each coordinate is monotonic) that goes from x_1 to $S_1 \cap O_1$, then from $S_1 \cap O_1$ to $S_2 \cap O_2$, and finally from $S_2 \cap O_2$ to x_2. The point here is that the second step can be made with $\mathrm{Card}(S_1 \cap O_1) = \mathrm{Card}(S_2 \cap O_2)$ completely disjoint paths, in the sense that they cross disjoint pieces of the regular partition. This is done by translating $S_1 \cap O_1$ onto $S_2 \cap O_2$ in $D\epsilon - \frac{2}{3}R^{-1}D^{\alpha}$ steps.

On each path, we have an inequality like (8.21), with

$$\epsilon - RD^{-\alpha} \times \frac{2D^{\alpha-1}}{3R} \geq \frac{\epsilon}{3}$$

instead of ϵ (we remove the paths from x_1 to $S_1 \cap O_1$, and from $S_2 \cap O_2$ to x_2, using that the variation of s on each of piece of the partition is at most $RD^{-\alpha}$. Since

$$\mathrm{Card}(S_1 \cap O_1) \geq L\left(\frac{D^{\alpha}}{3R}\right)^{k-1} \quad ,$$

we obtain, for every $D \geq (2\delta)^{-1}$,

$$\int_A (s - s_D)^2 \geq L_{k,\alpha}\epsilon^{2+k\alpha^{-1}}\delta^{-1-k\alpha^{-1}}R^{-k\alpha^{-1}-(k-1)}D^{-1-k\alpha^{-1}+(k-1)\alpha} \quad . \tag{8.22}$$

□

We now prove the analogous of Lemma 8.22 needed in the proof of Lemma 8.20. We prove it in a more general case, assuming only that $\mathcal{X}$ is a measurable metric space.

LEMMA 8.23. *Let $(\mathcal{X}, d, \mu)$ be a measurable metric space, $I \subset \mathcal{X}$ measurable with $\mu(I) \in (0;\infty)$, and s a measurable function $I \mapsto \mathbb{R}$. Assume that s is (α, R) hölderian,* i.e.

$$\forall x_1, x_2 \in I, \quad |s(x_1) - s(x_2)| \leq Rd(x_1,x_2)^{\alpha} \quad .$$

Assume that

$$s_I := \frac{1}{\mu(I)}\int_I s d\mu \quad \textit{and} \quad \mathrm{varia}_I(s) := \sup_I s - \inf_I s$$

both exist and are finite. Then, there exists some $x_0 \in I$ such that

$$\int_I (s - s_I)^2 \, d\mu \geq \frac{(\mathrm{varia}_I(s))^2 \, \mu\left(I \cap B_d\left(x_0, \left(\frac{\mathrm{varia}_I(s)}{4R}\right)^{1/\alpha}\right)\right)}{144} \quad . \tag{8.23}$$

PROOF OF LEMMA 8.23. There must be some point $x_0 \in I$ such that $s(x_0)$ is at a distance larger than $\frac{\mathrm{varia}_I(s)}{3}$ from s_I (otherwise, we must have $\mathrm{varia}_I(s) = 0$ and any $x_0 \in I$ is convenient).

If $x \in I$ and $d(x,x_0) \leq \left(\frac{\mathrm{varia}_I(s)}{4R}\right)^{1/\alpha}$, then $|s(x)-s(x_0)| \leq \frac{\mathrm{varia}_I(s)}{4}$ since s is (α,R) hölderian. By definition of x_0, we have $|s-s_I| \geq \frac{\mathrm{varia}_I(s)}{12}$ uniformly in $I \cap B_d\left(x_0, \left(\frac{\mathrm{varia}_I(s)}{4R}\right)^{1/\alpha}\right)$. The result follows. □

In Lemma 8.20, we use Lemma 8.23 in the following framework. Let $\mathcal{X} = \mathbb{R}^k$, $d = \|\cdot\|_\infty$, $\mu = \mathrm{Leb}$ and $I = B_d(c, r_I)$ for some $c \in \mathcal{X}$ and $r \geq 0$, we have for any $x_0 \in I$ and any $s \geq 0$,

$$\mu\left(B_d(x_0,s) \cap I\right) \geq \left(r_I \wedge s\right)^k \quad .$$

As a consequence, if $s : \mathbb{R}^k \mapsto \mathbb{R}$ is (α,R) hölderian (for the norm $\|\cdot\|_\infty$), we have $s_I < \infty$, $\mathrm{varia}_I(s) \leq R(2r_I)^\alpha < \infty$ and

$$\begin{aligned}\int_I \left(s(x)-s_I\right)^2 \mu(dx) &\geq \frac{\left(\mathrm{varia}_I(s)\right)^2 \left(r_I \wedge \left(\frac{\mathrm{varia}_I(s)}{4R}\right)^{\alpha^{-1}}\right)^k}{144} \\ &= \frac{\left(\mathrm{varia}_I(s)\right)^{2+k/\alpha}}{144 \times 4^{k\alpha^{-1}} R^{k\alpha^{-1}}} \quad . \end{aligned} \tag{8.24}$$

Optimality of the lower bound in $[0;1]$. We first have to prove the "optimality" of Lemma 8.22.

LEMMA 8.24. *Let I be a compact interval of $\mathbb{R}$. Let $R > 0$, $\alpha \in (0;1]$, $T \in [0; Ra^\alpha]$. Then, there exists an increasing function $s \in \mathcal{H}(\alpha,R)$ such that* $\mathrm{varia}_I(s) = T$ *and* $\int_I \left(s(t)-s_I\right)^2 dt \leq L(\alpha) R^{-\alpha^{-1}} T^{2+\alpha^{-1}}$.

PROOF OF LEMMA 8.24. Let a be the length of I; up to a translation, we can assume that $I = [0;a]$.

Consider $s(x) = R\left(x \wedge \epsilon\right)^\alpha$ with $\epsilon = (T/R)^{\alpha^{-1}} \leq a$, we have $s \in \mathcal{H}(\alpha,R)$ and $\mathrm{varia}_I(s) = T$. Straigthforward computations show that

$$\begin{aligned} s_I &= a^{-1}\int_0^\epsilon Rt^\alpha dt + a^{-1}\int_\epsilon^a R\epsilon^\alpha dt = \frac{R\epsilon^{\alpha+1}}{a(\alpha+1)} + R\epsilon^\alpha \frac{a-\epsilon}{a} \\ &= R\epsilon^\alpha - \frac{\alpha R\epsilon^{\alpha+1}}{a(\alpha+1)} = R\epsilon^\alpha\left(1 - \frac{\alpha}{\alpha+1}\frac{\epsilon}{a}\right) \in \left[\frac{R\epsilon^\alpha}{\alpha+1}; R\epsilon^\alpha\right] \end{aligned}$$

$$\begin{aligned} \int_0^\epsilon \left(s(t)-s_I\right)^2 dt &= \int_0^\epsilon \left(Rt^\alpha - s_I\right)^2 dt = \frac{R^2\epsilon^{2\alpha+1}}{2\alpha+1} - 2s_I R\frac{\epsilon^{\alpha+1}}{\alpha+1} + \epsilon s_I^2 \\ &= \frac{R^2\epsilon^{2\alpha+1}}{2\alpha+1} - 2R\epsilon^\alpha\left(1 - \frac{\alpha}{\alpha+1}\frac{\epsilon}{a}\right)R\frac{\epsilon^{\alpha+1}}{\alpha+1} + R^2\epsilon\left(\epsilon^\alpha - \frac{\alpha}{\alpha+1}\frac{\epsilon^{\alpha+1}}{a}\right)^2 \\ &= R^2\epsilon^{2\alpha+1}\left(\frac{1}{2\alpha+1} - \frac{2}{\alpha+1}\left(1 - \frac{\alpha}{\alpha+1}\frac{\epsilon}{a}\right) + \left(1 - \frac{\alpha}{\alpha+1}\frac{\epsilon}{a}\right)^2\right) \\ &\leq R^2\epsilon^{2\alpha+1}\left(\frac{1}{2\alpha+1} + 1\right) \quad . \end{aligned}$$

Moreover, we have

$$\begin{aligned} \int_\epsilon^a \left(s(t)-s_I\right)^2 dt &= (a-\epsilon)\left(R\epsilon^\alpha - s_I\right)^2 = (a-\epsilon)\left(\frac{R\alpha}{\alpha+1}\right)^2 \frac{\epsilon^{2\alpha+2}}{a^2} \\ &\leq \left(R\epsilon^\alpha - s_I\right)^2 = \left(\frac{R\alpha}{\alpha+1}\right)^2 \frac{\epsilon^{2\alpha+2}}{a} \leq \left(\frac{R\alpha}{\alpha+1}\right)^2 \epsilon^{2\alpha+1} \quad . \end{aligned}$$

Thus,

$$\int_I (s(t) - s_I)^2 \, dt \leq R^2 \epsilon^{2\alpha+1} \left(\frac{1}{2\alpha+1} + 1 + \frac{\alpha^2}{(\alpha+1)^2} \right) = L(\alpha) R^{-\alpha^{-1}} T^{2+\alpha^{-1}} \quad .$$

□

PROOF OF LEMMA 8.21. We use here Lemma 8.24, and base our proof upon the one of Lemma 8.19. Let $x_1 = 0$ and $x_2 = \lfloor D\delta \rfloor D^{-1} \in \mathcal{X}$ (because $\delta \in (0;1]$). Then, $\delta - D^{-1} < |x_1 - x_2| \leq \delta$ and both x_1, x_2 are limit points of the regular partition of $\mathcal{X}$ with D pieces. There are exactly

$$N(D) = |x_1 - x_2| \, D = \lfloor D\delta \rfloor > D\delta - 1 \geq \frac{\eta}{1+\eta} D\delta$$

intervals of the partition between x_1 and x_2.

Let $T \in [0; RD^{-\alpha}]$ to be chosen later. Define s_0 as the continuous function on $\mathcal{X}$ such that:

- $s_0(0) = 0$,
- for every $k = 0, \ldots, \lfloor D\delta \rfloor - 1$, on $[kD^{-1}; (k+1)D^{-1})$, $s_0 - s_0(kD^{-1})$ is equal to the increasing function of Lemma 8.24 with a variation T,
- s_0 is constant on $[x_2; 1]$.

The resulting function s_0 is increasing, so that $|s_0(x_1) - s_0(x_2)| = TN(D)$. It is more difficult to show that $s_0 \in \mathcal{H}(\alpha, R)$, we will do it last.

From Lemma 8.24, we also have

$$\int_{\mathcal{X}} (s_0(t) - s_{0,D}(t))^2 \, dt \leq N(D) L(\alpha) R^{-\alpha^{-1}} T^{2+\alpha^{-1}}$$

Taking $T = \epsilon N(D)^{-1} \leq \epsilon \delta^{-1} D^{-1}$, we have $|s_0(x_1) - s_0(x_2)| \geq \epsilon$ and

$$\begin{aligned} \int_{\mathcal{X}} (s_0(t) - s_{0,D}(t))^2 \, dt &\leq L(\alpha) R^{-\alpha^{-1}} \epsilon^{2+\alpha^{-1}} N(D)^{-2-\alpha^{-1}} \\ &\leq L(\alpha) R^{-\alpha^{-1}} \epsilon^{2+\alpha^{-1}} \left(\frac{\eta\delta}{1+\eta} \right)^{-1-\alpha^{-1}} D^{-1-\alpha^{-1}} \quad . \end{aligned}$$

It remains to prove $s_0 \in \mathcal{H}(\alpha, R)$. Given $0 \leq y_1 \leq y_2 \leq 1$, we want to prove: $|s_0(y_1) - s_0(y_2)| \leq R \, |y_1 - y_2|^\alpha$. If y_1 and y_2 belong to the same interval of the regular partition, this is satisfied by construction of s_0. Otherwise, we must use the exact definition of s_0 that follows from the proof of Lemma 8.24.

Let $p = D^{-1}$ be the size of an interval of the regular partition. For every $x \in [kp; (k+1)p)$ (with $0 \leq k \leq D-1$ an integer), we have $s_0(x) = kT + (R(x - kp)^\alpha) \wedge T$. Since the increments of s_0 are periodic with period p, we can assume that y_1 belongs to the first interval of the regular partition. Then,

$$s_0(y_1) = (Ry_1^\alpha) \wedge T \quad \text{and} \quad s_0(y_2) = s_0(kp + h_2) = kT + (Rh_2^\alpha) \wedge T$$

and $|y_1 - y_2| = kp + h_2 - y_1$. Since $s_0(y_2)$ does not increase when $Rh_2^\alpha > T$ whereas $|y_1 - y_2|$ does, we can assume that $Rh_2^\alpha \leq T$.

When $Ry_1^\alpha > T$, it suffices to consider $y_1 = p$ (it is not in the first interval of the partition, but it belongs to its closure and s_0 is continuous). A previous remark ("periodicity") remains this case to $y_1 = 0$. Finally, the only case to consider is when $0 \leq Ry_1^\alpha \leq T$ and $0 \leq Rh_2^\alpha \leq T$. We then have

$$|s_0(y_1) - s_0(y_2)| = kT + R(h_2^\alpha - y_1^\alpha) \quad \text{and} \quad |y_1 - y_2| = kp + h_2 - y_1 \quad .$$

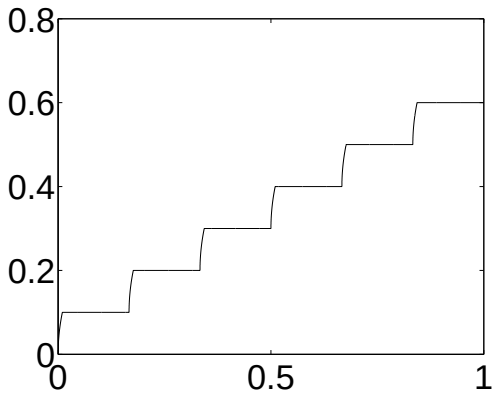


FIGURE 8.1. The function s of the proof of Lemma 8.21, with $D = 6$, $\alpha = 0.5$, $R = 1$, $\delta = 0.9$, $\epsilon = 0.5$.

We deduce the following condition on T: for every $1 \leq k \leq D-1$ and $0 \leq t_1, t_2 \leq 1$ (with $t_2 = (R/T)^{\alpha^{-1}} h_2$ and the same for t_1 and y_1):

$$T\left(k + (t_2^\alpha - t_1^\alpha)\right) \leq R\left(kp + T^{\alpha^{-1}} R^{-\alpha^{-1}} (t_2 - t_1)\right)^\alpha \quad .$$

Let $z = (T/R)^{\alpha^{-1}}$. This may be written as

$$z\left(k + (t_2^\alpha - t_1^\alpha)\right)^{\alpha^{-1}} \leq kp + z(t_2 - t_1)$$

which is equivalent to

$$z \leq \inf_{1 \leq k \leq D-1,\, 0 \leq t_1, t_2 \leq 1} \left\{ \frac{kp}{\left(k + (t_2^\alpha - t_1^\alpha)\right)^{\alpha^{-1}} - (t_2 - t_1)} \right\} \quad .$$

Keeping $k \geq 1$ fixed, the infimum is attained when the denominator $\left(k + (t_2^\alpha - t_1^\alpha)\right)^{\alpha^{-1}} - (t_2 - t_1)$ is maximal, *i.e.* when $t_1 = 0$ and $t_2 = 1$ (it is increasing in t_2, and decreasing then increasing in t_1; thus, the maximum is attained either at $(0,1)$ or at $(1,1)$; since $\alpha^{-1} \geq 1$, it is at $(0,1)$). We obtain the following condition on T:

$$\left(\frac{T}{R}\right)^{\alpha^{-1}} \leq \inf_{1 \leq k \leq D-1} \left\{ \frac{kp}{(k+1)^{\alpha^{-1}} - 1} \right\} \quad .$$

Since $\alpha^{-1} \geq 1$, $x \mapsto x^{\alpha^{-1}}$ is convex, so $\frac{(k+1)^{\alpha^{-1}} - 1}{(k+1) - 1}$ is increasing in k. Thus, the infimum is attained for $k = D - 1$, and the condition becomes

$$T \leq \left(\frac{D-1}{D^{\alpha^{-1}} - 1}\right)^\alpha Rp^\alpha \leq L(\alpha) R D^{-1}$$

so that the choice $T = \epsilon N(D)^{-1}$ is possible as soon as $\epsilon \leq L(\alpha) R D^{-1} \lfloor D\delta \rfloor$. □

CHAPTER 9

On the constant in front of global penalties

RÉSUMÉ. Ce chapitre aborde la question de la calibration de pénalités globales par rééchantillonnage, telles que les complexités de Rademacher ou les pénalités définies par Fromont [**Fro04**]. Il y a un facteur au moins 2 entre la borne théorique «pessimiste» et la borne observée dans différentes études de simulations. Nous montrons que la borne observée est valable sous une hypothèse additionnelle de symétrie, tandis que la borne «pessimiste» est atteinte dans un cas asymétrique limite. Il s'agit d'un phénomène hautement non-asymptotique, qui indique que la théorie ne peut rejoindre la pratique qu'au prix d'hypothèses supplémentaires, sans doute beaucoup moins restrictives que la symétrie. Ces résultats indiquent également qu'il peut être nécessaire de calibrer de telles pénalités en utilisant les données. On ne peut pas se contenter des bornes théoriques générales pour obtenir une procédure optimale.

9.1. Introduction

In this chapter, we consider global resampling penalties in classification. Following the introduction of Chap. 7 (and using the same notations), the ideal global penalty is

$$\mathrm{pen}_{\mathrm{id,g}}(m) := \sup_{t\in S_m} (P-P_n)\gamma(t) \geq (P-P_n)\gamma(\widehat{s}_m) \ . \tag{9.1}$$

We then call global penalty any estimator of $\mathrm{pen}_{\mathrm{id,g}}(m)$. Defining

$$\mathcal{F}_m := \{\xi \in \mathcal{X}\times\mathcal{Y} \mapsto \gamma(t,\xi) \quad \text{s.t.} \quad t\in S_m\} \ ,$$

the ideal global penalty can be written

$$\widehat{C}_n(\mathcal{F}) := \sup_{f\in\mathcal{F}} \{(P-P_n)(f)\}$$

for some class $\mathcal{F}$ of functions. Global penalties are thus *global complexity measures* of the classes $(\mathcal{F}_m)_{m\in\mathcal{M}_n}$.

The *Rademacher complexity* (independently introduced by Koltchinskii [**Kol01**] and Bartlett, Boucheron and Lugosi [**BBL02**]) is now a common global complexity measure in learning theory. It can be defined as follows:

$$\widehat{R}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(\xi_i)\right\} \,\middle|\, \xi_{1\dots n}\right] \tag{9.2}$$

where $\epsilon_1,\dots,\epsilon_n$ are i.i.d. Rademacher variables (*i.e.* equally distributed between $+1$ and -1), independent from the sample $\xi_{1\dots n}$. A similar complexity measure is the *Gaussian complexity* (see

[**BM02**] and the references therein), defined as

$$\widehat{G}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n} G_i f(\xi_i)\right\}\,\middle|\, \xi_{1\dots n}\right] \tag{9.3}$$

where $G_1,\dots,G_n$ are i.i.d. standard gaussian variables, independent from the sample $\xi_{1\dots n}$.

More recently, Fromont [**Fro04, Fro07**] defined more general complexities, called *bootstrap penalties*, which are based upon Efron's resampling heuristics. They can be written

$$\widehat{B}_n^{(Z)}(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i f(\xi_i)\right\}\,\middle|\, \xi_{1\dots n}\right] \tag{9.4}$$

where $(Z_1,\dots,Z_n)\in\mathbb{R}^n$ is any random vector independent from the sample $\xi_{1\dots n}$. With the notations of Sect. 7.1, the weights Z_i would be written $1-W_i$. Notice that Fromont's results are restricted to some particular weight vectors: i.i.d. symmetric, or exchangeable with $\sum_i Z_i = 0$ (for instance Efron's bootstrap (Efr): $W_i = 1 - Z_i$ multinomial). The first case includes the Rademacher and Gaussian complexity cases:

$$\widehat{R}_n(\mathcal{F}) = \widehat{B}_n^{(\mathrm{Rad})}(\mathcal{F}) \qquad \widehat{G}_n(\mathcal{F}) = \widehat{B}_n^{(\mathcal{N})}(\mathcal{F}) \ .$$

In order to use these complexities in penalization procedures, we would have to show that they are close to $\mathrm{pen}_{\mathrm{id,g}}(m)$ for every $m\in\mathcal{M}_n$ with large probability (see (7.4) and (7.5) in Sect. 7.2.1). Assuming that they satisfy concentration inequalities (like the ones of Fromont [**Fro07**], or those of Chap. 7), we need a comparison of their expectations. This is why, in this chapter, we are interested in the ratio

$$R_Z(\mathcal{F}) := \frac{\mathbb{E}\left[\widehat{B}_n^{(Z)}(\mathcal{F})\right]}{\mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right]} = \frac{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^n Z_i f(\xi_i)\right\}\right]}{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^n \left(\mathbb{E}\left[f(\xi_i)\right] - f(\xi_i)\right)\right\}\right]} \ . \tag{9.5}$$

If we knew exactly $R_Z(\mathcal{F})$, we would be able to use $R_Z(\mathcal{F})^{-1}\widehat{B}_n^{(Z)}(\mathcal{F})$ as a penalty, which is unbiased. Then, the resulting penalization procedure would be asymptotically optimal (and satisfy an oracle inequality with constant almost one).

However, in most cases, we only know that $R_Z(\mathcal{F})$ belongs to some interval $[a,b]$ of $(0,\infty)$. With this knowledge, the most reasonable choice of penalty is $a^{-1}\widehat{B}_n^{(Z)}(\mathcal{F})$, which ensures that the penalty is larger than $\mathrm{pen}_{\mathrm{id,g}}$ with large probability. Then, we are able to derive an oracle inequality, in which the variance term is overestimated within a factor $R_Z(\mathcal{F})a^{-1}$. If this factor is small in most cases, this does not matter. But if $b/a > 1$ (even asymptotically) and $R_Z(\mathcal{F})$ is close to b, then we have a *suboptimal* model selection procedure (in the sense of Sect. 5.2).

The aim of this chapter is to investigate whether this may happen or not, in particular in binary classification. It appears that the values taken by $R_Z(\mathcal{F})$ can differ from a factor at least two according to the class P and the distribution of the sample. This phenomenon occurs with both Rademacher complexities and Efron's boostrap complexities $\widehat{B}_n^{(\mathrm{Efr})}$, but it is highly non-asymptotic. Moreover, the classes $\mathcal{F}$ and distributions P we consider have a very "asymmetric" design. We can then conjecture that $R_Z(\mathcal{F})$ should be constant over a wide set of classes $\mathcal{F}$ and distributions P, which are the most used in practice.

Another interest of studying the ratio $R_Z(\mathcal{F})$ is that it appears in the confidence balls built in Chap. 10. In particular, we prove that we can not drop entirely the symmetry assumption on the data distribution.

The rest of the chapter is organized as follows. We first give several lower bounds on $R_Z(\mathcal{F})$ in Sect. 9.2. Then, we give upper bounds in particular frameworks in Sect. 9.3. We compare these lower and upper bounds in Sect. 9.4. Finally, the proofs are given in Sect. 9.5.

9.2. Lower bounds on $R_Z(\mathcal{F})$

9.2.1. A factor 2 between theory and practice. We first recall all the bounds used by Fromont [**Fro07**] in order to build bootstrap penalties with several weights.

When $Z_1, \ldots, Z_n$ are i.i.d. symmetric (*e.g.* Rademacher or Gaussian as in $\widehat{R}_n(\mathcal{F})$ and $\widehat{G}_n(\mathcal{F})$), a classical symmetrization tool shows that

$$R_Z(\mathcal{F}) \geq \frac{\mathbb{E}\,|Z_1|}{2} = \mathbb{E}\,(\,Z_1\,)_+ \tag{9.6}$$

(see for instance Lemma 1 in [**Fro07**] and the references therein). In the Rademacher case, the right-hand side is equal to $1/2$, so that Rademacher penalties should be equal to $2\widehat{R}_n(\mathcal{F})$.

When the weights $(Z_i)_{1\leq i\leq n}$ are exchangeable and $\sum_i Z_i = 0$ a.s., another lower bound comes from the proof of Prop. 2 of [**Fro07**]:

$$R_Z(\mathcal{F}) \geq \mathbb{E}\,(\,Z_1\,)_+ \quad . \tag{9.7}$$

In the bootstrap case, $(1 - Z_i)_{1\leq i\leq n} \sim \mathcal{M}(n; n^{-1}, \ldots, n^{-1})$ so that $1 - Z_1$ is binomial with parameters (n, n^{-1}). The right-hand side is then equal to $(1 - n^{-1})^n \sim_{n\to\infty} e^{-1}$, so that Efron's bootstrap penalties should be equal to $e\widehat{B}_n^{(\mathrm{Efr})}(\mathcal{F})$.

However, in simulation experiments, Lozano [**Loz00**] and Fromont [**Fro07**] prefer to use $\widehat{R}_n(\mathcal{F})$ as Rademacher penalty, and $\widehat{B}_n^{(\mathrm{Efr})}(\mathcal{F})$ as Efron's (global) boostrap penalty. In addition, Fromont gives two arguments in favour of these constants (in Rk. 3, following Thm. 3 of [**Fro07**]): an experimental computation of the ratio $R_Z(\mathcal{F})$, and the asymptotic results of Giné and Zinn [**GZ90**] and Præstgaard and Wellner [**PW93**] which suggest that

$$R_Z(\mathcal{F}) \approx \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^2 \right)^{1/2} . \tag{9.8}$$

In both Rademacher and Efron's bootstrap cases, the right-hand side of (9.8) is close to 1.

There is thus a gap between the theoretical lower bounds (9.6) and (9.7) and the experimental values of $R_Z(\mathcal{F})$, which are within a factor 2 or more. On the one hand, using the theoretical bounds will often lead to overpenalization and suboptimal model selection. On the other hand, using the experimental values may lead to a strong overfitting if the theoretical bounds happen to be sharp. The next subsection shows that accurate theoretical bounds can be obtained with some assumptions on P and $\mathcal{F}$.

9.2.2. Tight theoretical bounds. In Chap. 10, we are interested in quantities of the form

$$\mathbb{E}\left[\phi\left(\frac{1}{n} \sum_{i=1}^{n} Z_i \mathbf{Y}^i \right) \,\middle|\, \xi_{1\ldots n} \right] \qquad \text{for estimating} \qquad \phi\left(\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{Y}^i - \mathbb{E}\left[\mathbf{Y}^i \right] \right) \right)$$

with $\mathbf{Y}^1, \ldots, \mathbf{Y}^n \in \mathbb{R}^K$, $\phi : \mathbb{R}^K \mapsto \mathbb{R}$ satisfying some properties and $Z \in \mathbb{R}^n$ independent from $\mathbf{Y}$. For instance, we can take $\phi : x \in \mathbb{R}^K \mapsto \sup_k \{-x_k\}$ in the results of Sect. 10.2.2 (Prop. 10.2 and 10.3) to obtain bounds on $R_Z(\Pi)$ with

$$\Pi = \left\{ \pi_k : x \in R^K \mapsto x_k \quad ; \quad 1 \leq k \leq K \right\} \qquad \text{and} \qquad \xi_i = \mathbf{Y}^i \ .$$

Notice that in Chap. 10, we use these bounds for building confidence balls. This is why we really need a tight estimate of $R_Z(\Pi)$. With the results of Sect. 9.2.1, we would obtain much less powerful confidence regions for each given confidence level.

Gaussian classes. We now extend these results to more general families $\mathcal{F}$. First, if $(f(\xi))_{f\in\mathcal{F}}$ is a gaussian process such that $\sup_{f\in\mathcal{F}}\{f(\xi)-\mathbb{E}[f(\xi)]\}$ is measurable and has finite expectation, then for any random $Z\in\mathbb{R}^n$ independent from $\xi_{1\dots n}$,

$$R_Z(\mathcal{F})=\mathbb{E}\left[\sqrt{\frac{1}{n}\sum_{i=1}^n Z_i^2}\right] \quad . \tag{9.9}$$

When $\mathcal{F}$ is finite, this is exactly Prop. 10.2. Otherwise, it follows from the caracterization of gaussian processes by their covariance structure.

Equation (9.9) thus gives an example in which (9.8) is attained. Of course, this assumption never holds in classification where $f(\xi)\in\{0,1\}$ (with the 0-1 loss) or at least bounded. Its main interest is when $\xi\in\mathbb{R}^K$ is gaussian and $\mathcal{F}$ is a class of linear forms.

Symmetric classes. Secondly, following Prop. 10.3, we obtain tight bounds on $R_Z(\mathcal{F})$ under a symmetry assumption on $\mathcal{F}$.

PROPOSITION 9.1. *Let $\xi_1,\dots,\xi_n$ be i.i.d. random variables with values in some measurable space Ξ and $\mathcal{F}$ a class of measurable functions $\Xi\mapsto\mathbb{R}$ such that*

$$(f(\xi_1)-\mathbb{E}[f(\xi_1)])_{f\in\mathcal{F}}\overset{(d)}{=}(\mathbb{E}[f(\xi_1)]-f(\xi_1))_{f\in\mathcal{F}} \quad . \tag{9.10}$$

Let $Z\in\mathbb{R}^n$ be some random vector independent from $\xi_{1\dots n}$, such that $\sum_i Z_i=0$ a.s. Then, we have the following:

(i)

$$R_Z(\mathcal{F})\geq\frac{1}{n}\sum_{i=1}^n\mathbb{E}|Z_i| \quad . \tag{9.11}$$

(ii) *if moreover $|Z_i-x_0(Z_1,\dots,Z_n)|=a$ a.s. for some $a\in\mathbb{R}$ and some measurable $x_0:\mathbb{R}^n\mapsto\mathbb{R}$, then*

$$R_Z(\mathcal{F})\leq a+\mathbb{E}|x_0(Z_1,\dots,Z_n)| \quad . \tag{9.12}$$

REMARK 9.1.

(1) When the weights are exchangeable, (9.11) improves (9.7), since the lower bound $\mathbb{E}(Z_1)_+$ is replaced by $\mathbb{E}|Z_1|=2\mathbb{E}(Z_1)_+$.

(2) The assumption that $\sum_i Z_i=0$ forbids the use of i.i.d. Rademacher or gaussian weights. However, we can use Prop. 9.1 with "centered Rademacher weights" (centRad) $Z_i=\epsilon_i-n^{-1}\sum_{j=1}^n\epsilon_j$ (for some i.i.d. Rademacher variables $\epsilon_{1\dots n}$ independent from $\xi_{1\dots n}$). We denote by $\widehat{R}'_n(\mathcal{F}):=\widehat{B}_n^{(\mathrm{centRad})}(\mathcal{F})$ the resulting complexity measure. Notice that (9.12) can be applied with $x_0=n^{-1}\sum_i Z_i$ and $a=1$, showing that (9.11) is sharp:

$$1+n^{-1/2}\geq R_{\mathrm{centRad}}(\mathcal{F})\geq 1-n^{-1/2} \quad .$$

This is also the case with Random hold-out weights (see Sect. 10.2).

(3) When for every $f\in\mathcal{F}$, $f(\xi)\in[0,1]$ a.s., the centered Rademacher complexity can be compared to the classical Rademacher complexity (see Sect. 9.5 for a complete proof):

$$\left|\mathbb{E}\left[\widehat{R}'_n(\mathcal{F})\right]-\mathbb{E}\left[\widehat{R}_n(\mathcal{F})\right]\right|\leq\frac{1}{4\sqrt{n}} \quad . \tag{9.13}$$

When $\mathcal{F}$ is the loss class of a model with VC-dimension V, global complexities are of order $\sqrt{V/n}$ so that the remainder terms in (9.13) can be neglected. In particular, $R_{\mathrm{Rad}}(\mathcal{F}) \approx 1$, which improves on the classical bound (9.6).

However, the symmetry assumption is quite strong. In the classification case, each variable $f(\xi)$ belongs to $\{0,1\}$ a.s., so that it is binomial with parameter $P(f)$. Then, (9.10) is equivalent to

$$P(f) \in \left\{0, \frac{1}{2}, 1\right\} \qquad \text{for every} \quad f \in \mathcal{F} \ .$$

In other words, this can only be applied to classes of perfectly true, perfectly wrong and uninformative classifiers. In particular, when the classification problem is ill-posed, any classifier satisfies $P(f) = 1/2$ and the symmetry assumption holds.

Binary classification. As a conclusion to this first section, let us focus on the binary classification framework, with the 0-1 loss. In Sect. 9.2.1, we recalled theoretical lower bounds on $R_Z(\mathcal{F})$ that can be applied for most model selection problems. They have the drawback of being twice smaller than experimental and asymptotic estimates of $R_Z(\mathcal{F})$. In Sect. 9.2.2, we gave tight bounds that avoid this factor 2, but the first one can't be applied to binary classification, and the second one is limited to a toy framework.

However, even if Prop. 9.1 is restricted to very particular classes $\mathcal{F}$, it gives theoretical evidence of the drawbacks of the bounds of Sect. 9.2.1, even non-asymptotically. It is then tempting to conjecture that this factor 2 is never necessary, at least when $f(\xi) \in \{0,1\}$ a.s. In Sect. 9.3, we show that this conjecture would be wrong.

9.3. Upper bounds on $R_Z(\mathcal{F})$

In view of Prop. 9.1 above, classes such that (9.6) and (9.7) are tight have to be asymmetric. We start with a very simple case, when $\mathcal{F}$ is reduced to two opposite points.

9.3.1. Two points classes.

PROPOSITION 9.2. *Let $\mathcal{F} = \{f_0, 1 - f_0\}$ such that $f_0(\xi)$ is a Bernoulli variable with parameter $p \in (0,1)$. Let $Z \in \mathbb{R}^n$ a random vector independent from $\xi_{1\ldots n}$ such that $\sum_i Z_i = 0$.*

Then, for every $n \in \mathbb{N}$,

$$(1-np)_+ \frac{1}{2n} \sum_{i=1}^{n} \mathbb{E}\,|Z_i| \leq R_Z(\mathcal{F}) \leq (1 + n(n-1)p)\, \frac{1}{2n} \sum_{i=1}^{n} \mathbb{E}\,|Z_i| \quad . \tag{9.14}$$

As a consequence, when p goes to zero and n is fixed,

$$R_Z(\mathcal{F}) \sim \frac{1}{2n} \sum_{i=1}^{n} \mathbb{E}\,|Z_i| \quad . \tag{9.15}$$

REMARK 9.2. The assumption $\sum_i Z_i = 0$ is unnecessary if one considers the class $\mathcal{F} = \{f_0\}$ and complexities of the form $\sup_{f\in\mathcal{F}} |\cdot|$. We can then apply Prop. 9.2 to "symmetric" Rademacher complexities

$$\widehat{R}_n^{\mathrm{sym}}(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(\xi_i) \right| \,\middle|\, \xi_{1\ldots n} \right]$$

and the following "symmetric" ideal global complexity

$$\widehat{C}_n^{\mathrm{sym}}(\mathcal{F}) := \sup_{f\in\mathcal{F}} |(P - P_n)(f)| \quad .$$

Although the example of Prop. 9.2 seems somehow artifical, (9.14) gives an example where (9.7) is sharp. Its main consequence is that the bounds of Sect. 9.2.2 can not be generalized without additional assumptions on $\mathcal{F}$ and P.

There may seem to be a contradiction between (9.15) and the asymptotic comparison (9.8). Actually, this is because the phenomenon of Prop. 9.2 is highly non-asymptotic. For a given sample size n, it only occurs when $0 < P(f_0) \ll n^{-2}$. Moreover, an experimental study showed that $R_Z(\mathcal{F})$ is close to its asymptotic value when p is not too close to 0 or 1, and the "minimal distance to 0" seems to be of order n^{-1}.

9.3.2. Independent classes. The example of the previous section is somehow disappointing, since it is limited to very small classes. We now extend it to some finite classes, under an independency assumption. The main example to have in mind here is a particular case of Chap. 10: $\Xi = \{0,1\}^K$ and $\mathcal{F}$ is the set Π of coordinates projections.

PROPOSITION 9.3. *Let $\mathcal{F} = \{ f, 1-f \text{ s.t. } f \in \mathcal{F}_1 \}$ such that $\mathcal{F}_1$ is finite and $(f(\xi))_{f\in\mathcal{F}_1}$ are independent Bernoulli variable with parameters $p_f \in (0,1/2]$. Let $Z \in \mathbb{R}^n$ a random vector independent from $\xi_{1\dots n}$ such that $\sum_i Z_i = 0$.*

Then, for every $n \in \mathbb{N}$, if $\sum_f p_f \leq (4n)^{-1}$,

$$\frac{\sum_f p_f - n\left(\sum_f p_f\right)^2}{\sup_f p_f + \sum_f p_f + n^2\left(\sum_f p_f\right)^2} \leq \frac{R_Z(\mathcal{F})}{n^{-1}\sum_{i=1}^n \mathbb{E}\,|Z_i|} \leq \frac{\sum_f p_f + n^2\left(\sum_f p_f\right)^2}{\sup_f p_f + \sum_f p_f - 3n\left(\sum_f p_f\right)^2} \ . \tag{9.16}$$

As a consequence, when $\sum_f p_f$ goes to zero and n is fixed,

$$R_Z(\mathcal{F}) \sim \frac{\sum_{f\in\mathcal{F}_1} p_f}{\sup_{f\in\mathcal{F}_1}(p_f) + \sum_{f\in\mathcal{F}_1} p_f} \times \frac{1}{n}\sum_{i=1}^n \mathbb{E}\,|Z_i| \ . \tag{9.17}$$

REMARK 9.3. The proof of Prop. 9.3 can be extended to functions $\phi : \mathbb{R}^{\mathrm{Card}(\mathcal{F}_1)} \mapsto \mathbb{R}$ more general than $\sup|\cdot|$, under assumptions made in Chap. 10. Moreover, like Prop. 9.2, we can drop the assumption $\sum_i Z_i = 0$ by considering directly complexities of the form $\sup_{f\in\mathcal{F}_1}|\cdot|$ and the class $\mathcal{F}_1$.
Notice that when $\mathrm{Card}(\mathcal{F}_1) = 1$, we recover Prop. 9.2 (up to some small enlargement of the remainder terms).

We now give some comments on the consequences of Prop. 9.3:

(1) Taking one of the $(p_f)_{f\in\mathcal{F}_1}$ equal to p, and the other ones negligible in front of p, we obtain a larger family of examples for which (9.7) is sharp. In particular, contrary to Prop. 9.2, we no longer assume that $\mathrm{Card}(\mathcal{F}) = 2$, but only that it is finite.

(2) In classification, the independency assumption is valid when $\mathcal{F}_1 = \{ f_0 \}$, or in some particular cases, *e.g.* $\mathcal{X} = [0,1]^K$, X uniform on $\mathcal{X}$, $Y \equiv 0$ and

$$\mathcal{F}_1 = \{ f_k : x \mapsto \mathbb{1}_{x_k \leq p} \text{ s.t. } 1 \leq k \leq K \} \ .$$

Thus, (9.11) can not be generalized to all classes of any finite cardinality.

(3) When all the p_f are equal to some $p \in (0,1/2]$, the limit value of $R_Z(\mathcal{F})$ when $p \to 0$ in (9.17) is $\mathrm{Card}(\mathcal{F})/(1+\mathrm{Card}(\mathcal{F}))$. There is thus some hope that the $1/2$ factor is unnecessary when $\mathrm{Card}(\mathcal{F})$ is large enough and the variables $f(\xi)$ not too asymmetric.

(4) As in Prop. 9.2, this counter-example is highly non-asymptotic, since we have to assume that $\sum_f p_f \ll n^{-2}$.

9.3.3. Rademacher complexities. In Prop. 9.2 and 9.3, we assume that $\sum_i Z_i = 0$, which forbids the use of Rademacher weights. We here consider this particular case, for which the behavior of $R_{\mathrm{Rad}}(\mathcal{F})$ when $p \to 0$ is quite different.

PROPOSITION 9.4. *Let $\mathcal{F} = \{ f_0, 1 - f_0 \}$ such that $f_0(\xi)$ is a Bernoulli variable with parameter $p \in (0,1)$. Let $Z \in \mathbb{R}^n$ be a random vector independent from $\xi_{1\dots n}$ such that $\mathbb{E}\left(\sum_i Z_i\right)_+ > 0$ (for instance, i.i.d. Rademacher variables).*

Then, for every $n \in \mathbb{N}$,

$$R_Z(\mathcal{F}) \geq \mathbb{E}\left(\sum_{i=1}^n Z_i\right)_+ \frac{1-np}{2np} - \frac{\sum_{i=1}^n \mathbb{E}\,|Z_i|}{2} \quad . \tag{9.18}$$

As a consequence, when p goes to zero and n is fixed,

$$R_Z(\mathcal{F}) \longrightarrow +\infty \quad . \tag{9.19}$$

REMARK 9.4. A similar phenomenon occurs with the framework of Prop. 9.3, where $\mathrm{Card}(\mathcal{F})$ can take any finite value.
When $\mathbb{E}\left(\sum_i Z_i\right)_+ = 0$ but $\mathbb{E}\left(\sum_i Z_i\right)_- > 0$, we have the same behaviour when $p \to 1$.

The behaviour of Rademacher complexities with highly asymmetric classes is thus very different from Efron's bootstrap penalties, or centered Rademacher complexities. This comes from the fact that $\sum_i Z_i$ is not a.s. equal to zero: the Rademacher complexity stays away from zero, whereas the true complexity $\widehat{C}_n(\mathcal{F}) \sim 2p$ goes to zero.

9.4. Discussion

In the binary classification framework, we have thus found instances of classes $\mathcal{F}$ and distributions P for which $R_Z(\mathcal{F})$ is very far from its asymptotic value. When $\sum_i Z_i = 0$ a.s. (for instance Efron's bootstrap and "centered Rademacher" complexities), the ratio R_Z vary within a factor at least two, between symmetric and asymmetric families. As a consequence, the classical bound (9.7) is unimprovable, but not always tight, even non-asymptotically. On the other hand, when $\mathbb{P}(\sum_i Z_i \neq 0) > 0$, the ratio R_Z can be infinitely larger with asymmetric families than with symmetric ones.

The first consequence of these facts is that for every fixed sample size n, no global resampling complexity $\widehat{B}_n^{(Z)}$ is proportional in expectation to the complexity $\widehat{C}_n$ over all classes $\mathcal{F}$ and all distributions P (even if we restrict to binary classification and 0-1 loss classes).

However, from the practical viewpoint, it appears that R_{Efr} and R_{Rad} is (almost) always close to 1. In addition, this always holds asymptotically. This is why we can hope to generalize (9.11) to classes $\mathcal{F}$ and distributions P satisfying mild assumptions.

In view of Sect. 9.3, we can conjecture that a limitation of the "asymmetry" of the class (*e.g.* $\sum_f p_f \geq n^{-1}$) may be sufficient. In classification, this means that we would assume that the classifiers in $\mathcal{F}$ are not uniformly good, which may be unintuitive. Since such classes have a uniformly small risk, underestimating their complexity can not have serious consequences on the risk. Thus, a proof of (9.11) under this mild "non-asymmetry assumption" would be of great interest.

On the contrary, Rademacher complexities overestimate the complexities of these "too good" classes, which may enlarge the risk of a model selection procedure based upon them. Added to the fact that global penalties can be much larger than the ideal one, this may be a serious limitation of global Rademacher penalties.

Alternatively, if the ratio $R_Z(\mathcal{F}_m)$ is unknown but almost constant over the family of models $m \in \mathcal{M}_n$, we can try to estimate it from the data. This can be done for instance with the "slope heuristics" (algorithm 3.1 in Sect. 3.4). But if the ratio $R_Z(\mathcal{F}_m)$ is strongly dependent from m, the bootstrap complexity $\widehat{B}_n^{(Z)}(\mathcal{F}_m)$ does not even estimate the shape of $\widehat{C}_n(\mathcal{F}_m)$, and this method will not work.

In this chapter, we focused on global penalties, as opposed to *local penalties*, which estimate $(P - P_n)(\widehat{s}_m)$. For instance, in Chap. 5, 6 and 7, we proposed resampling penalties of the form

$$\mathbb{E}\left[(P_n - P_n^W)(\widehat{s}_m^W) \,\middle|\, \xi_{1\dots n} \right] \ .$$

It would be interesting to determine whether these local penalties have the same drawback as global resampling penalties.

Finally, in the confidence region and multiple testing framework of Chap. 10, our results have more straightforward consequences. Indeed, we have shown that the concentration thresholds defined in Sect. 10.2 can not be used with any bounded data. However, the symmetry assumption (SA) is probably too restrictive. We could hope to generalize (9.11) when the sample size is larger than some n_0, which may quantify the "asymmetry" of the sample.

9.5. Proofs

PROOF OF PROP. 9.1. (i) Since $Z_1, \dots, Z_n$ is independent from $\xi_{1\dots n}$, for every sample $\xi_{1\dots n}$,

$$\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}\,|Z_i| \right) \widehat{C}_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n |Z_i| \,(P(f) - f(\xi_i)) \,\middle|\, \xi_{1\dots n} \right] \right\} \ .$$

Then, using Jensen inequality and integrating w.r.t. $\xi_{1\dots n}$,

$$\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}\,|Z_i| \right) \mathbb{E}\left[\widehat{C}_n(\mathcal{F}) \right] \le \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n |Z_i| \,(P(f) - f(\xi_i)) \right\} \right] \ .$$

We now use the symmetry assumption (9.10) and the independence of the ξ_i to derive

$$\begin{aligned}
\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}\,|Z_i| \right) \mathbb{E}\left[\widehat{C}_n(\mathcal{F}) \right] &\le \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n Z_i \,(f(\xi_i) - P(f)) \right\} \right] \\
&\le \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n Z_i f(\xi_i) \right\} + \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n Z_i \,(-P(f)) \right\} \right] \\
&= \widehat{B}_n^{(Z)}(\mathcal{F}) + \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n Z_i \right)_+ \right] \sup_{f \in \mathcal{F}} \{-P(f)\} \\
&\quad + \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n Z_i \right)_- \right] \sup_{f \in \mathcal{F}} \{P(f)\} \ .
\end{aligned}$$

When $\sum_i Z_i = 0$ a.s., this upper bound is equal to $\widehat{B}_n^{(Z)}$ and the result follows.

(ii) comes from similar arguments and the additive assumption:

$$\widehat{B}_n^{(Z)}(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n Z_i \,(f(\xi_i) - P(f)) \right\} \right]$$

$$\text{so that} \qquad \widehat{B}_n^{(Z)}(\mathcal{F}) \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n (Z_i - x_0(Z_{1\dots n}))\,(f(\xi_i)-P(f))\right\}\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n x_0(Z_{1\dots n})\,(f(\xi_i)-P(f))\right\}\right] = (a+\mathbb{E}\,|x_0(Z_{1\dots n})|)\,\widehat{C}_n(\mathcal{F})\ .$$

□

PROOF OF (9.13). Define $\overline{\epsilon} = n^{-1}\sum_i \epsilon_i$. Then,

$$\mathbb{E}\left[\widehat{R}'_n(\mathcal{F})\right] := \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n (\epsilon_i-\overline{\epsilon})\,f(\xi_i)\right\}\right] \le \mathbb{E}\left[\widehat{R}_n(\mathcal{F})\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\{-\overline{\epsilon}P_n(f)\}\right]$$
$$= \mathbb{E}\left[\widehat{R}_n(\mathcal{F})\right] + \mathbb{E}\left[(\overline{\epsilon})_-\sup_{f\in\mathcal{F}}P_n(f)\right] \le \widehat{R}_n(\mathcal{F}) + \mathbb{E}\,(\overline{\epsilon})_-$$

and similarly

$$\mathbb{E}\left[\widehat{R}'_n(\mathcal{F})\right] \ge \mathbb{E}\left[\widehat{R}_n(\mathcal{F})\right] - \mathbb{E}\,(\overline{\epsilon})_-\ .$$

Notice that the same reasoning can be applied to any Z. Then, using that the ϵ_i are symmetric and i.i.d., we have

$$\mathbb{E}\,(\overline{\epsilon})_- = \frac{1}{2}\mathbb{E}\,|\overline{\epsilon}| \le \frac{1}{2}\sqrt{\mathbb{E}\overline{\epsilon}^2} = \frac{1}{4\sqrt{n}}\ .$$

□

PROOF OF PROP. 9.2. Let B be some binomial variable with parameters (n,p). Because of the particular form of $\mathcal{F}$, we have

$$\mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right] = \mathbb{E}\,|(P-P_n)(f_0)| = \frac{\mathbb{E}\,|np-B|}{n} = \frac{2}{n}\mathbb{E}\,[(np-B)\,\mathbb{1}_{B\le np}]$$
$$\ge 2p\mathbb{P}(B=0) = 2p(1-p)^n \ge 2p(1-pn)_+\ .$$

When moreover $np\le 1$, the first inequality is an equality so that

$$2p(1-pn) \le \mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right] = 2p(1-p)^n \le 2p\ . \tag{9.20}$$

On the other hand,

$$\mathbb{E}\left[\widehat{B}_n^{(Z)}(\mathcal{F})\right] = \mathbb{E}\left[\max\left\{\frac{1}{n}\sum_{i=1}^n Z_i f_0(\xi_i); \frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n Z_i f_0(\xi_i)\right\}\right] = \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_i f_0(\xi_i)\right|$$
$$= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n Z_i f_0(\xi_i)\right|\mathbb{1}_{\sum_i f_0(\xi_i)=1}\right] + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n Z_i f_0(\xi_i)\right|\mathbb{1}_{\sum_i f_0(\xi_i)\ge 2}\right]$$
$$\le \mathbb{P}(B=1)\frac{1}{n}\sum_{i=1}^n\left|\frac{Z_i}{n}\right| + \mathbb{P}(B\ge 2)\frac{1}{n}\sum_{i=1}^n\mathbb{E}\,|Z_i|$$
$$\le \left[p(1-p)^{n-1} + \left(1-(1-p)^n - np(1-p)^{n-1}\right)\right]\frac{1}{n}\sum_{i=1}^n\mathbb{E}\,|Z_i|$$
$$\le \left[p + n(n-1)p^2\right]\frac{1}{n}\sum_{i=1}^n\mathbb{E}\,|Z_i|$$

and with the same decomposition,

$$\mathbb{E}\left[\widehat{B}_n^{(Z)}(\mathcal{F})\right] \geq \mathbb{P}(B=1)\frac{1}{n}\sum_{i=1}^n\left|\frac{Z_i}{n}\right| = p(1-p)^{n-1}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|Z_i\right| \geq p(1-(n-1)p)\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|Z_i\right| \ .$$

Combining this with (9.20), we derive

$$(1-np)_+\,\frac{1}{2n}\sum_{i=1}^n\mathbb{E}\left|Z_i\right| \leq R_Z(\mathcal{F}) \leq (1+n(n-1)p)\,\frac{1}{2n}\sum_{i=1}^n\mathbb{E}\left|Z_i\right| \ .$$

□

PROOF OF PROP. 9.3. For $\widehat{C}_n(\mathcal{F})$, we start as in the proof of Prop. 9.2:

$$\mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|(P-P_n)(f)|\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|np_f - B_f|\right]$$

where $(B_f)_{f\in\mathcal{F}_1}$ are independent binomial variables with parameters $(n,p_f)_{f\in\mathcal{F}_1}$. Hence, using that $sup_{f\in\mathcal{F}_1}p_f \leq 1/(2n)$,

$$\begin{aligned}
\mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right] &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|np_f|\,\mathbb{1}_{\forall f,\,B_f=0}\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|np_f - B_f|\,\mathbb{1}_{\sum_f B_f=1}\right] \\
&\qquad + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|np_f - B_f|\,\mathbb{1}_{\sum_f B_f\geq 2}\right] \\
&= \left(\sup_f p_f\right)\mathbb{P}\left(\forall f,\,B_f=0\right) + \frac{1}{n}\sum_f\left[(1-np_f)\mathbb{P}\left(B_f=1 \text{ and } \forall f'\neq f,\,B_{f'}=0\right)\right] \\
&\qquad + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|np_f - B_f|\,\mathbb{1}_{\sum_f B_f\geq 2}\right] \\
&= \sup_f p_f\prod_f(1-p_f)^n + \sum_f\left[p_f(1-np_f)(1-p_f)^{n-1}\prod_{f'\neq f}\left(1-p_{f'}\right)^n\right] \\
&\qquad + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}|np_f - B_f|\,\mathbb{1}_{\sum_f B_f\geq 2}\right] \ .
\end{aligned}$$

Using repeatedly that for every $a,b\geq 0$, $(1-a)(1-b)\geq 1-a-b$, we have

$$\prod_f(1-p_f)^n \geq 1-n\sum_f p_f$$

and for every $f\in\mathcal{F}_1$,

$$(1-np_f)(1-p_f)^{n-1}\prod_{f'\neq f}\left(1-p_{f'}\right)^n \geq 1-(n-1)p_f - n\sum_{f'}p_{f'} \ .$$

As a consequence,

$$\begin{aligned}
\mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right] &\geq \sup_f p_f\left(1-n\sum_f p_f\right) + \sum_f\left[p_f\left(1-(n-1)p_f - n\sum_{f'}p_{f'}\right)\right] \\
&\geq \sup_f p_f + \sum_f p_f - n\sum_f p_f\sup_f p_f - (n-1)\sum_f p_f^2 - n\left(\sum_f p_f\right)^2 \qquad (9.21)
\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\widehat{C}_n(\mathcal{F})\right] &\leq \sup_f p_f + \sum_f p_f + \mathbb{P}\left(\sum_f B_f \geq 2\right) \\ &\leq \sup_f p_f + \sum_f p_f + 1 - \left(1 - n\sum_f p_f\right) - n\sum_f \left[p_f\left(1-(n-1)p_f - n\sum_{f'\neq f} p_{f'}\right)\right] \\ &\leq \sup_f p_f + \sum_f p_f + n^2\left(\sum_f p_f\right)^2 \quad . \end{aligned} \tag{9.22}$$

We now focus on $\widehat{B}_n(\mathcal{F})$. As in the proof of Prop. 9.2,

$$\begin{aligned}\mathbb{E}\left[\widehat{B}_n(\mathcal{F})\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{1}{n}\sum_{i=1}^n Z_i f(\xi_i)\right|\right] \\ &= \sum_{i=1}^n \sum_{f\in\mathcal{F}_1} \mathbb{E}\frac{|Z_i|}{n}\mathbb{P}\left(f(\xi_i)=1 \text{ and } \forall (j,f')\neq(i,f),\ f'(\xi_j)=0\right) \\ &\qquad + \mathbb{E}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{1}{n}\sum_{i=1}^n Z_i f(\xi_i)\right| \mathbb{1}_{\sum_f\sum_i f(\xi_i)\geq 2}\right] \\ &= \sum_{i=1}^n \sum_{f\in\mathcal{F}_1} \mathbb{E}\frac{|Z_i|}{n} p_f(1-p_f)^{n-1}\prod_{f'\neq f}(1-p_{f'})^n \\ &\qquad + \mathbb{E}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{1}{n}\sum_{i=1}^n Z_i f(\xi_i)\right| \mathbb{1}_{\sum_f\sum_i f(\xi_i)\geq 2}\right] \quad . \end{aligned}$$

Using the same inequalities as for $\widehat{C}_n$, we obtain

$$\mathbb{E}\left[\widehat{B}_n(\mathcal{F})\right] \geq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\,|Z_i|\right)\left[\sum_f p_f - n\left(\sum_f p_f\right)^2\right] \tag{9.23}$$

and

$$\begin{aligned}\mathbb{E}\left[\widehat{B}_n(\mathcal{F})\right] &\leq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\,|Z_i|\right)\left[\sum_f p_f + \mathbb{P}\left(\sum_f\sum_i f(\xi_i)\geq 2\right)\right] \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\,|Z_i|\right)\left[\sum_f p_f + n^2\left(\sum_f p_f\right)^2\right] \quad . \end{aligned} \tag{9.24}$$

We now combine (9.21) with (9.24) to obtain

$$R_Z(\mathcal{F}) \leq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\,|Z_i|\right) \times \frac{\sum_f p_f + n^2\left(\sum_f p_f\right)^2}{\sup_f p_f + \sum_f p_f - 3n\left(\sum_f p_f\right)^2}$$

and (9.22) with (9.23) gives

$$R_Z(\mathcal{F}) \geq \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\,|Z_i| \right) \times \frac{\sum_f p_f - n \left(\sum_f p_f \right)^2}{\sup_f p_f + \sum_f p_f + n^2 \left(\sum_f p_f \right)^2} \ .$$

□

PROOF OF PROP. 9.4. From (9.20) in the proof of Prop. 9.2, we know that if $np \leq 1$,

$$2p(1-pn) \leq \mathbb{E}\left[\widehat{C}_n(\mathcal{F}) \right] = 2p(1-p)^n \leq 2p \ .$$

We now consider the Bootstrap complexity. Define $\overline{Z} = n^{-1} \sum_i Z_i$.

$$\begin{aligned}
\mathbb{E}\left[\widehat{B}_n^{(Z)}(\mathcal{F}) \right] &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i f(\xi_i) \right\} \right] \\
&= \mathbb{E}\left[\max \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i f_0(\xi_i) \, ; \, \overline{Z} - \frac{1}{n} \sum_{i=1}^{n} Z_i f_0(\xi_i) \right\} \right] \\
&= \mathbb{E}\left(\overline{Z} \right)_+ \mathbb{P}\left(\forall i,\, f_0(\xi_i) = 0 \right) \\
&\qquad + \mathbb{E}\left[\max \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i f_0(\xi_i) \, ; \, \overline{Z} - \frac{1}{n} \sum_{i=1}^{n} Z_i f_0(\xi_i) \right\} \mathbb{1}_{\exists i,\, f_0(\xi_i) \neq 0} \right] \\
&\geq \mathbb{E}\left(\overline{Z} \right)_+ (1-np)_+ - p \sum_{i=1}^{n} \mathbb{E}\,|Z_i| \ .
\end{aligned}$$

□

CHAPTER 10

Resampling-based confidence regions and multiple tests

This chapter is a joint work with Gilles Blanchard[1] and Étienne Roquain[2]. A short version of it has been published in the Proceedings of COLT'07 [**ABR07**].

RÉSUMÉ. Ce chapitre est consacré à l'étude de régions de confiance par rééchantillonnage pour la moyenne d'un vecteur aléatoire, dont les coordonnées ont une structure de dépendance inconnue. La dimension de ce vecteur peut être bien plus grande que le nombre d'observations, et nous cherchons un contrôle non-asymptotique du niveau de confiance. Le vecteur aléatoire est supposé soit gaussien, soit symétrique et borné. Nous considérons deux approches, la première fondée sur des inégalités de concentration, la seconde sur l'estimation directe de quantiles par rééchantillonnage. Dans la première, nous considérons une très grande famille de poids de rééchantillonnage, alors que les résultats de la seconde sont limités aux poids Rademacher. Ces résultats sont appliqués ensuite au problème de test multiple unilatéral ou bilatéral, pour lequel nous obtenons plusieurs procédures step-down par rééchantillonnage d'où découle un contrôle du Family-Wise Error Rate. Nous comparons ces différentes procédures dans une étude de simulation, et nous montrons qu'elles peuvent s'avérer meilleures que les méthodes de Bonferroni ou de Holm dès lors que le vecteur observé a des coordonnées suffisamment corrélées.

ABSTRACT. We study generalized bootstrapped confidence regions for the mean of a random vector whose coordinates have an unknown dependence structure. The dimensionality of the vector can possibly be much larger than the number of observations and we focus on a non-asymptotic control of the confidence level. The random vector is supposed to be either Gaussian or to have a symmetric bounded distribution. We consider two approaches, the first based on a concentration principle and the second on a direct bootstrapped quantile. The first one allows us to deal with a very large class of resampling weights while our results for the second are specific to Rademacher weights. We present an application of these results to the one-sided and two-sided multiple testing problem, in which we derive several resampling-based step-down procedures providing a non-asymptotic FWER control. We compare our different procedures in a simulation study, and we show that they can outperform Bonferroni's or Holm's procedures as soon as the observed vector has sufficiently correlated coordinates.

[1]Fraunhofer FIRST.IDA, Berlin, Germany.
[2]INRA Jouy-en-Josas, unité MIG, Jouy-en-Josas, France

10.1. Introduction

10.1.1. Goals and motivations. In this chapter, we assume that we observe a sample $\mathbf{Y} := (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$ of $n \geq 2$ i.i.d. observations of an integrable random vector $\mathbf{Y}^i \in \mathbb{R}^K$ with a dimension K possibly much larger than n. Let $\mu \in \mathbb{R}^K$ denote the common mean of the $\mathbf{Y}^i$; our main goal is to find a non-asymptotic $(1-\alpha)$-confidence region for μ, of the form:

$$\left\{ x \in \mathbb{R}^K \text{ s.t. } \phi\left(\overline{\mathbf{Y}} - x\right) \leq t_\alpha(\mathbf{Y}) \right\} \quad , \tag{10.1}$$

where $\phi : \mathbb{R}^K \to \mathbb{R}$ is a measurable function fixed in advance by the user (measuring a kind of distance), $\alpha \in (0,1)$, $t_\alpha : \left(\mathbb{R}^K\right)^n \to \mathbb{R}$ is a measurable data-dependent threshold, and $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}^i$ is the empirical mean of the sample $\mathbf{Y}$.

The form of the confidence region (10.1) is motivated by the following multiple testing problem: when we want to test simultaneously for all $1 \leq k \leq K$ the null hypotheses $H_{0,k}$: " $\mu_k \leq 0$" against $H_{1,k}$: "$\mu_k > 0$, a classical procedure consists in rejecting the $H_{0,k}$ corresponding to

$$\left\{ 1 \leq k \leq K \text{ s.t. } \overline{\mathbf{Y}}_k > t_\alpha(\mathbf{Y}) \right\} \quad . \tag{10.2}$$

The error of such a multiple testing procedure can be measured by the Family-Wise Error Rate (FWER) defined by the probability that at least one hypothesis is wrongly rejected. Denoting by $\mathcal{H}_0 = \{ k \text{ s.t. } \mu_k \leq 0 \}$ the set of coordinates corresponding to the true null hypotheses, the FWER of the procedure defined in (10.2) can be controlled as follows:

$$\begin{aligned} \mathbb{P}\left(\exists k \text{ s.t. } \overline{\mathbf{Y}}_k > t_\alpha(\mathbf{Y}) \text{ and } \mu_k \leq 0\right) &\leq \mathbb{P}\left(\exists k \in \mathcal{H}_0 \text{ s.t. } \overline{\mathbf{Y}}_k - \mu_k > t_\alpha(\mathbf{Y})\right) \\ &= \mathbb{P}\left(\sup_{k \in \mathcal{H}_0} \left\{ \overline{\mathbf{Y}}_k - \mu_k \right\} > t_\alpha(\mathbf{Y}) \right) \quad . \end{aligned}$$

Since μ_k is unknown under $H_{0,k}$, controlling the above probability by a level α is equivalent to establish a $(1-\alpha)$-confidence region for μ of the form (10.1) with $\phi(x) = \sup_{k \in \mathcal{H}_0}(x_k)$. Similarly, the same reasoning with $\phi = \sup_{\mathcal{H}_0} |\cdot|$ in (10.1) allows us to test $H_{0,k}$: "$\mu_k = 0$" against $H_{1,k}$: "$\mu_k \neq 0$", by choosing the rejection set $\left\{ 1 \leq k \leq K \text{ s.t. } \left|\overline{\mathbf{Y}}_k\right| > t_\alpha(\mathbf{Y}) \right\}$.

In our framework, we emphasize that:

- we aim at obtaining a *non-asymptotical* result valid for any fixed K and n, with K possibly much larger than the number of observations n.
- we do not want to make any assumptions on the dependency structure of the coordinates of $\mathbf{Y}^i$ (although we will consider some general assumptions over the distribution of $\mathbf{Y}$, for example that it is Gaussian).

In the Gaussian case, a traditional parametric method based on the direct estimation of the covariance matrix to derive a confidence region would not be appropriate in the situation where $K \gg n$, unless the covariance matrix is assumed to belong to some parametric model of lower dimension, which we explicitly don't want to postulate here. In this sense our approach is closer in spirit to non-parametric or semiparametric statistics.

Our viewpoint is motivated by practical applications, especially neuroimaging (Pantazis *et al.* [**PNBL05**], Darvas *et al.* [**DRP+05**], Jerbi *et al.* [**JLN+07**]). In a typical magnetoencephalography (MEG) experiment, each observation $\mathbf{Y}^i$ is a two or three dimensional brain activity map[3] of 15 000 points, or a time series of length T of such data, $50 \leq T \leq 1\,000$. The dimensionality K thus goes from 10^4 to 10^7. Such observations are repeated $n = 15$ up to 4 000 times, but this upper bound is very hard to attain (see Waberski *et al.* [**WGK+03**] for an experiment with

[3] actually, $\mathbf{Y}^i$ is the difference between brain activities with and without some stimulation. Then, non-zero means are locations at which the stimulation has a significant effect.

$n \geq 4\,000$). Typically, $n \leq 100 \ll K$. In such data, there are strong dependencies between locations (the 15 000 points are obtained by pre-processing data of 150 sensors) which are highly spatially non-uniform, as remarked by Pantazis *et al.* [**PNBL05**]. Moreover, there may be distant correlations, *e.g.* depending on neural connections inside the brain, so that we cannot make use of a simple parametric model.

Another field of applications for this work is genomics, particularly microarray data analysis, where it is common to observe samples of limited size (*e.g.* less than 100) of a vector in high dimension (*e.g.* more than 20 000, each coordinate corresponding to a specific gene), and where the dependency structure can be quite arbitrary (Dudoit, Shaffer and Boldric [**DSB03**], Ge, Dudoit and Speed [**GDS03**]).

10.1.2. Two approaches to our goal. The ideal threshold t_α in (10.1) is obviously the $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$. However, this quantity depends on the unknown dependency structure of the coordinates of $\mathbf{Y}^i$ and is therefore itself unknown.

We propose here to approach t_α by some resampling scheme: the heuristics of the resampling method (introduced by Efron [**Efr79**], generalized to exchangeable weighted bootstrap by Mason and Newton [**MN92**] and Præstgaard and Wellner [**PW93**]) is that the distribution of $\overline{\mathbf{Y}} - \mu$ is "close" to the one of

$$\overline{\mathbf{Y}}_{[W-\overline{W}]} := \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})\mathbf{Y}^i = \frac{1}{n}\sum_{i=1}^{n} W_i(\mathbf{Y}^i - \overline{\mathbf{Y}}) = \overline{\left(\mathbf{Y} - \overline{\mathbf{Y}}\right)}_{[W]} \ ,$$

conditionally to $\mathbf{Y}$, where $(W_i)_{1\leq i\leq n}$ are real random variables independent of $\mathbf{Y}$ called the *resampling weights*, and $\overline{W} = n^{-1}\sum_{i=1}^{n} W_i$. We emphasize that the family $(W_i)_{1\leq i\leq n}$ itself *need not be independent.*

Following this general idea, we investigate two different approaches to obtain non-asymptotic confidence regions:

(1) "Concentration approach": the expectations of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ can be precisely compared, and the processes $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ concentrate well around their expectations.

(2) "Quantile approach": the $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ conditionally to $\mathbf{Y}$ is close to the one of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$.

The first approach above is closely related to the notion of Rademacher complexity in learning theory, and our results in this direction are heavily inspired by the work of Fromont [**Fro04**], who studies general resampling schemes in a learning theoretical setting. It may also be seen to some extent as a generalization of cross-validation methods. For what concerns the second approach, we will restrict ourselves specifically to Rademacher weights in our analysis, because we rely heavily on a symmetrization principle.

10.1.3. Relation to previous work. Using resampling to construct confidence regions (see *e.g.* Efron [**Efr79**], Hall [**Hal92**], Hall and Mammen [**HM94**]) or multiple testing procedures (see *e.g.* Westfall and Young [**WY93**], Yekutieli and Benjamini [**YB99**], Pollard and van der Laan [**PvdL03**], Ge, Dudoit and Speed [**GDS03**], Romano and Wolf [**RW07**]) is a vast field of study in statistics. Roughly speaking, we can mainly distinguish between two main kinds of results:

- asymptotic results, which are based on the fact that the bootstrap process is asymptotically close to the original empirical process (see van der Vaart and Wellner [**vdVW96**]).

- exact randomized tests (see *e.g.* Romano [**Rom89, Rom90**], Romano and Wolf [**RW05**]), which are based on an invariance of the null distribution under a given transformation; the underlying idea can be traced back to Fisher's permutation test (see Fisher [**Fis35**]).

Because we focus on a non-asymptotic viewpoint, the asymptotic approach mentioned above is not adapted to the goals we have fixed.

Our "concentration approach" of the previous section is not directly related to either type of the above previous results, but, as already pointed out earlier, is strongly inspired by results coming from learning theory. On the other hand, what we called our "quantile approach" in the previous section is strongly related to exact randomization tests. Namely, we will only consider symmetric distributions: this is a specific instance of an invariance with respect to a transformation and will allow us to make use of distribution-preserving randomization via sign-flipping. The main difference with traditional exact randomization tests is that, because our first goal is to derive a confidence region, the vector of the means is unknown and therefore, so is the exact invariant transformation. Our contribution to this point is essentially to show that the true vector of the means can be replaced by the empirical one in the randomization, for the price of additional terms of smaller order in the threshold thus obtained. To our knowledge, this gives the first non-asymptotic approximation result on resampled quantiles with an unknown distribution mean.

10.1.4. Notations. Let us now define a few notations that will be useful throughout this chapter.

- A boldface letter indicates a matrix. This will almost exclusively concern the $K \times n$ data matrix $\mathbf{Y}$. A superscript index such as $\mathbf{Y}^i$ indicates the i-th column of a matrix.
- If $\mu \in \mathbb{R}^K$, $\mathbf{Y} - \mu$ is the matrix obtained by subtracting μ from each (column) vector of $\mathbf{Y}$. If $c \in \mathbb{R}$ and $W \in \mathbb{R}^n$, $W - c = (W_i - c)_{1 \leq i \leq n} \in \mathbb{R}^n$.
- If X is a random variable, $\mathcal{D}(X)$ is its distribution and $\mathrm{var}(X)$ is its variance.
- We denote by $\mathbb{E}_W[\cdot]$, the expectation operator over the distribution of the weight vector W only, *i.e.*, conditional to $\mathbf{Y}$. We use a similar notation $\mathbb{P}_W$ for the corresponding probability operator and $\mathbb{E}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}}$ for the same operations conditional to W. Since $\mathbf{Y}$ and W are always assumed to be independent, the operators $\mathbb{E}_W$ and $\mathbb{E}_Y$ commute by Fubini's theorem.
- The vector $\sigma = (\sigma_k)_{1 \leq k \leq K}$ is the vector of the standard deviations of the data: $\forall k, 1 \leq k \leq K$, $\sigma_k = \mathrm{var}^{1/2}(\mathbf{Y}_k^1)$.
- $\overline{\Phi}$ is the standard Gaussian upper tail function: if $X \sim \mathcal{N}(0,1)$, $\forall x \in \mathbb{R}$, $\overline{\Phi}(x) = \mathbb{P}(X \geq x)$.

Several properties may be assumed for the function $\phi : \mathbb{R}^K \to \mathbb{R}$:

- Subadditivity: $\forall x, x' \in \mathbb{R}^K, \quad \phi(x + x') \leq \phi(x) + \phi(x')$.
- Positive-homogeneity: $\forall x \in \mathbb{R}^K, \forall \lambda \in \mathbb{R}_+, \quad \phi(\lambda x) = \lambda \phi(x)$.
- Bounded by the p-norm, $p \in [1, \infty]$: $\forall x \in \mathbb{R}^K$, $|\phi(x)| \leq \|x\|_p$, where $\|x\|_p$ is equal to $(\sum_{k=1}^K |x_k|^p)^{1/p}$ if $p < \infty$ and $\max_k\{|x_k|\}$ otherwise.

Finally, different assumptions on the generating distribution of $\mathbf{Y}$ can be made:

(GA) The Gaussian assumption: the $\mathbf{Y}^i$ are Gaussian vectors.

(SA) The symmetry assumption: the $\mathbf{Y}^i$ are symmetric with respect to μ *i.e.* $\mathbf{Y}^i - \mu \sim \mu - \mathbf{Y}^i$.

(BA)(p, M) The bounded assumption: $\left\|\mathbf{Y}^i - \mu\right\|_p \leq M$ a.s.

In this chapter, our primary focus is on the Gaussian framework (GA), because we obtain more accurate results under this assumption. In addition, we always assume that we know some upper bound on a p-norm of σ for some $p > 0$.

The chapter is organized as follows. We first build confidence regions with two different techniques: Sect. 10.2 deals with the concentration method with general weights, and Sect. 10.3 with a quantile approach with Rademacher weights. We then focus on the multiple testing problem in Sect. 10.4. Finally, we illustrate our results on both confidence regions and multiple testing in Sect. 10.5 by a simulation study. Sect. 10.6 gives discussions and concluding remarks. All the proofs are given in Sect. 10.7.

10.2. Confidence region using concentration

10.2.1. Main result. We consider in this chapter a general *resampling weight vector* W, that is, a $\mathbb{R}^n$-valued random vector $W = (W_i)_{1\le i\le n}$ independent of $\mathbf{Y}$ and satisfying the following properties: for all $i \in \{1,\dots,n\}$ $\mathbb{E}\left[W_i^2\right] < \infty$ and $n^{-1}\sum_{i=1}^n \mathbb{E}\left|W_i - \overline{W}\right| > 0$. In this section, we mainly consider an *exchangeable resampling weight vector*, that is, a resampling weight vector W such that $(W_i)_{1\le i\le n}$ has an exchangeable distribution (*i.e.* invariant under any permutation of the indices). Several examples of exchangeable resampling weight vectors are given in Sect. 10.2.4, where we also tackle the question of choosing a resampling. Non-exchangeable weight vectors are studied in Sect. 10.2.5.

Four constants that depend only on the distribution of W appear in the results below (the fourth one is defined only for a particular class of weights). They are defined as follows and computed for classical resamplings in Tab. 10.1:

$$A_W := \mathbb{E}\left|W_1 - \overline{W}\right| \tag{10.3}$$

$$B_W := \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)^2\right)^{\frac{1}{2}}\right] \tag{10.4}$$

$$C_W := \left(\frac{n}{n-1}\mathbb{E}\left[\left(W_1 - \overline{W}\right)^2\right]\right)^{\frac{1}{2}} \tag{10.5}$$

$$D_W := a + \mathbb{E}\left|\overline{W} - x_0\right| \quad \text{if } \forall i,\ |W_i - x_0| = a \text{ a.s. (with } a > 0, x_0 \in \mathbb{R}). \tag{10.6}$$

Note that these quantities are positive for an exchangeable resampling weight vector W:

$$0 < A_W \le B_W \le C_W\sqrt{1 - 1/n}.$$

Moreover, if the weights are i.i.d., we have $C_W = \mathrm{var}(W_1)^{\frac{1}{2}}$. We can now state the main result of this section:

THEOREM 10.1. *Fix $\alpha \in (0,1)$ and $p \in [1,\infty]$. Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be any function subadditive, positive-homogeneous and bounded by the p-norm, and let W be an exchangeable resampling weight vector.*

(1) *If $\mathbf{Y}$ satisfies (GA), then*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]}{B_W} + \|\sigma\|_p\,\overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{nB_W} + \frac{1}{\sqrt{n}}\right] \tag{10.7}$$

holds with probability at least $1-\alpha$. The same bound holds for the lower deviations, i.e. *with inequality* (10.7) *reversed and the additive term replaced by its opposite.*

(2) *If* $\mathbf{Y}$ *satisfies (BA)*(p, M) *and (SA), then*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]}{A_W} + \frac{2M}{\sqrt{n}}\sqrt{\log(1/\alpha)} \tag{10.8}$$

holds with probability at least $1-\alpha$*. If moreover the weights satisfy the assumption of* (10.6)*, then*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) > \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]}{D_W} - \frac{M}{\sqrt{n}}\sqrt{1 + \frac{A_W^2}{D_W^2}}\sqrt{2\log(1/\alpha)} \tag{10.9}$$

holds with probability at least $1-\alpha$*.*

Inequalities (10.7), (10.8) and (10.9) give thresholds such that the corresponding regions of the form (10.1) are confidence regions of level at least $1-\alpha$.

In specific situations, it can be the case that an alternate analysis of the problem can lead to deriving a deterministic threshold t_α such that $\mathbb{P}(\phi\left(\overline{\mathbf{Y}} - \mu\right) > t_\alpha) \leq \alpha$. In this case, we would ideally like to take the "best of two approaches" and consider the minimum of t_α and the resampling-based thresholds considered above. In the Gaussian case, the following corollary establishes that we can combine the concentration threshold corresponding to (10.7) with t_α to obtain a threshold that is very close to the minimum of the two.

COROLLARY 10.1. *Fix* $\alpha, \delta \in (0,1)$*,* $p \in [1,\infty]$ *and take* ϕ *and* W *as in Theorem 10.1. Suppose that* $\mathbf{Y}$ *satisfies (GA) and that* $t_{\alpha(1-\delta)}$ *is a real number such that* $\mathbb{P}\left(\phi\left(\overline{\mathbf{Y}} - \mu\right) > t_{\alpha(1-\delta)}\right) \leq \alpha(1-\delta)$*. Then with probability at least* $1-\alpha$*,* $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ *is upper bounded by the minimum between* $t_{\alpha(1-\delta)}$ *and*

$$\frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_p C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right). \tag{10.10}$$

REMARK 10.1.

(1) Cor. 10.1 is more precisely a consequence of Prop. 10.4 (ii).

(2) The important point to notice in Corollary 10.1 is that, since the last term of (10.10) becomes negligible with respect to the rest when n grows large, we can choose δ to be quite small (for instance $\delta = 1/n$), and obtain a threshold very close to the minimum between t_α and the threshold corresponding to (10.7). Therefore, this result is more subtle than just considering the minimum of two testing thresholds each taken at level $1 - \frac{\alpha}{2}$, as would be obtained by a direct union bound.

(3) For instance, if $\phi = \sup(\cdot)$ (resp. $\sup|\cdot|$), Cor. 10.1 may be applied with t_α equal to the classical Bonferroni threshold (obtained using a simple union bound over coordinates)

$$t_{\mathrm{Bonf},\alpha} := \frac{1}{\sqrt{n}}\|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{K}\right)\left(\text{resp. } t'_{\mathrm{Bonf},\alpha} := \frac{1}{\sqrt{n}}\|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{2K}\right)\right).$$

We thus obtain a confidence region almost equal to Bonferroni's for small correlations and better than Bonferroni's for strong correlations (see simulations in Sect. 10.5).

The proof of Thm. 10.1 involves results which are of self interest: the comparison between the expectations of the two processes $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ and $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and the concentration of these processes around their means. This is examinated in the two following subsections. Then, we give some elements for a wise choice of resampling weight vectors among several classical examples. The last subsection tackles the practical issue of computation time, and proposes two ways of solving it.

10.2.2. Comparison in expectation. In this section, we compare $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right]$. We note that these expectations exist in the Gaussian and the bounded case provided that ϕ is measurable and bounded by a p-norm. Otherwise, in particular in Prop. 10.2 and 10.3, we assume that these expectations exist. In the Gaussian case, these quantities are equal up to a factor that depends only on the distribution of W:

PROPOSITION 10.2. *Let $\mathbf{Y}$ be a sample satisfying (GA) and W a resampling weight vector. Then, for any measurable positive-homogeneous function $\phi : \mathbb{R}^K \to \mathbb{R}$, we have the following equality:*

$$B_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] = \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \quad . \tag{10.11}$$

REMARK 10.2. (1) In general, we can compute the value of B_W by simulation. For some classical weights, we give bounds or exact expressions (see Tab. 10.1 and Sect. 10.7.4).

(2) In a non-Gaussian framework, the constant B_W is still relevant, at least asymptotically: in their Thm. 3.6.13, van der Vaart and Wellner [**vdVW96**] use the limit of B_W when n goes to infinity as a normalizing constant.

(3) If the weights satisfy $\sum_{i=1}^n (W_i - \overline{W})^2 = n$ a.s., then (10.11) holds for any function ϕ (and $B_W = 1$).

When the sample is only symmetric we obtain the following inequalities:

PROPOSITION 10.3. *Let $\mathbf{Y}$ be a sample satisfying (SA), W an exchangeable resampling weight vector and $\phi : \mathbb{R}^K \to \mathbb{R}$ any subadditive, positive-homogeneous function.*

(i) *We have the general following lower bound:*

$$A_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] \leq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \quad . \tag{10.12}$$

(ii) *Moreover, if the weights satisfy the assumption of* (10.6)*, we have the following upper bound*

$$D_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] \geq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \quad . \tag{10.13}$$

REMARK 10.3. (1) The bounds (10.12) and (10.13) are tight for Rademacher and Random hold-out ($n/2$) weights, but far less optimal in some other cases like Leave-one-out (see Sect. 10.2.4 for details).

(2) When $\mathbf{Y}$ is not assumed to be symmetric and $\overline{W} = 1$ a.s., Prop. 2 of Fromont [**Fro07**] shows that (10.12) holds with $\mathbb{E}(W_1 - \overline{W})_+$ instead of A_W. Therefore, the symmetry of the sample allows us to get a tighter result (for instance twice sharper with Efron or Random hold-out (q) weights). According to Chap. 9 (in particular Prop. 9.2, 9.3 and 9.4), (10.12) does not hold in general. However, we conjecture that (10.12) could be generalized (up to some small additional term) when $\mathbf{Y}$ is not "too asymmetric".

10.2.3. Concentration around the expectation. In this section we present concentration results for the two processes $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ and $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ in the Gaussian framework.

PROPOSITION 10.4. *Let $p \in [1, +\infty]$, $\mathbf{Y}$ a sample satisfying (GA) and $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the p-norm.*

(i) *For all $\alpha \in (0,1)$, with probability at least $1-\alpha$ the following holds:*

$$\phi\left(\overline{\mathbf{Y}}-\mu\right) < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] + \frac{\|\sigma\|_p \, \overline{\Phi}^{-1}(\alpha/2)}{\sqrt{n}} \quad , \tag{10.14}$$

and the same bound holds for the corresponding lower deviations.

Efron	$2\left(1-\frac{1}{n}\right)^n = A_W \leq B_W \leq \sqrt{\frac{n-1}{n}} \quad C_W = 1$
Efr., $n \to +\infty$	$\frac{2}{e} = A_W \leq B_W \leq 1 = C_W$
Rademacher	$1-\frac{1}{\sqrt{n}} \leq A_W \leq B_W \leq \sqrt{1-\frac{1}{n}} \quad C_W = 1 \leq D_W \leq 1+\frac{1}{\sqrt{n}}$
Rad., $n \to +\infty$	$A_W = B_W = C_W = D_W = 1$
R. h.-o. (q)	$A_W = 2\left(1-\frac{q}{n}\right) \quad B_W = \sqrt{\frac{n}{q}-1}$
	$C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q}-1} \quad D_W = \frac{n}{2q} + \left\lvert 1-\frac{n}{2q}\right\rvert$
R. h.-o. $(n/2)$ $(2\vert n)$	$A_W = B_W = D_W = 1 \quad C_W = \sqrt{\frac{n}{n-1}}$
Leave-one-out	$\frac{2}{n} = A_W \leq B_W = \frac{1}{\sqrt{n-1}} \quad C_W = \frac{\sqrt{n}}{n-1} \quad D_W = 1$

TABLE 10.1. Resampling constants for classical resampling weight vector.

(ii) *Let W be some exchangeable resampling weight vector. Then, for all $\alpha \in (0,1)$, with probability at least $1-\alpha$ the following holds:*

$$\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] + \frac{\|\sigma\|_p C_W \overline{\Phi}^{-1}(\alpha/2)}{n} \ , \tag{10.15}$$

and the same bound holds for the corresponding lower deviations.

The bound (10.14) with a remainder in $n^{-1/2}$ is classical. The last one (10.15) is much more interesting since it it illustrates one of the key properties of resampling: the "stabilization effect". Indeed, the resampling quantity $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ concentrates around its expectation at the rate $C_W n^{-1} = o\left(n^{-1/2}\right)$ for most of the weights (see Sect. 10.2.4 and Tab. 10.1 for more details). Thus, compared to the original process, it is almost deterministic and equal to $B_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right]$. In an asymptotic viewpoint, this may be understood through Edgeworth expansions. Indeed, it is well-known (see for instance Hall [**Hal92**]) that when ϕ is smooth enough, the first non-zero term in the Edgeworth expansion of $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] - \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ is at least of order n^{-1}.

REMARK 10.4. Combining expression (10.11) and Prop. 10.4 (ii), we derive that for a Gaussian sample $\mathbf{Y}$ and any $p \in [1,\infty]$, the following upper bound holds with probability at least $1-\alpha$:

$$\mathbb{E}\left\|\overline{\mathbf{Y}}-\mu\right\|_p < \frac{\mathbb{E}_W\left[\left\|\overline{\mathbf{Y}}_{[W-\overline{W}]}\right\|_p\right]}{B_W} + \frac{\|\sigma\|_p C_W}{n B_W}\overline{\Phi}^{-1}(\alpha/2) \ , \tag{10.16}$$

and a similar lower bound holds. This gives a control with high probability of the L^p-risk of the estimator $\overline{\mathbf{Y}}$ of the mean $\mu \in \mathbb{R}^K$ at the rate $C_W B_W^{-1} n^{-1}$.

10.2.4. Resampling weight vectors. In this section, we consider the question of choosing some appropriate exchangeable resampling weight vector W when using Thm. 10.1 or Cor. 10.1. We define the following classical weights:

(1) **Rademacher**: W_i i.i.d. Rademacher variables, *i.e.* $W_i \in \{-1,1\}$ with equal probabilities.
(2) **Efron**: W has a multinomial distribution with parameters $(n; n^{-1}, \ldots, n^{-1})$.
(3) **Random hold-out** (q) (R. h.-o.), $q \in \{1,\ldots,n\}$: $W_i = \frac{n}{q}\mathbb{1}_{i\in I}$, where I is uniformly distributed on subsets of $\{1,\ldots,n\}$ of cardinality q. These weights may also be called cross validation weights, or leave-$(n-q)$-out weights. A classical choice is $q = n/2$ (when $2|n$). When $q = n-1$, these weights are called **leave-one-out** weights.

For these classical weights, exact or approximate values for the quantities A_W, B_W, C_W and D_W (defined by equations (10.3) to (10.6)) can be easily derived (see Tab. 10.1). Proofs are given

Resampling	$C_W B_W^{-1}$ (accuracy)	$\mathrm{Card}\,(\mathrm{supp}\,\mathcal{L}(W))$ (complexity)
Efron	$\leq \frac{1}{2}\left(1-\frac{1}{n}\right)^{-n} \xrightarrow[n\to\infty]{} \frac{e}{2}$	$\binom{2n-1}{n-1} \propto n^{-\frac{1}{2}} 4^n$
Rademacher	$\leq \left(1-n^{-1/2}\right)^{-1} \xrightarrow[n\to\infty]{} 1$	2^n
R. h.-o. $(n/2)$	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	$\binom{n}{n/2} \propto n^{-1/2} 2^n$
Leave-one-out	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	n

TABLE 10.2. Choice of the resampling weight vectors: accuracy-complexity tradeoff.

in Sect. 10.1, where several other weights are considered. Now, to use Thm. 10.1 or Cor. 10.1, we have to choose a particular resampling weight vector. In the Gaussian case, we propose the following accuracy and complexity criteria:

- first, relation (10.7) suggests that the quantity $C_W B_W^{-1}$ can be proposed as *accuracy* index for W. Namely, this index enters directly in the deviation term of the corresponding upper bound and the smaller the index is, the sharper the bound.
- second, an upper bound on the computational burden to compute exactly the resampling quantity is given by the cardinality of the support of $\mathcal{D}(W)$, thus providing a *complexity* index.

These two criteria are estimated in Tab. 10.2 for classical weights. For any exchangeable weight vector W, we have $C_W B_W^{-1} \geq [n/(n-1)]^{1/2}$ and the cardinality of the support of $\mathcal{D}(W)$ is greater than n. Therefore, the *leave-one-out weights* satisfy the best accuracy-complexity trade-off among exchangeable weights.

REMARK 10.5 (Link to leave-one-out prediction risk estimation). Consider using $\overline{\mathbf{Y}}$ for *predicting* a new data point $\mathbf{Y}^{n+1} \sim \mathbf{Y}^1$ (independent of $\mathbf{Y} = (Y^1, \ldots, Y^n)$). The corresponding L^p-prediction risk is given by $\mathbb{E}\left\|\overline{\mathbf{Y}} - \mathbf{Y}^{n+1}\right\|_p$. For Gaussian variables, this prediction risk is proportional to the L^p-risk: $\mathbb{E}\left\|\overline{\mathbf{Y}} - \mu\right\|_p = (n+1)^{\frac{1}{2}} \mathbb{E}\left\|\overline{\mathbf{Y}} - \mathbf{Y}^{n+1}\right\|_p$, so that the estimator of the L^p-risk proposed in Remark 10.4 leads to an estimator of the prediction risk. In particular, using leave-one-out weights and denoting by $\overline{\mathbf{Y}}^{(-i)}$ the mean of the $(\mathbf{Y}^j, j \neq i, 1 \leq j \leq n)$, we have then established that the leave-one-out estimator

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\overline{\mathbf{Y}}^{(-i)} - \mathbf{Y}^i\right\|_p$$

correctly estimates the prediction risk (up to the factor $(1-1/n^2)^{\frac{1}{2}} \sim 1$).

10.2.5. Practical computation of the thresholds. In practice, the exact computation of the resampling quantity $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ can still be too complex for the weights define above. To address this issue, we consider here two possible ways: first, we can use non-exchangeable weights with a lower complexity index and for which the exact computation is tractable. Alternatively, we propose to use a Monte-Carlo approximation, as is often done in practice to compute resampled quantities. In both cases, the thresholds have to be made slightly larger in order to keep the level larger than $1-\alpha$. This is detailed in the two paragraphs below.

V-fold cross-validation weights. In order to reduce the computation complexity, we can use "piecewise exchangeable" weights instead: consider a regular partition $(B_j)_{1\leq j\leq V}$ of $\{1, \ldots, n\}$ (where $V \in \{2, \ldots, n\}$ and $V|n$), and define the weights $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ with J uniformly distributed on $\{1, \ldots, V\}$. These weights are called the **(regular) V-fold cross validation** weights (VFCV).

Applying our results to the process $(\widetilde{\mathbf{Y}}^j)_{1\leq j\leq K}$ where $\widetilde{\mathbf{Y}}^j = \frac{V}{n}\sum_{i\in B_j}\mathbf{Y}^i$ is the empirical mean of $\mathbf{Y}$ on block B_j, we can show that Thm. 10.1 can be extended to (regular) V-fold cross validation weights with the following resampling constants [4]:

$$A_W = \frac{2}{V} \quad B_W = \frac{1}{\sqrt{V-1}} \quad C_W = \frac{\sqrt{n}}{V-1} \quad D_W = 1 \ .$$

With VFCV weights, the complexity index is only V, but we loose a factor $[(n-1)/(V-1)]^{1/2}$ in the accuracy index. The most accurate weights are leave-one-out ones ($V = n$), whereas the 2-fold ones are the best from the computational viewpoint. The choice of V thus relies on the balance between those two terms and depends on the particular features of each problem.

More general non-exchangeable weights are studied in Sect. 10.7.5. In this section, we focused on regular V-fold cross-validation weights because of they are both simple and efficient.

Monte-Carlo approximation. When we use a Monte-Carlo approximation to evaluate

$$\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \ ,$$

we draw randomly a small number B of i.i.d. weight vectors $W^1,\ldots,W^B$ and compute

$$\frac{1}{B}\sum_{k=1}^{B}\phi\left(\overline{\mathbf{Y}}_{\left[W^k-\overline{W^k}\right]}\right) \ .$$

This method is quite standard in the bootstrap literature and can be improved in several ways (see for instance [**Hal92**], appendix II). In Prop. 10.5 below, we propose an explicit correction of the concentration thresholds that takes into account $B < \infty$ for bounded weights.

PROPOSITION 10.5. *Let $B \geq 1$ and $W^1,\ldots,W^B$ be i.i.d. exchangeable resampling weight vectors such that $W_1^1 - \overline{W^1} \in [a;b]$ a.s. Let $p \in [1,+\infty]$, $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the p-norm.*

If $\mathbf{Y}$ is a fixed *sample and for every $k \in \{1,\ldots,K\}$, M_k is a median of $\left(\mathbf{Y}_k^i\right)_{1\leq i\leq n}$, then, for every $\beta \in (0;1)$,*

$$\begin{aligned}\frac{1}{B}\sum_{k=1}^{B}\phi\left(\overline{\mathbf{Y}}_{\left[W^k-\overline{W^k}\right]}\right) &\geq \mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \\ &\quad - \frac{b-a}{n}\sqrt{\frac{\ln(\beta^{-1})}{2B}}\left\|\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - M_k\right|\right)_k\right\|_p\end{aligned} \tag{10.17}$$

holds with probability at least $1-\beta$.

If $\mathbf{Y}$ is generated according to a distribution satisfying (GA), then, for every $\beta \in (0;1)$ and any deterministic $\nu \in \mathbb{R}^K$,

$$\left\|\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - M_k\right|\right)_k\right\|_p \leq \mathbb{E}\left\|\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - \nu_k\right|\right)_k\right\|_p + \|\sigma\|_p\,\overline{\Phi}^{-1}(\beta/2)\sqrt{n} \tag{10.18}$$

holds with probability at least $1-\beta$.

For instance, with Rademacher weights, we can use (10.17) with $b-a = 2$ and $\beta = \delta\alpha$ ($\delta \in (0,1)$). Then, in the thresholds built upon Thm. 10.1 and Cor. 10.1, one can replace $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ by its Monte-Carlo approximation at the price of changing α into $(1-\delta)\alpha$,

[4] When V does not divide n and the blocks are no longer regular, Thm. 10.1 can also be generalized, but the constants have more complex expressions. See Sect. 10.7.5.

and adding

$$\frac{2}{B_W n}\sqrt{\frac{\ln(1/(\delta\alpha))}{2B}}\left\|\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - M_k\right|\right)_k\right\|_p \tag{10.19}$$

to the threshold.

Note that (10.17) holds conditionally to the observed sample (hence independently from the unknown law of $\mathbf{Y}$), so that B can be chosen in function of $\mathbf{Y}$ in (10.19). Therefore, we can choose B with the following strategy: first, compute a rough estimate $t_{\text{est},\alpha}$ of the final threshold (*e.g.* if $\phi = \|\cdot\|_\infty$ and $\mathbf{Y}$ is gaussian, take the Bonferroni threshold $\|\sigma\|_\infty n^{-1/2}\overline{\Phi}^{-1}(\alpha/(2K))$ or the single test threshold $\|\sigma\|_\infty n^{-1/2}\overline{\Phi}^{-1}(\alpha/2)$). Second, choose B such that (10.19) is much smaller than $t_{\text{est},\alpha}$.

REMARK 10.6. In the Gaussian case, (10.18) gives a theoretical upper bound on the additive term (if one can bound the expectation term). This is only useful to ensure that the correction (10.19) is negligible for reasonable values of B.

10.3. Confidence region using resampled quantiles

In this section, we consider a different approach to construct confidence regions, directly based on the estimation of the quantile via resampling. Remember that our setting is non-asymptotic, so that the standard asymptotic approaches cannot be applied here. For this reason, we based our approach on ideas coming from exact randomized tests and consider here the case where $\mathbf{Y}^1$ has a symmetric distribution and where W is an i.i.d Rademacher weight vector, that is, W_i i.i.d. with $W_1 \in \{-1, 1\}$ with equal probabilities.

10.3.1. Main result. The idea here is to approximate the quantiles of the distribution $\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}} - \mu\right)\right)$ by the quantiles of the corresponding resampling-based distribution:

$$\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\,\middle|\,\mathbf{Y}\right) \quad . \tag{10.20}$$

For this, we take advantage of the symmetry of each $\mathbf{Y}^i$ around its mean. Let us define for a function ϕ the resampled empirical quantile by:

$$q_\alpha(\phi, \mathbf{Y}) := \inf\left\{x \in \mathbb{R} \text{ s.t. } \mathbb{P}_W\left[\phi(\overline{\mathbf{Y}}_{[W]}) > x\right] \le \alpha\right\} \quad .$$

The following lemma, close in spirit to exact test results, easily derives from the "symmetrization trick", *i.e.* from taking advantage of the distribution invariance of the data via sign-flipping.

LEMMA 10.6. *Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Then the following holds:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}} - \mu) > q_\alpha\left(\phi, \mathbf{Y} - \mu\right)\right] \le \alpha. \tag{10.21}$$

Of course, since $q_\alpha\left(\phi, \mathbf{Y} - \mu\right)$ still depends on the unknown μ, we cannot use this threshold to get a confidence region of the form (10.1). Therefore, following the general philosophy of resampling, we propose to replace μ by $\overline{\mathbf{Y}}$ in $q_\alpha\left(\phi, \mathbf{Y} - \mu\right)$. The main technical result of this section quantifies the price to pay to perform this operation:

THEOREM 10.2. *Fix $\delta, \alpha_0 \in (0, 1)$. Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Let $f : \left(\mathbb{R}^K\right)^n \to [0, \infty)$ be a nonnegative measurable function on the set of the data sample. Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be a nonnegative, subadditive, positive-homogeneous function. Denote $\widetilde{\phi}(x) = \max\left(\phi(x), \phi(-x)\right)$. The following holds:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha_0(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}\right) + \gamma_1(\alpha_0\delta) f(\mathbf{Y})\right] \le \alpha_0 + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y})\right] \quad , \tag{10.22}$$

where

$$\gamma_1(\eta) = \frac{2\overline{\mathcal{B}}\left(n, \frac{\eta}{2}\right) - n}{n}$$

and

$$\overline{\mathcal{B}}(n,\eta) = \max\left\{ k \in \{0,\dots,n\} \middle| 2^{-n} \sum_{i=k}^{n} \binom{n}{i} \geq \eta \right\} \ ,$$

is the upper quantile function of a Binomial $(n, \frac{1}{2})$ *variable.*

REMARK 10.7. Note that from Hoeffding's inequality, we have

$$\frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha\delta}{2}\right) - n} \geq \left(\frac{n}{2\ln\left(\frac{2}{\alpha\delta}\right)} \right)^{1/2} \ .$$

We can use this in (10.22) to derive a more explicit (but slightly less accurate) inequality.

By iteration of Thm. 10.2 we obtain the following corollary:

COROLLARY 10.7. *Fix* J *a positive integer,* $(\alpha_i)_{i=0,\dots,J-1}$ *a finite sequence in* $(0,1)$ *and* $\beta, \delta \in (0,1)$. *Let* $\mathbf{Y}$ *be a data sample satisfying assumption (SA). Let* $\phi : \mathbb{R}^K \to \mathbb{R}$ *be a nonnegative, subadditive, positive-homogeneous function and* $f : \left(\mathbb{R}^K\right)^n \to [0,\infty)$ *be a nonnegative function on the set of data samples. Then the following holds:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}} - \mu) > q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \sum_{i=1}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}) \right]$$
$$\leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}) \right] \ , \quad (10.23)$$

where, for $k \geq 1$, $\gamma_k = n^{-k} \prod_{i=0}^{k-1} \left(2\overline{\mathcal{B}}\left(n, \frac{\alpha_i \delta}{2}\right) - n \right)$.

The rationale behind this result is that the sum appearing inside the probability in (10.23) should be interpreted as a series of corrective terms of decreasing order of magnitude, since we expect the sequence γ_k to be sharply decreasing. Looking at Hoeffding's bound, this will be the case if the levels are such that $\alpha_i \gg \exp(-n)$.

Looking at (10.23), we still have to deal with the trailing term on the right-hand-side to obtain a useful result. We did not succeed in obtaining a self-contained result based on the symmetry assumption (SA) alone. However, to upper-bound the trailing term, we can assume some additional regularity assumption on the distribution of the data. For example, if the data are Gaussian or bounded, we can apply the results of the previous section (or apply some other device like Bonferroni's bound (10.11)). Explicit formulas for the resulting thresholds are given in Sect. 10.4 and 10.5 (with $J = 1$). We want to emphasize that the bound used in this last step does not have to be particularly sharp: since we expect (in favorable cases) γ_J to be very small, the trailing probability term on the right-hand side as well as the contribution of $\gamma_J f(\mathbf{Y})$ to the left-hand side should be very minor. Therefore, even a coarse bound on this last term should suffice.

10.3.2. Practical computation of the resampled quantile. Since the above results use Rademacher weight vectors, the exact computation of the quantile q_α requires in principle 2^n iterations and thus is too complex as n becomes large. Therefore, it might be relevant to consider a block-wise Rademacher resampling scheme. For this, let $(B_j)_{1 \leq j \leq V}$ be a regular partition of

$\{1,\dots,n\}$ and for all $i \in B_j$, $W_i = W_j^B$, where $(W_j^B)_{1\le j\le V}$ are i.i.d. Rademacher. This is equivalent to applying the previous method to the block-averaged sample $(\widetilde{Y}_1,\dots,\widetilde{Y}_V)$, where $\widetilde{Y}_j$ is the average of the $(Y_i)_{i\in B_j}$. Because the $\widetilde{Y}_j$ are i.i.d. variables, all of the previous results carry over when replacing n by V. However, this results in a loss of accuracy in Theorem 10.2 (and then in Corollary 10.7).

Another way to address this computation complexity issue is to consider Monte-Carlo quantile approximation: let $\mathbf{W}$ denote a $n \times B$ matrix of i.i.d. Rademacher weights (independent of all other variables), and define

$$\widetilde{q}_\alpha(\phi,\mathbf{Y},\mathbf{W}) = \inf\left\{ x\in\mathbb{R} \text{ s.t. } \frac{1}{B}\sum_{j=1}^{B} \mathbb{1}_{\phi\left(\overline{\mathbf{Y}}_{[\mathbf{W}^j]}\right)\ge x} \le \alpha \right\},$$

that is, $\widetilde{q}_\alpha$ is defined just as q_α except that the true distribution $\mathbb{P}_W$ of the Rademacher weight vector is replaced by the empirical distribution constructed from the columns of $\mathbf{W}$, $\widetilde{\mathbb{P}}_W = B^{-1}\sum_{j=1}^{B}\delta_{\mathbf{W}^j}$. The following result then holds:

PROPOSITION 10.8. *Consider the same conditions as in Thm. 10.2 except the function f can now be a function of both $\mathbf{Y}$ and $\mathbf{W}$. We have:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu) > \widetilde{q}_{\alpha_0(1-\delta)}\left(\phi,\mathbf{Y}-\overline{\mathbf{Y}},\mathbf{W}\right) + \gamma(\mathbf{W},\alpha_0\delta) f(\mathbf{Y},\mathbf{W})\right] \le \widetilde{\alpha}_0 + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > f(\mathbf{Y},\mathbf{W})\right],$$

where $\widetilde{\alpha}_0 = \frac{\lfloor B\alpha_0\rfloor+1}{B+1} \le \alpha_0 + \frac{1}{B+1}$ *and*

$$\gamma(\mathbf{W},\eta) := \max\left\{ y\ge 0 \text{ s.t. } \frac{1}{B}\sum_{j=1}^{B}\mathbb{1}_{|\overline{W^j}|\ge y} \ge \eta \right\}.$$

is the $(1-\eta)$-quantile of $|\overline{W}|$ under the empirical distribution $\widetilde{\mathbb{P}}_W$.

Note that for practical purposes, we can choose $f(\mathbf{W},\mathbf{Y})$ to depend on $\mathbf{Y}$ only and use another type of bound to control the last term on the right-hand side, see discussion in the previous section. The above result tells us that if we replace in Theorem 10.2 the true quantile by an empirical quantile based on B i.i.d. weight vectors, and the factor γ_1 is similarly replaced by an empirical quantile of $|\overline{W}|$, then we lose at most $(B+1)^{-1}$ in the corresponding covering probability. Furthermore, it can be seen easily that if α_0 is taken to be a positive multiple of $(B+1)^{-1}$, then there is no loss in the final covering probability (*i.e.* $\widetilde{\alpha}_0 = \alpha_0$).

10.4. Application to multiple testing

In this section, we describe how the results of Sect. 10.2 and 10.3 can be used to derive multiple testing procedures. We focus on the two following multiple testing problems:

- *One-sided problem*: test simultaneously the null hypotheses $H_{0,k}$: "$\mu_k \le 0$" against $H_{1,k}$: "$\mu_k > 0$", $1 \le k \le K$
- *Two-sided problem*: test simultaneously the null hypotheses $H_{0,k}$: "$\mu_k = 0$" against $H_{1,k}$: "$\mu_k \ne 0$", $1 \le k \le K$.

In this context, we precise the link between confidence regions and multiple testing, and explain how to improve our resampling-based thresholds. We first introduce a few more notations:

- $\mathcal{H} := \{1, \dots, K\}$, $\mathcal{H}_0 := \{1 \leq k \leq K \text{ s.t. } H_{0,k} \text{ is true}\}$ and $\mathcal{H}_0^c$ its complementary in $\mathcal{H}$. Note that $\mathcal{H}_0$ is course unknown since the goal of multiple testing is in fact precisely to estimate this set.
- For any $x \in \mathbb{R}$, the bracket $[x]$ denotes either x in the one-sided context or $|x|$ in the two-sided context.
- Reordering the coordinates of $\overline{\mathbf{Y}}$

$$\left[\overline{\mathbf{Y}}_{\sigma(1)}\right] \geq \left[\overline{\mathbf{Y}}_{\sigma(2)}\right] \geq \cdots \geq \left[\overline{\mathbf{Y}}_{\sigma(K)}\right]$$

with a permutation σ of $\mathcal{H}$, we define for every $i \in \{1, \dots, K\}$, $\mathcal{C}_i(\mathbf{Y}) := \{\sigma(j) \text{ s.t. } j \geq i\}$ the set which contains the $K - i + 1$ smaller coordinates of $\overline{\mathbf{Y}}$. In particular, $\mathcal{C}_1 = \mathcal{H}$.
- For any $\mathcal{C} \subset \{1, \dots, K\}$,

$$T(\mathcal{C}) := \sup_{k \in \mathcal{C}} \left[\overline{\mathbf{Y}}_k - \mu_k\right] \qquad T'(\mathcal{C}) := \sup_{k \in \mathcal{C}} \left[\overline{\mathbf{Y}}_k\right]$$

We remark that $T(\mathcal{H}) \geq T(\mathcal{H}_0) \geq T'(\mathcal{H}_0)$ in general and $T(\mathcal{H}_0) = T'(\mathcal{H}_0)$ in the two-sided context.

10.4.1. Multiple testing and connection with confidence regions. A multiple testing procedure is a (measurable) function of $\mathbf{Y}$,

$$R(\mathbf{Y}) \subset \mathcal{H} \ ,$$

that rejects the null hypotheses $H_{0,k}$ with $k \in R(\mathbf{Y})$. For such a multiple testing procedure R, a type I error arises as soon as R rejects at least one hypothesis which is in fact true. The family-wise error rate of R is then the probability that at least one type I error occurs:

$$\mathrm{FWER}(R) := \mathbb{P}\left(R(\mathbf{Y}) \cap \mathcal{H}_0 \neq \emptyset\right) \ .$$

Given a level $\alpha \in (0, 1)$, our goal is then to build a multiple testing procedure R with

$$\mathrm{FWER}(R) \leq \alpha. \tag{10.24}$$

Of course, choosing the procedure $R = \emptyset$ (*i.e.* the procedure which rejects no null hypothesis), satisfies trivially the problem. Therefore, provided that (10.24) holds, we want the average number of rejected false null hypotheses, that is

$$\mathbb{E}\left[R(\mathbf{Y}) \cap \mathcal{H}_0^c\right] \ , \tag{10.25}$$

to be as large as possible.

A common way to build a multiple testing procedure is to reject the null hypotheses $H_{0,k}$ corresponding to

$$R(\mathbf{Y}) = \left\{1 \leq k \leq K \text{ s.t. } \left[\overline{\mathbf{Y}}_k\right] > t\right\} \ , \tag{10.26}$$

where t is a (possibly data-dependent) threshold. From now on, we will restrict our attention to multiple testing procedures of the previous form. In this case, the deterministic threshold that maximises (10.25) provided that (10.24) holds is obviously the $1 - \alpha$ quantile of the distribution of $T'(\mathcal{H}_0)$. However, the latter quantile cannot be directly accessed, because it depends both on the unknown dependency structure between the coordinates of $\mathbf{Y}^i$ and on the unknown set $\mathcal{H}_0$. The aim of the following sections (10.4.2, 10.4.3, 10.4.4) will be to approach this quantity.

This should be compared to the confidence region context, where the smallest deterministic threshold for which (10.1) holds with $\phi = \sup[\cdot]$ is the $(1 - \alpha)$ quantile of the distribution of $T(\mathcal{H})$. Since $T(\mathcal{H}) \geq T'(\mathcal{H}_0)$, we observe following:

(1) The thresholds that give confidence regions of the form (10.1) with $\phi = \sup[\cdot]$ also give multiple testing procedures with a FWER smaller than α (following the thresholding procedure (10.26)). Therefore, we can directly derived from Sect. 10.2 and 10.3 resampling-based multiple testing procedures that control the FWER.
(2) One might expect to be able to find better (*i.e.* smaller) thresholds in the multiple testing framework than in the confidence region framework. Therefore, when $\mathcal{H}_0^c$ is "large", $T(\mathcal{H})$ is "significantly larger" than $T'(\mathcal{H}_0)$ and then procedures based on upper bounding $T(\mathcal{H})$ are conservative. A method commonly used to address this issue is to consider step-down procedures. This is examinated in the following section.

10.4.2. Background on step-down procedures. We review in this section known facts on step-down procedure (see Romano and Wolf [**RW05**]). We consider here thresholds $\mathbf{t}$ of the following general form:

$$\mathbf{t} : \mathcal{C} \subset \mathcal{H} \mapsto \mathbf{t}(\mathcal{C}) \in \mathbb{R} \ .$$

We call such a threshold a *subset-based threshold* since it gives a value to each subset of $\mathcal{H}$. A subset-based threshold is said to be *non-decreasing* if for all subsets $\mathcal{C}$ and $\mathcal{C}'$, we have

$$\mathcal{C} \subset \mathcal{C}' \quad \Rightarrow \quad \mathbf{t}(\mathcal{C}) \leq \mathbf{t}(\mathcal{C}') \ .$$

In our setting, a non-decreasing subset-based threshold is easily obtained by taking a supremum over a subset $\mathcal{C}$ of coordinates. In particular, the thresholds derived from Sect. 10.2 (resp. Sect. 10.3) define non-decreasing subset-based thresholds, by taking $\phi = \sup_{\mathcal{C}}[\cdot]$ (resp. $\phi = 0 \vee \sup_{\mathcal{C}}[\cdot]$).

Definition 10.1 (Step-down procedure with subset-based threshold). Let $\mathbf{t}$ be a non-decreasing subset-based threshold and note for all i, $t_i = \mathbf{t}(\mathcal{C}_i)$. The step-down procedure with threshold $\mathbf{t}$ rejects

$$\left\{ 1 \leq k \leq K \text{ s.t. } \left[\overline{\mathbf{Y}}_k\right] \geq t_{\hat{\ell}} \right\}$$

where $\hat{\ell} = \max\left\{ 1 \leq i \leq K \text{ s.t. } \forall j \leq i, \left[\overline{\mathbf{Y}}_{\sigma(j)}\right] \geq t_j \right\}$ when the latter maximum exists, and the procedure rejects no null hypothesis otherwise.

A step-down procedure of the above form can be computed using the following iterative algorithm:

Algorithm 10.1.

(1) Init: define $R_0 := \emptyset$, $\mathcal{E}_0 := \mathcal{H}$.
(2) Iteration $i \geq 1$: put $\mathcal{E}_i := \mathcal{E}_{i-1} \backslash R_{i-1}$ and $R_i = \left\{ k \in \mathcal{E}_i \text{ s.t. } \left[\overline{\mathbf{Y}}_k\right] \geq \mathbf{t}(\mathcal{E}_i) \right\}$.
If $R_i = \emptyset$, stop and reject the null hypotheses corresponding to:

$$R(\mathbf{Y}) := \{ \sigma(k) \text{ s.t. } k \in \cup_{j \leq i-1} R_j \} \ .$$

Otherwise, go to iteration $i + 1$.

We recall here Thm. 1 of Romano and Wolf [**RW05**], adapted to our setting:

Theorem 10.3 (Romano and Wolf, 2005). *Let $\mathbf{t}$ be a non-decreasing subset-based threshold. Then the step-down procedure R of threshold $\mathbf{t}$ satisfies,*

$$\text{FWER}(R) \leq \mathbb{P}(T(\mathcal{H}_0) \geq \mathbf{t}(\mathcal{H}_0)). \tag{10.27}$$

As a consequence, Algorithm 10.1 with any threshold derived from Sect. 10.2 (resp. Sect. 10.3) with $\phi = \sup_{\mathcal{H}_0}[\cdot]$ (resp. $\phi = 0 \vee \sup_{\mathcal{H}_0}[\cdot]$) gives a multiple testing procedure with control of the FWER. We detail this in the following section.

10.4.3. Using our confidence regions to build step-down procedures. Using Thm. 10.3 and Cor. 10.1 (wherein we use the Bonferroni threshold), we derive:

COROLLARY 10.9. *Fix $\alpha, \delta \in (0,1)$. Let W be an exchangeable resampling weight vector and suppose that $\mathbf{Y}$ satisfies (GA). Then, in the one-sided context, the step-down procedure with the following subset-based threshold controls the FWER at level α:*

$$\mathcal{C} \mapsto \min\left(\frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{\mathrm{Card}(\mathcal{C})} \right), \frac{\mathbb{E}_W\left[\sup_{k\in\mathcal{C}} \left\{ \left(\overline{\mathbf{Y}}_{[W-\overline{W}]} \right)_k \right\} \right]}{B_W} + \varepsilon(\alpha,\delta,n) \right)$$

where $\varepsilon(\alpha,\delta,n) = \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2} \right) + \frac{\|\sigma\|_\infty C_W}{nB_W} \overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2} \right)$.

Using Thm. 10.3 and 10.2 (with $\alpha_0 = \alpha(1-\gamma)$ and f equal to the Bonferroni threshold at level $\alpha\gamma/2$), we derive:

COROLLARY 10.10. *Fix $\alpha, \gamma, \delta \in (0,1)$. Let W be a Rademacher weight vector and suppose that $\mathbf{Y}$ satisfies (GA). Then, in the one-sided context, the step-down procedure with the following subset-based threshold controls the FWER at level α:*

$$\mathcal{C} \mapsto q_{\alpha(1-\delta)(1-\gamma)}\left(0 \vee \sup_{\mathcal{C}}(\cdot), \mathbf{Y} - \overline{\mathbf{Y}} \right) + \varepsilon'(\alpha,\delta,\gamma,n,\mathrm{Card}(\mathcal{C}))$$

where $\varepsilon'(\alpha,\delta,\gamma,n,k) = \frac{2\overline{\mathcal{B}}(n,\alpha(1-\gamma)\delta/2)-n}{n} \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\alpha\gamma}{2k} \right)$.

Of course, analogues of Cor. 10.9 and 10.10 can also be derived for the two-sided problem.

REMARK 10.8.
(1) Note that the above (data-dependent) subset-based thresholds are translation-invariant because $\mathbf{Y} - \overline{\mathbf{Y}}$ is. Therefore, large values of non-zero means μ_k will not enlarge these thresholds.
(2) Both subset-based thresholds of Cor. 10.9 and 10.10 are built in order to improve "Bonferroni's subset-based threshold"

$$\mathcal{C} \mapsto \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\alpha}{\mathrm{Card}(\mathcal{C})} \right) \quad .$$

Therefore, the corresponding step-down procedures are expected to perform better than Holm's procedure (*i.e.* the step-down version of Bonferroni's procedure, see [**Hol79**]).

10.4.4. Uncentered quantile approach for two-sided testing. We now focus specifically on the two-sided multiple testing problem. A fundamental consequence of Thm. 10.3 is that only a weak control (*i.e.*, when $\mathcal{C} = \mathcal{H}_0$) of $T'(\mathcal{C}) = \sup_{k\in\mathcal{C}} |\overline{\mathbf{Y}}_k|$ is needed to obtain a step-down procedure with a strong control (*i.e.*, for arbitrary mean $\mu \in \mathbb{R}^K$) of the FWER. In this situation, the main problem dealt with in Sect. 10.3 disappears: namely, under the hypothesis that $\mathcal{H}_0 = \mathcal{C}$, by definition all the coordinates contributing to the supremum in $T'(\mathcal{C})$ are assumed to have zero mean, and therefore, following the reasoning in Lemma 10.6, a direct exact quantile approach is possible.

COROLLARY 10.11. *Let W be a Rademacher weight vector and suppose that $\mathbf{Y}$ satisfies (SA). Then for two-sided testing, the step down procedure with the subset-based threshold*

$$\mathcal{C} \mapsto q_\alpha\left(\sup_{\mathcal{C}} |\cdot|, \mathbf{Y} \right)$$

controls the FWER at level α.

Note the differences of this result with our main approach (*i.e.*, the analogue of Cor. 10.10 in the two-sided setting):

- there is no additional trailing term ε' and no "shrinking" in the level of the computed empirical quantile.
- the data is not recentered around the empirical expectation to compute the quantile.

In the following, we will call the threshold $q_\alpha\left(\sup_{\mathcal{C}} |\cdot|, \mathbf{Y}\right)$ the "uncentered quantile", while the threshold built using our main approach (including the additional term) will be called "recentered quantile threshold" for brevity.

To understand the practical consequences of these differences, let us consider an informal and qualitative argumentation.

- if $\mathcal{C} = \mathcal{H}_0$, then the empirical mean $\overline{\mathbf{Y}}$ should be close to 0. Hence, if we assume that replacing $\overline{\mathbf{Y}}$ by 0 does not change the centered quantile significantly, we conclude that the uncentered quantile threshold will be smaller (hence better) than the recentered quantile threshold, since the latter has the same form but at a slightly shrunk level and has an additional term ε'. In this situation the uncentered quantile will actually achieve the exact level (up to 2^{-n}).
- if on the other hand there are some coordinates with a large non-zero mean in the set $\mathcal{C}$ (by which we mean having a large signal-to-noise (SNR) ratio), then these coordinates will on average have a large absolute value and hence make the uncentered quantile significantly larger; in this case the signal will contribute to the uncentered quantile more than the noise. By contrast, and as remarked earlier, the recentered quantile threshold is translation invariant and thus not affected by the relative strength of the signal. Hence, in this situation, it is likely that the recentered quantile threshold will be smaller.

While the second situation above appears to be detrimental to the uncentered quantile, this disadvantage will in some sense be "automatically corrected" by the step-down procedure. Namely, if some coordinates have a large SNR, they will certainly contribute significantly to the uncentered quantile threshold at the first step of the step-down procedure; however even if this threshold is relatively large, it will still allow to eliminate at the first step precisely those coordinates having a very large mean. This will result in an important improvement of the threshold at the second iteration, and so on, until all coordinates with a large SNR have been weeded out, so that in the end we actually end up very close to the situation described in the first point.

Hence, the conclusion from this qualitative discussion is that, in a situation where some of the coordinates have a large SNR, we expect that the uncentered quantile will be less accurate (*i.e.*, larger) than the centered quantile threshold in the first iteration(s) of the step-down procedure, but that it will then improve along the iterations and eventually prevail in the race. (This behavior will be confirmed by our simulations in the next section.)

At this point, it seems that the step-down using the uncentered quantile is both simpler and more effective than our main approach and thus should always be preferred. However, this qualitative discussion also gives us another insight: the step-down procedure based on the uncentered quantile may need more iterations to converge since the first steps result in an inaccurate threshold. In order to fix this drawback, we propose to use the leverage of the recentered quantile for the first step in order to weed out in one single step most of coordinates having a large SNR, and then continue subsequently with the uncentered threshold in the next steps for more accuracy. We thus obtain the following algorithm:

ALGORITHM 10.2.

(1) Reject the null hypotheses corresponding to:

$$R_0 := \left\{ k \text{ s.t. } \left|\overline{\mathbf{Y}}_k\right| \geq q_{\alpha(1-\delta)(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \overline{\mathbf{Y}}) + \varepsilon'(\alpha, \delta, \gamma, n, K) \right\}$$

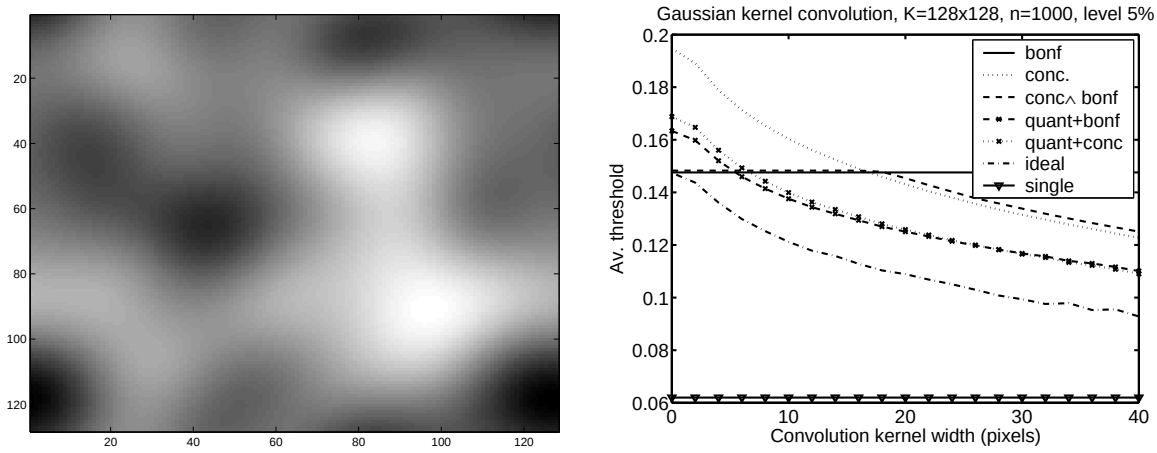


FIGURE 10.1. Left: example of a 128x128 pixel image obtained by convolution of Gaussian white noise with a (toroidal) Gaussian filter with width $b = 18$ pixels. Right: average thresholds obtained for the different approaches, see text.

(2) If $R_0 = \mathcal{H}$ then stop.
Otherwise, consider the set of the remaining coordinates $\mathcal{H} \backslash R_0$ and apply on it the step-down Algorithm 10.1 with the subset-based threshold

$$\mathcal{C} \mapsto q_{\alpha(1-\gamma)} \left(\sup_{\mathcal{C}} |\cdot| \, , \mathbf{Y} \right) \quad .$$

PROPOSITION 10.12. *Fix $\alpha, \gamma, \delta \in (0,1)$. Let W be a Rademacher weight vector and suppose that $\mathbf{Y}$ satisfies (GA). In the two-sided context, the Algorithm 10.2 gives a multiple testing procedure with a FWER smaller than α.*

What we expect is that the above algorithm will yield essentially the same final result as the one of Cor. 10.11 (up to some small loss in the level), while requiring less iterations. In numerical applications such as neuroimaging with a large number of images, where one iteration can take up to one day, this can result in a significant improvement.

10.5. Simulations

For simulations we consider data of the form $\mathbf{Y}_t = \mu_t + \mathbf{G}_t$, where t belongs to an $m \times m$ discretized 2D torus of $K = m^2$ "pixels", identified with $\mathbb{T}_m^2 = (\mathbb{Z}/m\mathbb{Z})^2$, and $\mathbf{G}$ is a centered Gaussian vector obtained by 2D discrete convolution of an i.i.d. standard Gaussian field ("white noise") on $\mathbb{T}_m^2$ with a function $F : \mathbb{T}_m^2 \to \mathbb{R}$ such that $\sum_{t \in \mathbb{T}_m^2} F^2(t) = 1$. This ensures that $\mathbf{G}$ is a stationary Gaussian process on the discrete torus, it is in particular isotropic with $\mathbb{E}\left[\mathbf{G}_t^2\right] = 1$ for all $t \in \mathbb{T}_m^2$.

In the simulations below we consider for the function F a "Gaussian" convolution filter of bandwidth b on the torus:

$$F_b(t) = C_b \exp\left(-d(0,t)^2/b^2\right) \quad ,$$

where $d(t,t')$ is the standard distance on the torus and C_b is a normalizing constant. Note that for actual simulations it is more convenient to work in the Fourier domain and to apply the inverse DFT which can be computed efficiently. We then compare the different thresholds obtained by the methods proposed in this work for varying values of b. Remember that the only information available to the algorithms is the bound on the marginal variance; the form of the function F_b itself is of course unknown.

10.5.1. Confidence balls. On Fig. 10.1 we compare the thresholds obtained when $\phi = \sup|\cdot|$, which corresponds to L^∞ confidence balls. Remember that these thresholds can be also directly used in the two-sided multiple testing situation (see Sect. 10.4). We use the different approaches proposed in this work, with the following parameters: the dimension is $K = 128^2 = 16\,384$, the number of data points per sample is $n = 1\,000$ (much smaller than K, so that we really are in a non-asymptotic framework), the width b takes even values in the range $[0, 40]$, the overall level is $\alpha = 0.05$.

Recall that the Bonferroni threshold is

$$t'_{\mathrm{Bonf},\alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{2K}\right) \ .$$

For the concentration threshold (10.7)

$$t_{\mathrm{conc},\alpha} := \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]}{B_W} + \|\sigma\|_p \overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{nB_W} + \frac{1}{\sqrt{n}}\right] \ ,$$

we used Rademacher weights. For the "compound" threshold of Cor. 10.1 (with Bonferroni as deterministic reference threshold)

$$t_{\mathrm{conc}\wedge\mathrm{bonf},\alpha} := \min\left\{ t'_{\mathrm{Bonf},\alpha}, \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_p C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right)\right\} \ ,$$

we used $\delta = 0.1$. For the quantile approach (10.23)

$$t_{\mathrm{quant+bonf},\alpha} := q_{\alpha_0(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}\right) + \frac{2\overline{\mathcal{B}}\left(n, \frac{\alpha_0\delta}{2}\right) - n}{n} t'_{\mathrm{Bonf},\alpha-\alpha_0}$$

$$t_{\mathrm{quant+conc},\alpha} := q_{\alpha_0(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}\right) + \frac{2\overline{\mathcal{B}}\left(n, \frac{\alpha_0\delta}{2}\right) - n}{n} t_{\mathrm{conc},\alpha-\alpha_0}\left(\mathbf{Y}\right) \ ,$$

we used $J = 1$, $\alpha_0 = 0.9\alpha$ $(= (1-\gamma)\alpha$ with $\gamma = 0.1)$, $\delta = 0.1$ and took f either equal to the Bonferroni or the concentration threshold, respectively (these values of $\alpha_0, \alpha, \gamma, \delta$ will stay unchanged for all the experiments presented here, including in the next section). Finally, for comparison purposes, we included in the figure the threshold corresponding to $K = 1$ (estimation of a single coordinate mean)

$$t_{\mathrm{single},\alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{2}\right) \ ,$$

and an estimation of the true quantile (actually, an empirical quantile over 1 000 samples), *i.e.* $t_{\mathrm{ideal},\alpha}$ the $1-\alpha$ quantile of the distribution of $\phi(\mathbf{Y} - \mu)$.

Each point represents an average over 50 experiments (except of course for $t'_{\mathrm{Bonf},\alpha}$, $t_{\mathrm{single},\alpha}$ and $t_{\mathrm{ideal},\alpha}$). The quantiles or expectation with Rademacher weights were estimated by Monte-Carlo with 1 000 draws (without the additional term introduced in Sect. 10.2.5). On the figure we did not include standard deviations: they are quite low, of the order of 10^{-3}, although it is worth noting that the quantile threshold has a standard deviation roughly twice as large as the concentration threshold (we did not investigate at this point what part of this variation is due to the MC approximation).

We also computed the quantile threshold $q_\alpha(\phi, \mathbf{Y} - \overline{\mathbf{Y}})$ without second-order term: it is so close to $t_{\mathrm{ideal},\alpha}$ that we would not distinguish them on Fig. 10.1.

The overall conclusion of this first preliminary experiment is that the different thresholds proposed in this work are relevant in the sense that they are smaller than the Bonferroni threshold

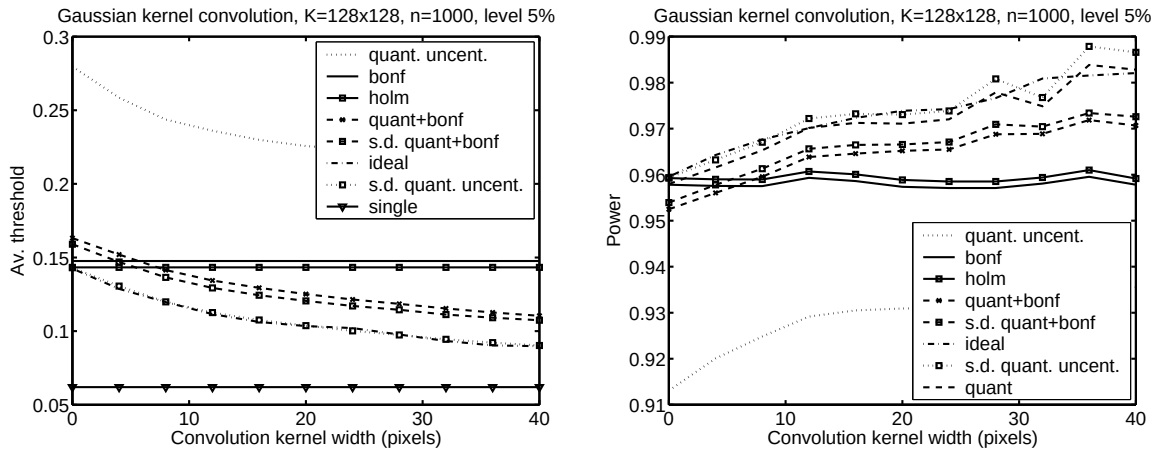


FIGURE 10.2. Multiple testing problem with μ defined by (10.28) for different approaches, see text. Left: average thresholds. Right: power, defined by (10.30).

provided the vector has strong enough correlations. As expected, the quantile approach appears to lead to tighter thresholds. (However, this might not be always the case for smaller sample sizes because of the additional term ε'.) One advantage of the concentration approach is that the "compound" threshold (10.10) can "fall back" on the Bonferroni threshold when needed, at the price of a minimal threshold increase.

10.5.2. Multiple testing. We now focus on the multiple testing problem. We present here only the two-sided case because the one-sided case gives similar results, except that we can not use the "uncentered quantile" method of Cor. 10.11.

We consider the experiment of the previous section, with the following choice for the vector of means:

$$\forall (i,j) \in \{0,\dots,127\}^2, \qquad \mu_{(i,j)} = \frac{(64-j)_+}{64} \times 20 t'_{\mathrm{Bonf},\alpha} \ . \tag{10.28}$$

In this situation, note that the half of the null hypotheses are true while the non-zero means are increasing linearly from 0 to $20t'_{\mathrm{Bonf},\alpha}$. The thresholds obtained are given on Figure 10.2 (100 simulations). The ideal threshold $t_{\mathrm{ideal},\alpha}$ is now derived from the $1-\alpha$ quantile of the distribution of $T'(\mathcal{H}_0) = \sup_{\mathcal{H}_0} \left|\overline{\mathbf{Y}}\right|$. We did not report $t_{\mathrm{conc},\alpha}$ and $t_{\mathrm{conc}\wedge\mathrm{bonf},\alpha}$ in order to simplify Fig. 10.2 (their values are unchanged, since these thresholds are translation invariant). In addition to the previous thresholds, we consider:

- the uncentered quantile:

$$t_{\mathrm{quant.uncent.},\alpha} := q_\alpha \left(\sup |\cdot|, \mathbf{Y}\right) \tag{10.29}$$

  and its step-down version $t_{\mathrm{s.d.quant.uncent.},\alpha}$ (see Cor. 10.11).
- the step-down version $t_{\mathrm{s.d.quant+bonf},\alpha}$ of $t_{\mathrm{quant+bonf},\alpha}$.
- Holm threshold $t_{\mathrm{Holm},\alpha}$ (*i.e.* the step-down version of the Bonferroni procedure).

On the right-hand-side of Fig. 10.2, we evaluated the powers of the different thresholds $t_\alpha(\mathbf{Y})$, defined as follows:

$$\mathrm{Power}(t_\alpha) := \frac{\mathrm{Card}\left\{1 \le k \le K \text{ s.t. } \mu_k \neq 0 \text{ and } |\mathbf{Y}_k| > t_\alpha(\mathbf{Y})\right\}}{\mathrm{Card}\left\{1 \le k \le K \text{ s.t. } \mu_k \neq 0\right\}} \ . \tag{10.30}$$

This experiment shows that:

(1) for single-step resampling-based procedures:

- the single-step procedure based on our quantile approach ("quant+bonf") can outperform Holm's procedure as soon as the the coordinates of the vector are sufficiently correlated.
- the single-step procedure based on the uncentered quantile ("quant. uncent") has bad performance.

(2) for step-down resampling-based procedures:

- the step-down procedure based on our quantile approach ("s.d. quant+bonf") can outperform Holm's procedure as soon as the the coordinates of the vector are sufficiently correlated (obvious from the point 1).
- the step-down procedure based on the uncentered quantile ("s.d. quant+bonf") seems to be the most efficient thresholds of the step-down procedures considered here.

However, when K and n are large, each iteration of the step-down algorithm for the uncentered quantiles may be quite long to compute[5] while our quantile approach ("quant+Bonf") provides in only one step a quite good accuracy. Following Sect. 10.4.4, these two methods can be combined (see Algorithm 10.2, called here "mixed approach"), resulting in a speed-accuracy trade-off.

We illustrate this with a specific simulation study: consider the same simulation framework as above unless that the bandwidth b is now fixed at 30, the size of the sample is $n = 100$ and the means are given by: $\forall (i,j) \in \{0,\dots,127\}^2$, $\mu_{(i,j)} = f(i+128j)$, where

$$\forall k \in \{0,\dots,8\,192\}, \qquad f(k) = 50 t'_{\mathrm{Bonf},\alpha} \times \exp\left(-\frac{(8\,192-k)_+}{8\,192}\log(100)\right), \tag{10.31}$$

and $f(k) = 0$ for the other values of k. In this situation, the non-zero means are increasing log-linearly from $0.5\, t'_{\mathrm{Bonf},\alpha}$ to $50\, t'_{\mathrm{Bonf},\alpha}$. With 100 simulations, we computed in Tab. 10.3 the average number of iterations in the step-down algorithm 10.1 for the above step-down procedures. Additionally, on Fig. 10.3, the power is given as a function of the number of iterations in the step-down algorithm for the different approaches.

We can read the following results:

- The "mixed approach" needs on average significantly less iterations to converge.
- In the case of a very strict computation time constraint, it is possible to stop the step-down procedures early after a fixed number of iterations. Stopping the mixed approach procedure after only 2 iterations results in an average power that is virtually undistinguishable from the power obtained without early stopping. By contrast 3 iterations are needed for the step-down with uncentered quantile threshold.

While these results are certainly specific to the particular simulation setup we used, they illustrate that the informal and qualitative analysis we presented in Sect. 10.4.4 appears to be correct. In particular, the fact that the mixed approach appears to give already very satisfactory results after the two first iterations reinforces the interpretation that the first step (using the recentered quantile threshold with remainder term) rules out at once all coordinates with a large SNR while the second step (using the exact, uncentered quantile) improves the precision once these high-SNR coordinates have been eliminated.

Therefore, this mixed approach can be an interesting alternative to the uncentered quantile approach when several long iterations in the step-down algorithm are expected. This situation arises typically when the signal (non-zero means) has a wide dynamic range (in our above simulation, the signal-to-noise ratio for non-true null hypotheses varies between 0.25 and 25).

[5] typically one day in the neuroimaging framework.

Holm's procedure	" s.d. quant+bonf"	"s.d. quant. uncent."	"mixed approach"
3.25	3.13	4.92	3.94

TABLE 10.3. Multiple testing problem with μ corresponding to (10.31) for different step-down approaches. Average number of iterations in the step-down algorithm.

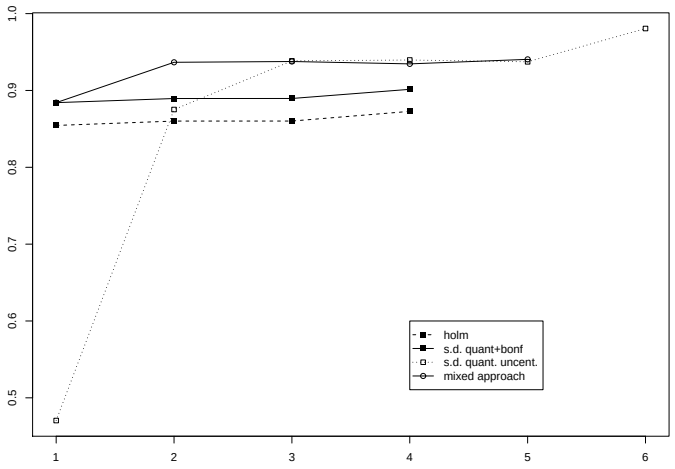


FIGURE 10.3. Multiple testing problem with μ corresponding to (10.31) for different step-down approaches. Power as a function of the number of iterations in the step-down algorithm.

10.6. Discussion and concluding remarks

10.6.1. Confidence regions and tests. In this paper we have first constructed confidence regions of the form (10.1) and presented an application of this result to (multiple) testing. Because of the duality between confidence regions and tests, a natural question is whether conversely, one could construct tests first and deduce confidence regions. In particular, testing the (single) null hypothesis $H_{\mu_0} : \mu = \mu_0$ is very simple using an exact symmetrization test: using directly Lemma 10.6 we know that the test $T_{\mu_0,\phi}$ rejecting H_{μ_0} if $\phi(\mathbf{Y} - \mu_0) > q_\alpha(\phi, \mathbf{Y} - \mu_0)$ has significance level bounded by α. We can construct from this the confidence region

$$\mathcal{F}_\phi(\mathbf{Y}, 1-\alpha) = \left\{ \mu_0 \in \mathbb{R}^K \text{ s.t. } T_{\mu_0,\phi} \text{ does not reject } H_{\mu_0} \right\} .$$

This method avoids completely the problems linked to the direct construction of a confidence region that we faced in Sect. 10.3; furthermore, the above confidence region is almost exactly of level $1-\alpha$ (up to 2^{-n}), while the region constructed in Sect. 10.3 is certainly more conservative. Nevertheless, argue that the approach developed in Sect. 10.3 is much more practically relevant:

- the region $\mathcal{F}_\phi(\mathbf{Y}, 1-\alpha)$ constructed above by test inversion is not of the form (10.1), that is, it is not a "ϕ-ball" around the empirical mean. However, it might be required by external contraints, for example for further analysis, that the confidence region should be of this form.
- more generally, it does not seem clear at all what shape the above region would take or even if would enjoy some desirable properties such as convexity. This seems very

impractical, particularly in high dimension, where regions which cannot be described under a simple form seem very difficult to handle. In fact, it seems actually very difficult to obtain any explicit description of this region short of calculating $T_{\mu_0,\phi}$ for every point μ_0 on a discretized grid of $\mathbb{R}^K$, which becomes intractable for both computational burden and memory usage as soon as K is large.

10.6.2. FWER versus FDR in multiple testing. It can legitimately be asked if the FWER is in fact an appropriate measure of type I error. Namely, the false discovery rate (FDR), introduced in Benjamini and Hochberg [**BH95**] and defined as the average proportion of wrongly rejected hypotheses among all the rejected hypotheses, appears to have recently become a *de facto* standard, in particular in the setting of a large number of hypotheses to test as we consider here. One reason for the popularity of FDR is that it is a less strict measure of error as the FWER and to this extent, FDR-controlled procedures reject more hypotheses than FWER-controlled ones. We give two reasons why the FWER is still a quantity of interest to investigate. First, the FDR is not always relevant, in particular for neuroimaging data. Indeed, in this context the signal is often strong over some well-known large areas of the brain (*e.g.* the motor and visual cortex). Therefore, if for instance 95 percent of the detected locations belong to these well-known areas, FDR control (at level 5%) does not provide evidence for any new true discovery. On the contrary, FWER control is more conservative, but each detected location outside these well-known areas is a new true discovery with high probability. Secondly, assuming the FDR or a related quantity is nevertheless the endgoal, it can be very useful to consider a two-step procedure, where the first step consists in a FWER-controlled multiple test. Namely, this first step can be used as a means to estimate the FDR or the FDP (false discovery proportion) of another procedure used in the second step and thus fine-tune the parameters of this second step for the desired goal. This approach has been for example advocated by Perone Pacifico *et al.* [**PPGVW04**] with application to neuroimaging data as well.

10.6.3. About the variances of the coordinates. In the concentration approach and in the Gaussian case, the derived thresholds depend explicitly on the p-norm of the vector of standard deviations $\sigma = (\sigma_k)_k$ (an upper bound on this quantity can be used as well). While we have left aside the problem of determining this parameter if no prior information is available, there is at least a simple solution available: build (using standard techniques) an individual upper confidence bound for each σ_k, then combine these different confidence bounds with the Bonferroni method. While this naive method will not take into account the possible dependence between the coordinates for the estimation of σ itself, it will generally only contribute a lower order term in the final threshold defined by (10.7).

A second and potentially more crucial problem is that, since the confidence regions proposed in this paper are balls rather than ellipsoids, these regions will — inevitably — be conservative when the variances of the coordinates are very different. The standard way to address this issue is to consider studentized data. While this would solve this heteroscedasticity issue, it also voids the assumption of independent datapoints – a crucial assumption in all of our proofs. Therefore, generalizing our approach to studentized observations is an important (and probably challenging) direction for future research.

10.6.4. Conclusion. In this chapter, we proposed two approaches to build non-asymptotic resampling-based confidence regions for a correlated random vector:

- The first one is strongly inspired by results coming from learning theory and is based on a concentration argument. An advantage of this method is that it allows to use a very

large class of resampling weights. However, these concentration-based thresholds have relatively conservative deviation terms and they are better than the Bonferroni threshold only if there are very strong correlations in the data. Therefore, using this method when we do not have any prior knowledge on the correlations can be too risky. To address this issue, we propose (under the Gaussian assumption) to combine the corresponding concentration threshold with the Bonferroni threshold to obtain a threshold very close to the minimum of the two (using the so-called "stabilization property" of the resampling).

- The second method is closer to the idea of randomization tests: it estimates directly the quantile of $\phi(\overline{\mathbf{Y}} - \mu)$ using a symmetrization argument (it is therefore restricted to Rademacher weights). The point is that an exact approach is not possible because we have to replace the unknown parameter μ by the empirical mean $\overline{\mathbf{Y}}$. Therefore, the derived thresholds have a remainder term, but it is quite small when n is sufficiently large (typically $n \geq 1\,000$).

Our simulations have shown that for confidence regions in supremum norm, the confidence balls obtained with the second method are better than the regions based on the Bonferroni threshold, when there are important correlations between the coordinates. Moreover, it seems that the quantile threshold without the remainder term is very close to the ideal quantile, so that we may conjecture that the additional term is unnecessary (or at least too large).

Finally, we have used the two previous methods to derive step-down multiple testing procedures that control the FWER when testing simultaneously the means of a (Gaussian) random vector (in the one-sided or two-sided context). Because these procedures use translation-invariant thresholds, the number of iterations in the step-down algorithm is generally small. Moreover, they can outperform Holm's procedure when the coordinates of the observed vector has strong enough correlations. However, these procedures are somewhat too conservative because of the remainder terms (in the quantile approach, the remainder terms arise as a consequence of empirically recentering the data).

In the two-sided context, an exact step-down procedure based on the resampled quantiles of the *uncentered* data is valid and turns out to be more accurate than the above methods (because no remainder term is then necessary). However, this exact method needs generally more iterations in the step-down algorithm. Therefore, we propose to combine our quantile approach with the latter exact method to get a faster procedure with (almost) the same accuracy.

Again, we may conjecture that the step-down procedure using the recentred quantile without the additional term (or at least with a smaller term) still controls the FWER for a fixed n. This would give an accurate procedure in both two-sided and one-sided contexts, and the latter would be faster than the exact step-down procedure in the two-sided context. This is certainly an interesting direction for future work.

10.7. Proofs

10.7.1. Confidence regions using concentration. In this section, we prove all the statements of Sect. 10.2 except computations of resampling weight constants (made in Sect. 10.7.4) and statements with non-exchangeable resampling weights (made in Sect. 10.7.5).

Comparison in expectation.

PROOF OF PROP. 10.2. Denoting by $\mathbf{\Sigma}$ the common covariance matrix of the $\mathbf{Y}^i$, we have $\mathcal{D}(\overline{\mathbf{Y}}_{[W-\overline{W}]}|W) = \mathcal{N}\big(0, (n^{-1}\sum_{i=1}^n (W_i-\overline{W})^2) n^{-1}\mathbf{\Sigma}\big)$, and the result follows because $\mathcal{D}(\overline{\mathbf{Y}}-\mu) = \mathcal{N}(0, n^{-1}\mathbf{\Sigma})$ and ϕ is positive-homogeneous. □

PROOF OF PROP. 10.3. (i). By independence between W and $\mathbf{Y}$, exchangeability of W and the positive homogeneity of ϕ, for every realization of $\mathbf{Y}$ we have:

$$A_W \phi\left(\overline{\mathbf{Y}}-\mu\right) = \phi\left(\mathbb{E}_W\left[\frac{1}{n}\sum_{i=1}^n \left|W_i-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\right]\right) \quad .$$

Then, by convexity of ϕ,

$$A_W \phi\left(\overline{\mathbf{Y}}-\mu\right) \leq \mathbb{E}_W\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left|W_i-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\right)\right] \quad .$$

We integrate with respect to $\mathbf{Y}$, and use the symmetry of the $\mathbf{Y}^i$ with respect to μ and again the independence between W and $\mathbf{Y}$ to show finally that

$$\begin{aligned} A_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] &\leq \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left|W_i-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\right)\right] \\ &= \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left(W_i-\overline{W}\right)\left(\mathbf{Y}^i-\mu\right)\right)\right] = \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] . \end{aligned}$$

(ii) comes from:

$$\begin{aligned} \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{W-\overline{W}}\right) &= \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i-\overline{W})(\mathbf{Y}^i-\mu)\right) \\ &\leq \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i-x_0)(\mathbf{Y}^i-\mu)\right) + \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n (x_0-\overline{W})(\mathbf{Y}^i-\mu)\right) \quad . \end{aligned}$$

Then, by symmetry of the $\mathbf{Y}^i$ with respect to μ and independence between W and $\mathbf{Y}$, we get

$$\begin{aligned} \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{W-\overline{W}}\right) &\leq \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n |W_i-x_0|\left(\mathbf{Y}^i-\mu\right)\right) + \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n \left|x_0-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\right) \\ &\leq \left(a+\mathbb{E}\left|\overline{W}-x_0\right|\right)\mathbb{E}\phi\left(\overline{\mathbf{Y}}-\mu\right) \quad . \end{aligned}$$

□

Concentration inequalities.

PROOF OF PROP. 10.4. We use here concentration principles applied to a supremum of Gaussian random vectors, following closely the approach in Massart [**Mas07**], Sect. 3.2.4. The essential ingredient is the Gaussian concentration theorem of Cirel'son, Ibragimov and Sudakov [**CIS76**] (and recalled in [**Mas07**], Thm. 3.8; see also Thm. 8.1 in Sect. 8.5), stating that if F is a Lipschitz function on $\mathbb{R}^N$ with constant L, then for the standard Gaussian measure on $\mathbb{R}^N$ we have $\mathbb{P}\left[F \geq \mathbb{E}\left[F\right]+t\right] \leq 2\overline{\Phi}(t/L)$.

Let us denote by $\mathbf{A}$ a square root of the common covariance matrix of the $\mathbf{Y}^i$. If $\mathbf{G}$ is a $K\times n$ matrix with standard centered i.i.d. Gaussian entries, then $\mathbf{AG}$ has the same distribution as $\mathbf{Y}-\mu$. We let for all $\zeta \in \left(\mathbb{R}^K\right)^n$, $T_1(\zeta) := \phi\left(\frac{1}{n}\sum_{i=1}^n \mathbf{A}\zeta_i\right)$ and $T_2(\zeta) := \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i-\overline{W})\mathbf{A}\zeta_i\right)\right]$.

From the Gaussian concentration theorem recalled above, to reach the conclusion we just need to prove that T_1 (resp. T_2) is a Lipschitz function with constant $\|\sigma\|_p/\sqrt{n}$ (resp. $\|\sigma\|_p C_W/n$) with

respect to the Euclidean norm $\|\cdot\|_{2,Kn}$ on $\left(\mathbb{R}^K\right)^n$. Let $\zeta, \zeta' \in \left(\mathbb{R}^K\right)^n$ and denote by $(a_k)_{1\leq k\leq K}$ the rows of $\mathbf{A}$. Using that ϕ is 1-Lipschitz with respect to the p-norm (because it is subadditive and bounded by the p-norm), we get

$$\begin{aligned} |T_1(\zeta) - T_1(\zeta')| &\leq \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{A}(\zeta_i - \zeta_i') \right\|_p \\ &\leq \left\| \left(\left\langle a_k, \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\rangle \right)_k \right\|_p . \end{aligned}$$

For each coordinate k, by Cauchy-Schwartz's inequality and since $\|a_k\|_2 = \sigma_k$, we deduce

$$\left| \left\langle a_k, \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\rangle \right| \leq \sigma_k \left\| \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2 .$$

Therefore, we get

$$\left|T_1(\zeta) - T_1(\zeta')\right| \leq \|\sigma\|_p \left\| \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2 \leq \frac{\|\sigma\|_p}{\sqrt{n}} \left\|\zeta - \zeta'\right\|_{2,Kn} \ ,$$

using the convexity of $x \in \mathbb{R}^K \mapsto \|x\|_2^2$, and we obtain (i). For T_2, we use the same method as for T_1:

$$\begin{aligned} \left|T_2(\zeta) - T_2(\zeta')\right| &\leq \|\sigma\|_p \, \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2 \\ &\leq \frac{\|\sigma\|_p}{n} \sqrt{\mathbb{E}\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2^2} \ . \end{aligned} \tag{10.32}$$

Note that since $\left(\sum_{i=1}^n (W_i - \overline{W})\right)^2 = 0$, we have $\mathbb{E}(W_1 - \overline{W})(W_2 - \overline{W}) = -C_W^2/n$. We now develop $\left\|\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i')\right\|_2^2$ in the Euclidean space $\mathbb{R}^K$:

$$\begin{aligned} \mathbb{E}\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2^2 &= C_W^2 \left(1 - n^{-1}\right) \sum_{i=1}^n \left\|\zeta_i - \zeta_i'\right\|_2^2 - \frac{C_W^2}{n} \sum_{i\neq j} \left\langle \zeta_i - \zeta_i', \zeta_j - \zeta_j' \right\rangle \\ &= C_W^2 \sum_{i=1}^n \left\|\zeta_i - \zeta_i'\right\|_2^2 - \frac{C_W^2}{n} \left\| \sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2^2 . \end{aligned}$$

Consequently,

$$\mathbb{E}\left\| \sum_{i=1}^n \left(W_i - \overline{W}\right)\left(\zeta_i - \zeta_i'\right) \right\|_2^2 \leq C_W^2 \sum_{i=1}^n \left\|\zeta_i - \zeta_i'\right\|_2^2 = C_W^2 \left\|\zeta - \zeta'\right\|_{2,Kn}^2 . \tag{10.33}$$

Combining expression (10.32) and (10.33), we find that T_2 is $\|\sigma\|_p C_W/n$-Lipschitz. □

REMARK 10.9. The proof of Proposition 10.4 is still valid under the weaker assumption (instead of exchangeability of W) that $\mathbb{E}\left[(W_i - \overline{W})(W_j - \overline{W})\right]$ can only take two possible values depending on whether or not $i = j$.

Main results.

PROOF OF THM. 10.1. The case (BA)(p, M) and (SA) is obtained by combining Prop. 10.3 and McDiarmid's inequality (Prop. 8.7 in Sect. 8.5). The (GA) case is a straightforward consequence of Prop. 10.2 and the proof of Prop. 10.4. □

PROOF OF COR. 10.1. From Prop. 10.4 (i), with probability at least $1-\alpha(1-\delta)$, $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ is upper bounded by the minimum between $t_{\alpha(1-\delta)}$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right]+\frac{\|\sigma\|_p \overline{\Phi}^{-1}(\alpha(1-\delta)/2)}{\sqrt{n}}$ (because these thresholds are deterministic). In addition, Prop. 10.2 and Prop. 10.4 (ii) give that with probability at least $1-\alpha\delta$, $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] \leq \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right]}{B_W}+\frac{\|\sigma\|_p C_W}{B_W n}\overline{\Phi}^{-1}(\alpha\delta/2)$. The result follows by combining the two last expressions. □

Monte-Carlo approximation.

PROOF OF PROP. 10.5. The idea of the proof is to apply McDiarmid's inequality conditionally to $\mathbf{Y}$. For any realizations W and W' of the resampling weight vector and any $\nu \in \mathbb{R}^k$,

$$\begin{aligned}\left|\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)-\phi\left(\overline{\mathbf{Y}}_{[W'-\overline{W'}]}\right)\right| &\leq \phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}-\overline{\mathbf{Y}}_{[W'-\overline{W'}]}\right)\\ &\leq \frac{b-a}{n}\left\|\left(\sum_{i=1}^n\left|\mathbf{Y}_k^i-\nu_k\right|\right)_k\right\|_p\end{aligned}$$

since ϕ is sub-additive and bounded by the p-norm and $W_i-\overline{W}\in[a;b]$ a.s.

The sample $\mathbf{Y}$ being deterministic, we can take $\nu=M$ which realizes the infimum. Since $W^1,\ldots,W^B$ are independent, McDiarmid's inequality (Prop. 8.7, Sect. 8.5) gives (10.17).

When $\mathbf{Y}$ satisfies (GA), a proof very similar to the one of (10.14) in Prop. 10.4 can be applied to the remainder term with any deterministic ν. We then obtain (10.18). □

REMARK 10.10. When the weights are unbounded, one can replace (10.19) by any upper bound on

$$\sqrt{\frac{\ln(1/(\delta\alpha))}{2B}}\sup_{W,W'}\left\{\phi\left(\overline{\mathbf{Y}}_{[W-W'-\overline{W}+\overline{W'}]}\right)\right\}\ ,$$

which is data-dependent but hard to compute in general.

10.7.2. Quantiles. Remember the following inequality coming from the definition of the quantile q_α : for any fixed $\mathbf{Y}$

$$\mathbb{P}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right)>q_\alpha(\phi,\mathbf{Y})\right]\leq\alpha\leq\mathbb{P}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right)\geq q_\alpha(\phi,\mathbf{Y})\right]\ . \tag{10.34}$$

PROOF OF LEMMA 10.6. We introduce the notation $\mathbf{Y}\bullet W=\mathbf{Y}.\mathrm{diag}(W)$ for the matrix obtained by multiplying the i-th column of $\mathbf{Y}$ by W_i, $i=1,\ldots,n$. We have

$$\begin{aligned}\mathbb{P}_{\mathbf{Y}}\left[\phi(\overline{\mathbf{Y}}-\mu)>q_\alpha(\phi,\mathbf{Y}-\mu)\right] &= \mathbb{E}_W\left[\mathbb{P}_{\mathbf{Y}}\left[\phi\big(\overline{(\mathbf{Y}-\mu)}_{[W]}\big)>q_\alpha(\phi,(\mathbf{Y}-\mu)\bullet W)\right]\right]\\ &= \mathbb{E}_{\mathbf{Y}}\left[\mathbb{P}_W\left[\phi\left(\overline{(\mathbf{Y}-\mu)}_{[W]}\right)>q_\alpha(\phi,\mathbf{Y}-\mu)\right]\right]\leq\alpha\ .\end{aligned} \tag{10.35}$$

The first equality is due to the fact that the distribution of $\mathbf{Y}$ satisfies assumption (SA), hence the distribution of $(\mathbf{Y}-\mu)$ invariant by reweighting by (arbitrary) signs $W\in\{-1,1\}^n$. In the second equality we used Fubini's theorem and the fact that for any arbitrary signs W as above $q_\alpha(\phi,(\mathbf{Y}-\mu)\bullet W)=q_\alpha(\phi,\mathbf{Y}-\mu)$; finally the last inequality comes from (10.34). □

PROOF OF THM. 10.2. Put $\gamma_1=\gamma_1(\alpha_0\delta)$ for short and define the event

$$\Omega=\left\{\mathbf{Y}\text{ s.t. }q_{\alpha_0}(\phi,\mathbf{Y}-\mu)\leq q_{\alpha_0(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}})+\gamma_1 f(\mathbf{Y})\right\}\ .$$

Then we have using (10.35):

$$\begin{aligned}\mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu)>q_{\alpha_0(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}})+\gamma_1 f(\mathbf{Y})\right] &\leq \mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu)>q_{\alpha_0}(\phi,\mathbf{Y}-\mu)\right]+\mathbb{P}\left[\mathbf{Y}\in\Omega^c\right]\\ &\leq \alpha_0+\mathbb{P}\left[\mathbf{Y}\in\Omega^c\right]\ .\end{aligned} \tag{10.36}$$

We now concentrate on the event Ω^c. Using the subadditivity of ϕ, and the fact that $\overline{(\mathbf{Y}-\mu)}_{[W]} = \overline{(\mathbf{Y}-\overline{\mathbf{Y}})}_{[W]} + \overline{W}(\overline{\mathbf{Y}}-\mu)$, we have for any fixed $\mathbf{Y} \in \Omega^c$:

$$\begin{aligned}\alpha_0 &\leq \mathbb{P}_W\left[\phi(\overline{(\mathbf{Y}-\mu)}_{[W]}) \geq q_{\alpha_0}(\phi, \mathbf{Y}-\mu)\right] \leq \mathbb{P}_W\left[\phi(\overline{(\mathbf{Y}-\mu)}_{[W]}) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y})\right] \\ &\leq \mathbb{P}_W\left[\phi(\overline{(\mathbf{Y}-\overline{\mathbf{Y}})}_{[W]}) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y}-\overline{\mathbf{Y}})\right] + \mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right] \\ &\leq \alpha_0(1-\delta) + \mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right].\end{aligned}$$

For the first and last inequalities we have used (10.34), and for the second inequality the definition of Ω^c. From this we deduce that

$$\Omega^c \subset \left\{\mathbf{Y} \text{ s.t. } \mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right] \geq \alpha_0\delta\right\}.$$

Now using the homogeneity of ϕ, and the fact that both ϕ and f are nonnegative:

$$\begin{aligned}\mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right] &= \mathbb{P}_W\left[|\overline{W}| > \frac{\gamma_1 f(\mathbf{Y})}{\phi(\mathrm{sign}(\overline{W})(\overline{\mathbf{Y}}-\mu))}\right] \leq \mathbb{P}_W\left[|\overline{W}| > \frac{\gamma_1 f(\mathbf{Y})}{\widetilde{\phi}(\overline{\mathbf{Y}}-\mu)}\right] \\ &= 2\mathbb{P}_W\left[\frac{1}{n}(2B_{n,\frac{1}{2}} - n) > \frac{\gamma_1 f(\mathbf{Y})}{\widetilde{\phi}(\overline{\mathbf{Y}}-\mu)}\right],\end{aligned}$$

where $B_{n,\frac{1}{2}}$ denotes a binomial $(n, \frac{1}{2})$ variable (independent of $\mathbf{Y}$). From the two last displays and the definition of γ_1, we conclude

$$\Omega^c \subset \left\{\mathbf{Y} \text{ s.t. } \widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > f(\mathbf{Y})\right\},$$

which, put back in (10.36), leads to the desired conclusion. □

Proof of Cor. 10.7. Define the function

$$g_0(\mathbf{Y}) = q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y}-\overline{\mathbf{Y}}) + \left(\sum_{i=1}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y})\right),$$

and for $k = 1, \ldots, J$,

$$g_k(\mathbf{Y}) = \gamma_k^{-1}\left(\sum_{i=k}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y})\right),$$

with the convention $g_J = f$. For $0 \leq k \leq J-1$, applying Thm. 10.2 with the function g_{k+1} yields the relation

$$\mathbb{P}_W\left[\phi(\overline{\mathbf{Y}}-\mu) > g_k(\mathbf{Y})\right] \leq \alpha_k + \mathbb{P}_W\left[\phi(\overline{\mathbf{Y}}-\mu) > g_{k+1}(\mathbf{Y})\right].$$

Therefore,

$$\mathbb{P}_W\left[\phi(\overline{\mathbf{Y}}-\mu) > g_0(\mathbf{Y})\right] \leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > f(\mathbf{Y})\right],$$

as announced. □

Proof of Prop. 10.8. Let us first prove that an analogue of Lemma 10.6 holds with q_{α_0} replaced by $\widetilde{q}_{\alpha_0}$. First, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{W}}\mathbb{P}_{\mathbf{Y}}\left[\phi(\overline{\mathbf{Y}}-\mu) > \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y}-\mu, \mathbf{W})\right] \\ &\qquad = \mathbb{E}_{W'}\mathbb{E}_{\mathbf{W}}\mathbb{P}_{\mathbf{Y}}\left[\phi\big(\overline{(\mathbf{Y}-\mu)}_{[W']}\big) > \widetilde{q}_{\alpha_0}(\phi, (\mathbf{Y}-\mu)\bullet W', \mathbf{W})\right] \\ &\qquad = \mathbb{E}_{\mathbf{Y}}\mathbb{P}_{\mathbf{W},W'}\left[\phi\big(\overline{(\mathbf{Y}-\mu)}_{[W']}\big) > \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y}-\mu, W'\bullet\mathbf{W})\right],\end{aligned}$$

where W' denotes a Rademacher vector independent of all other random variables and $W' \bullet \mathbf{W} = \mathrm{diag}(W').\mathbf{W}$ denotes the matrix obtained by multiplying the i-th row of $\mathbf{W}$ by W'_i, $i = 1, \ldots, n$. Note that $(W', W' \bullet \mathbf{W}) \sim (W', \mathbf{W})$. Therefore, by definition of the quantile $\widetilde{q}_{\alpha_0}$, the latter quantity is equal to

$$\mathbb{E}_{\mathbf{Y}} \mathbb{P}_{\mathbf{W}, W'} \left[\frac{1}{B} \sum_{j=1}^{B} \mathbb{1}_{\phi((\overline{\mathbf{Y}-\mu})_{[\mathbf{W}^j]}) \geq \phi((\overline{\mathbf{Y}-\mu})_{[W']})} \leq \alpha_0 \right] \leq \frac{\lfloor B\alpha_0 \rfloor + 1}{B+1},$$

where the last step comes from Lemma 10.13 taken from Romano and Wolf [**RW05**] (see below).

The rest of the proof is similar to the one of Thm. 10.2, where $\mathbb{P}_W$ is replaced by the empirical distribution based on $\mathbf{W}$, $\widetilde{\mathbb{P}}_W = \frac{1}{B} \sum_{j=1}^{B} \delta_{\mathbf{W}^j}$. Thus, (10.34) becomes for any fixed $\mathbf{Y}, \mathbf{W}$:

$$\widetilde{\mathbb{P}}_W \left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right) > \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y}, \mathbf{W}) \right] \leq \alpha_0 \leq \widetilde{\mathbb{P}}_W \left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right) \geq \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y}, \mathbf{W}) \right].$$

Then, the role of Ω is taken by

$$\widetilde{\Omega} := \left\{ \mathbf{Y}, \mathbf{W} \text{ s.t. } \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, \mathbf{W}) \leq \widetilde{q}_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{W}) + \gamma f(\mathbf{Y}, \mathbf{W}) \right\},$$

where we put $\gamma = \gamma(\mathbf{W}, \alpha_0 \delta)$ for short. We then have similarly to (10.36):

$$\mathbb{P}_{\mathbf{Y}, \mathbf{W}} \left[\phi(\overline{\mathbf{Y}} - \mu) > \widetilde{q}_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma f(\mathbf{Y}, \mathbf{W}) \right] \leq \frac{\lfloor B\alpha_0 \rfloor + 1}{B+1} + \mathbb{P}_{\mathbf{Y}, \mathbf{W}} \left[\widetilde{\Omega}^c \right],$$

and following further the proof of Theorem 10.2, we obtain

$$\widetilde{\Omega}^c \subset \left\{ \mathbf{Y}, \mathbf{W} \,\middle|\, \widetilde{\mathbb{P}}_W \left[|\overline{W}| > \frac{\gamma f(\mathbf{Y}, \mathbf{W})}{\widetilde{\phi}\left(\overline{\mathbf{Y}} - \mu\right)} \right] \geq \alpha_0 \delta \right\},$$

which gives the result. □

We have used the following Lemma:

LEMMA 10.13 (Essentially Lemma 1 of Romano and Wolf [**RW05**]). *Let Z_0, Z_1, ..., Z_B be exchangeable real-valued random variables. Then for all $\alpha \in (0, 1)$,*

$$\mathbb{P} \left[\frac{1}{B} \sum_{j=1}^{B} \mathbb{1}_{Z_j \geq Z_0} \leq \alpha \right] \leq \frac{\lfloor B\alpha \rfloor + 1}{B+1} \leq \alpha + \frac{1}{B+1}.$$

The first inequality becomes an equality if $Z_i \neq Z_j$ a.s. For example, it is the case if the Z_is are i.i.d. variables from a distribution without atoms.

We provide a proof for completeness.

PROOF OF LEMMA 10.13. Let U denote a random variable uniformly distributed in $\{0, \ldots, B\}$ and independent of the Z_is. We then have

$$\begin{aligned}
\mathbb{P} \left[\frac{1}{B} \sum_{j=1}^{B} \mathbb{1}_{Z_j \geq Z_0} \leq \alpha \right] &= \mathbb{P} \left[\sum_{j=0}^{B} \mathbb{1}_{Z_j \geq Z_0} \leq B\alpha + 1 \right] \\
&= \mathbb{P}_U \mathbb{P}_{(Z_i)} \left[\sum_{j=0}^{B} \mathbb{1}_{Z_j \geq Z_U} \leq B\alpha + 1 \right] \\
&= \mathbb{P}_{(Z_i)} \mathbb{P}_U \left[\sum_{j=0}^{B} \mathbb{1}_{Z_j \geq Z_U} \leq \lfloor B\alpha \rfloor + 1 \right] \leq \frac{\lfloor B\alpha \rfloor + 1}{B+1}.
\end{aligned}$$

Note that the last inequality is an equality if the Z_is are a.s. distinct. □

10.7.3. Multiple testing.

PROOF OF THM. 10.3, FROM ROMANO AND WOLF [**RW05**]. We use the notations of Def. 10.1. If the procedure rejects at least one true null hypothesis, we may consider $j_0 = \min\{j \leq \hat{\ell} \text{ s.t. } H_{\sigma(j)} \text{ is true }\}$. By definition of a step-down procedure, we have $[\overline{\mathbf{Y}}_{\sigma(j_0)}] \geq t_{j_0}$. By definition of j_0, we have $\mathcal{H}_0 \subset \mathcal{C}_{j_0}$ so that, since $\mathbf{t}$ is non-decreasing, $\mathbf{t}(\mathcal{C}_{j_0}) \geq \mathbf{t}(\mathcal{H}_0)$. Finally, we can obtain (10.27) as follows:

$$\begin{aligned}\mathrm{FWER}(R) &\leq \mathbb{P}\left(\exists j_0 \text{ s.t. } H_{\sigma(j_0)} \text{ is true and } \left[\overline{\mathbf{Y}}_{\sigma(j_0)}\right] \geq \mathbf{t}(\mathcal{H}_0)\right) \\ &\leq \mathbb{P}\left(T'(H_0) \geq \mathbf{t}(\mathcal{H}_0)\right) \\ &\leq \mathbb{P}\left(T(H_0) \geq \mathbf{t}(\mathcal{H}_0)\right) \ .\end{aligned}$$

□

PROOF OF PROP. 10.12. First note that

$$q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right) \leq q_{\alpha(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \mu) \ .$$

Recall that from the proof of Thm. 10.2, with probability larger than $1 - \alpha\gamma$ we have

$$q_{\alpha(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \mu) \leq q_{\alpha(1-\delta)(1-\gamma)}\left(\|\cdot\|_\infty, \mathbf{Y} - \overline{\mathbf{Y}}\right) + \varepsilon'(\alpha, \delta, \gamma, n, K) \ .$$

Take $\mathbf{Y}$ in the previous event, where the above inequality holds. If the global procedure rejects at least one true null hypothesis, we note j_0 the first time that this occurs ($j_0 = 0$ if it is in the first step). There are two cases:

- if $j_0 = 0$ then we have
$$T(\mathcal{H}_0) \geq q_{\alpha(1-\delta)(1-\gamma)}\left(\|\cdot\|_\infty, \mathbf{Y} - \overline{\mathbf{Y}}\right) + \varepsilon'(\alpha, \delta, \gamma, n, K) \geq q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right)$$
- if $j_0 \geq 1$, all the null hypotheses rejected at the first step are false. Following the proof of Thm. 10.3, $T(\mathcal{H}_0) \geq q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right)$.

In both cases, $T(\mathcal{H}_0) \geq q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right)$, which occurs with probability smaller than $\alpha(1-\gamma)$. □

10.7.4. Exchangeable resampling computations. In this section, we compute constants A_W, B_W, C_W and D_W (defined by (10.3) to (10.6)) for some exchangeable resamplings. This implies all the statements in Tab. 10.1. We first define several additional exchangeable resampling weights:

- **Bernoulli** (p), $p \in (0;1)$: pW_i i.i.d. with a Bernoulli distribution of parameter p. A classical choice is $p = \frac{1}{2}$.
- **Efron** (q), $q \in \{1\ldots,n\}$: $qn^{-1}W$ has a multinomial distribution with parameters $(q; n^{-1}, \ldots, n^{-1})$. A classical choice is $q = n$.
- **Poisson** (μ), $\mu \in (0;+\infty)$: μW_i i.i.d. with a Poisson distribution of parameter μ. A classical choice is $\mu = 1$.

Notice that $\overline{Y}_{[W-\overline{W}]}$ and all the resampling constants are invariant under translation of the weights, so that Bernoulli (1/2) weights are completely equivalent to Rademacher weights in this chapter.

LEMMA 10.14. (1) *Let W be Bernoulli (p) weights with $p \in (0,1)$. Then,*

$$2(1-p) - \sqrt{\frac{1-p}{pn}} \le A_W \le B_W \le \sqrt{\frac{1}{p}-1}\sqrt{1-\frac{1}{n}}$$
$$C_W = \sqrt{\frac{1}{p}-1} \qquad \text{and} \qquad D_W \le \frac{1}{2p} + \left|\frac{1}{2p}-1\right| + \sqrt{\frac{1-p}{np}} \ .$$

(2) *Let W be Efron (q) weights with $q \in \{1,\dots,n\}$. Then,*

$$A_W \le B_W \le \sqrt{\frac{n-1}{n}} \qquad \text{and} \qquad C_W = 1 \ .$$

Moreover, if $q \le n$,

$$A_W = 2\left(1-\frac{1}{n}\right)^q \ .$$

(3) *Let W be Poisson (μ) weights with $\mu > 0$. Then,*

$$A_W \le B_W \le \frac{1}{\sqrt{\mu}}\sqrt{1-\frac{1}{n}} \qquad \text{and} \qquad C_W = \frac{1}{\sqrt{\mu}} \ .$$

Moreover, if $\mu = 1$,

$$\frac{2}{e} - \frac{1}{\sqrt{n}} \le A_W \ .$$

(4) *Let W be Random hold-out (q) weights with $q \in \{1,\dots,n\}$. Then,*

$$A_W = 2\left(1-\frac{q}{n}\right) \qquad B_W = \sqrt{\frac{n}{q}-1}$$
$$C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q}-1} \qquad \text{and} \qquad D_W = \frac{n}{2q} + \left|1-\frac{n}{2q}\right| \ .$$

PROOF OF LEMMA 10.14.

General case. We first only assume that W is exchangeable. Then, from the concavity of $\sqrt{\cdot}$ and the triangular inequality, we have

$$\begin{aligned}\mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \sqrt{\mathbb{E}\left(\overline{W} - \mathbb{E}[W_1]\right)^2} &\le \mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \mathbb{E}\left|\overline{W} - \mathbb{E}[W_1]\right| \\ &\le A_W \le B_W \le \sqrt{\frac{n-1}{n}} C_W \ . \end{aligned} \tag{10.37}$$

Independent weights. We now assume that the W_i are i.i.d. Then,

$$\mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \frac{\sqrt{\mathrm{var}(W_1)}}{\sqrt{n}} \le A_W \qquad \text{and} \qquad C_W = \sqrt{\mathrm{var}(W_1)} \ . \tag{10.38}$$

Bernoulli. These weights are i.i.d. with $\mathrm{var}(W_1) = p^{-1} - 1$, $\mathbb{E}[W_1] = 1$ and

$$\mathbb{E}\left|W_1 - 1\right| = p\left(p^{-1}-1\right) + (1-p) = 2(1-p) \ .$$

With (10.37) and (10.38), we obtain the bounds for A_W, B_W and C_W. Moreover, Bernoulli (p) weights satisfy the assumption of (10.6) with $x_0 = a = (2p)^{-1}$. Then,

$$D_W = \frac{1}{2p} + \mathbb{E}\left|\overline{W} - \frac{1}{2p}\right| \le \frac{1}{2p} + \left|1-\frac{1}{2p}\right| + \mathbb{E}\left|\overline{W} - 1\right| \le \frac{1}{2p} + \frac{1}{p}\left|\frac{1}{2} - p\right| + \sqrt{\frac{1-p}{np}} \ .$$

Efron. We have $\overline{W} = 1$ a.s. so that

$$C_W = \sqrt{\frac{n}{n-1}}\,\mathrm{var}(W_1) = 1 \ .$$

If moreover $q \leq n$, then $W_i < 1$ implies $W_i = 0$ and

$$\begin{aligned} A_W &= \mathbb{E}\,|W_1 - 1| = \mathbb{E}\,[W_1 - 1 + 2\mathbb{1}_{W_1=0}] \\ &= 2\mathbb{P}(W_1 = 0) = 2\left(1 - \frac{1}{n}\right)^q \ . \end{aligned}$$

The result follows from (10.37).

Poisson. These weights are i.i.d. with $\mathrm{var}(W_1) = \mu^{-1}$, $\mathbb{E}[W_1] = 1$. Moreover, if $\mu \leq 1$, $W_i < 1$ implies $W_i = 0$ and

$$\mathbb{E}\,|W_1 - 1| = 2\mathbb{P}(W_1 = 0) = 2e^{-\mu} \ .$$

With (10.37) and (10.38), the result follows.

Random hold-out. These weights are such that $\{W_i\}_{1\leq i\leq n}$ is deterministic, with $\overline{W} = 1$. Then, A_W, B_W and C_W can be directly computed. Moreover, they satisfy the assumption of (10.6) with $x_0 = a = n/(2q)$. The computation of D_W is straightforward. □

10.7.5. Non-exchangeable weights. In Sect. 10.2.5, we consider non-exchangeable weights in order to reduce the complexity of computation of expectations w.r.t. the resampling randomness. Then, we are mainly interested in non-exchangeable weights with small support. This is why we focus on the two following cases:

(1) deterministic weights
(2) V-fold weights ($V \in \{2, \ldots, n\}$): let $(B_j)_{1\leq j\leq V}$ be a partition of $\{1, \ldots, n\}$ and $W^B \in \mathbb{R}^V$ an exchangeable resampling weight vector of size V. Then, for any $i \in \{1, \ldots, n\}$ with $i \in B_j$, define $W_i = W^B_j$.

We will often assume that the partition $(B_j)_{1\leq j\leq V}$ is "regular", *i.e.* that V divides n and $\mathrm{Card}(B_j) = n/V$ for every $j \in \{1, \ldots, V\}$. When V does not divide n, the B_j can be chosen approximatively of the same size. Remember that we always assume in this paper that W is independent from the data $\mathbf{Y}$, even in the non-exchangeable case.

In the following, we make use of five constants that depend only on the resampling scheme: B_W and D_W stay unchanged (see definitions (10.4) and (10.6)), we modify the definitions of A_W and C_W (notice that we stay consistent with (10.3) and (10.5) when W is exchangeable), and we introduce a fifth constant E_W (which is equal to A_W in the exchangeable case):

$$A_W := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left|W_i - \overline{W}\right| \tag{10.39}$$

$$C_W := \sqrt{n} B_W \qquad \text{if } W \text{ is deterministic} \tag{10.40}$$

$$C_W := \sqrt{\max_j \mathrm{Card}(B_j)}\, C_{W^B} + \sqrt{n}\mathbb{E}\left|\overline{W^B} - \overline{W}\right| \qquad \text{if } W \text{ is } V\text{-fold} \tag{10.41}$$

$$E_W := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}|W_i - \overline{W})|\right)^2} \ . \tag{10.42}$$

We can now state the main theorem of this section.

THEOREM 10.4. *Let W be either a deterministic or V-fold resampling weight vector, and define the constants A_W, B_W, C_W, D_W and E_W by* (10.39), (10.4), (10.40), (10.41), (10.6) *and* (10.42).

Then, all the results of Thm. 10.1 and Cor. 10.1 hold, with only a slight modification in (10.8)*:*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]}{A_W} + \frac{M}{\sqrt{n}}\sqrt{1+\frac{A_W^2}{E_W^2}}\sqrt{2\log(1/\alpha)} \ .$$

PROOF OF THM. 10.4. In the Gaussian case, we use the same proof as Thm. 10.1 and Cor. 10.1, but we replace the concentration result (10.15) by the one of Prop. 10.15.

In the bounded case, the proof is identical (it relies on McDiarmid's inequality), but we no longer have $A_W = E_W$ because the weights are non-exchangeable. □

When V divides n, we can compute the constants for regular V-fold weights:

$$A_W = E_W = A_{W^B} \qquad B_W = B_{W^B} \qquad C_W = \sqrt{\frac{n}{V}} C_{W^B} \ .$$

We now give two natural examples of non-exchangeable weights:

(1) **Hold-out** (q): $W_i = \frac{n}{q}\mathbb{1}_{i\in I}$ for some deterministic subset $I \subset \{1,\dots,n\}$ of cardinality q. A classical choice is $q = \lfloor n/2 \rfloor$.

(2) **V-fold cross validation**, $V \in \{2,\dots,n\}$: V-fold weights with W^B leave-one-out (which is often called cross-validation). More precisely, $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$, J uniform on $\{1,\dots,V\}$), $(B_j)_{1\le j\le V}$ partition of $\{1,\dots,n\}$.

The terms "hold-out", "cross-validation" and "V-fold cross-validation" refer to slightly different procedures which inspired these weights. In these two cases, we can compute the resampling constants:

(1) **Hold-out** (q):

$$A_W = 2\left(1-\frac{q}{n}\right) \qquad B_W = E_W = \sqrt{\frac{n}{q}-1}$$

$$C_W = \sqrt{n\left(\frac{n}{q}-1\right)} \qquad \text{and} \qquad D_W = \frac{n}{2q} + \left|1-\frac{n}{2q}\right| \ .$$

(2) **V-fold cross validation** (possibly non-regular):

$$A_W = \frac{2}{V-1}\sum_{j=1}^{V}\frac{\mathrm{Card}(B_j)}{n}\left(1-\frac{\mathrm{Card}(B_j)}{n}\right)$$

$$B_W = \frac{1}{V-1}\sum_{j=1}^{V}\sqrt{\frac{\mathrm{Card}(B_j)}{n}\left(1-\frac{\mathrm{Card}(B_j)}{n}\right)}$$

$$C_W = \sqrt{\max_j \mathrm{Card}(B_j)}\frac{\sqrt{V}}{V-1} + \frac{\sqrt{n}}{V-1}\sum_{j=1}^{V}\left|\frac{\mathrm{Card}(B_j)}{n}-\frac{1}{V}\right|$$

$$D_W = \frac{1}{V-1}\sum_{j=1}^{V}\left(\frac{1}{2}+\left|\frac{1}{2}-\frac{\mathrm{Card}(B_j)}{n}\right|\right)$$

$$E_W = \frac{2}{V-1}\sqrt{\sum_{j=1}^{V}\frac{\mathrm{Card}(B_j)}{n}\left(1-\frac{\mathrm{Card}(B_j)}{n}\right)^2} \ .$$

When the partition $(B_j)_{1\leq j\leq V}$ is almost regular, *i.e.* $\max_j \left|\mathrm{Card}(B_j) - nV^{-1}\right| \leq 1$ and $n \gg V \geq 3$, then $C_W B_W^{-1} \leq \sqrt{n/(V-1)}\,(1+o(1))$ which is close to its value in the "regular" case. This means that the concentration thresholds behave as in the regular case provided that n is large enough. The proofs of these results are given at the end of this section. Before this, we give analogues of the results of Sect. 10.2.2 and 10.2.3 in the non-exchangeable case.

Expectations. Although we stated the results of Sect. 10.2.2 with exchangeable weights, the proofs of Prop. 10.2 and 10.3 remain unchanged with non-exchangeable weights, with A_W defined by (10.39). Moreover, Lemma 8.4 in Sect. 8.4.1 shows how to generalize any result on expectations from exchangeable to non-exchangeable weights.

Concentration inequalities. Whereas Prop. 10.4 deals only with exchangeable weights, we can derive a similar result for deterministic and V-fold exchangeable weights. This is the object of the following result.

PROPOSITION 10.15. *Let $p \in [1,+\infty]$, $\mathbf{Y}$ a sample satisfying (GA) and $\phi : \mathbb{R}^K \to \mathbb{R}$ any subadditive function, bounded by the p-norm. Let W be some resampling weight vector among*

(i) *Deterministic weights.*
(ii) *V-fold exchangeable resampling weight for some $V \in \{2,\dots,n\}$.*

Then, for all $\alpha \in (0,1)$, (10.15) and the corresponding lower bound hold with C_W defined by (10.40) (deterministic case) or (10.41) (V-fold case).

PROOF OF PROP. 10.15.

(i) Deterministic weights. We can use (10.14) and the corresponding lower bound with $B_W\sigma$ instead of σ since

$$\overline{\mathbf{Y}}_{[W-\overline{W}]} = \frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)\mathbf{Y}^i \overset{(d)}{=} B_W(\overline{\mathbf{Y}} - \mu) \ .$$

The result follows with $C_W = \sqrt{n}B_W$.

(ii) V-fold weights. The proof is widely inspired from the one of Prop. 10.4. We have to compute the Lipschitz constant of T_2 defined by

$$T_2(\zeta) = \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)\mathbf{A}\zeta_i\right) \ .$$

For all $\zeta, \zeta' \in \mathbb{R}^K$, we use the triangular inequality and the same arguments as in the proof of Prop. 10.4:

$$\begin{aligned}
\left|T_2(\zeta) - T_2(\zeta')\right| &\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)\mathbf{A}(\zeta_i - \zeta_i')\right\|_p \\
&\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W^B}\right)\mathbf{A}(\zeta_i - \zeta_i')\right\|_p + \mathbb{E}\left|\overline{W^B} - \overline{W}\right|\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{A}(\zeta_i - \zeta_i')\right\|_p \\
&\leq \frac{\|\sigma\|_p}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n (W_i - \overline{W^B})(\zeta_i - \zeta_i')\right\|_2^2} + \mathbb{E}\left|\overline{W^B} - \overline{W}\right|\frac{\|\sigma\|_p}{\sqrt{n}}\left\|\zeta - \zeta'\right\|_{2,Kn}
\end{aligned}$$

Using the exchangeability of the W^B, we show that

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{n}(W_i-\overline{W^B})(\zeta_i-\zeta_i')\right\|_2^2 &= \mathbb{E}\left\|\sum_{j=1}^{V}\left(W_j^B-\overline{W^B}\right)\sum_{i\in B_j}\left(\zeta_i-\zeta_i'\right)\right\|_2^2\\ &\leq C_{W^B}^2\sum_{j=1}^{V}\left\|\sum_{i\in B_j}\left(\zeta_i-\zeta_i'\right)\right\|_2^2\\ &\leq C_{W^B}^2\sum_{j=1}^{V}\mathrm{Card}(B_j)\sum_{i\in B_j}\left\|\zeta_i-\zeta_i'\right\|_2^2\end{aligned}$$

by convexity of $\|\cdot\|_2^2$. Finally, this implies that T_2 is Lipschitz of parameter

$$\frac{\|\sigma\|_p}{n}\sqrt{\max_j \mathrm{Card}(B_j)}C_{W^B}+\frac{\|\sigma\|_p}{\sqrt{n}}\mathbb{E}\left|\overline{W^B}-\overline{W}\right| \quad .$$

□

Computation of the constants. We first remark that the following statements are straightforward:

- if W is deterministic, $B_W = E_W$.
- if W is regular V-fold exchangeable,

$$A_W = E_W = A_{W^B} \qquad B_W = B_{W^B} \qquad C_W = \sqrt{\frac{n}{V}}C_{W^B}.$$

In the hold-out (q) case, we compute A_W, B_W and D_W exactly as in the Random hold-out (q) case.

In the general V-fold cross-validation case, we use the following trick: conditionally to the index J of the removed block, W is a deterministic hold-out $(n-\mathrm{Card}(B_J))$ weight multiplied by a factor $c(J)=\frac{V(n-\mathrm{Card}(B_J))}{(V-1)n}$. This allows to compute A_W, B_W and D_W from the hold-out case: for instance,

$$\begin{aligned}A_W &= \frac{1}{V}\sum_{j=1}^{V}\left[2c(J)\left(1-\frac{q}{n}\right)\right]\\ &= \frac{2}{V-1}\sum_{j=1}^{V}\frac{\mathrm{Card}(B_j)}{n}\left(1-\frac{\mathrm{Card}(B_j)}{n}\right) \quad .\end{aligned}$$

This also shows

$$\mathbb{E}\left|\overline{W^B}-\overline{B}\right| = \frac{1}{V}\sum_{j=1}^{V}\left|\frac{V}{V-1}\frac{n-\mathrm{Card}(B_J)}{n}-1\right|$$

from which we obtain C_W. The computation of E_W is done directly by noting that

$$\mathbb{E}\left|W_j^B-\overline{W}\right| = \frac{V}{V-1}\mathbb{E}\left|\mathbb{1}_{j\neq J}-1+\frac{\mathrm{Card}(B_J)}{n}\right| = \frac{2}{V-1}\left(1-\frac{\mathrm{Card}(B_j)}{n}\right) \quad ,$$

$$E_W^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\left|W_i-\overline{W}\right|)^2 = \sum_{j=1}^{V}\frac{\mathrm{Card}(B_j)}{n}(\mathbb{E}\left|W_j^B-\overline{W}\right|)^2$$

$$E_W^2 = \left(\frac{2}{V-1}\right)^2 \sum_{j=1}^{V} \frac{\mathrm{Card}(B_j)}{n}\left(1-\frac{\mathrm{Card}(B_j)}{n}\right)^2 \ .$$

We now prove the last statement about "almost regular" VFCV: when $\max_j \mathrm{Card}(B_j) \leq nV^{-1}+1$,

$$\begin{aligned} C_W &\leq \sqrt{\frac{n}{V}+1}\frac{\sqrt{V}}{V-1} + \frac{V\sqrt{n}}{n(V-1)} \\ &\leq \frac{\sqrt{n}}{V-1}\left(1+\sqrt{\frac{V}{n}}+\frac{V}{n}\right) \ . \end{aligned}$$

If moreover $V^{-1}+n^{-1} \leq 1/2$ and the partition is almost regular, we have:

$$\begin{aligned} B_W &\geq \frac{V}{V-1}\sqrt{\left(\frac{1}{V}-\frac{1}{n}\right)\left(1-\frac{1}{V}+\frac{1}{n}\right)} \\ &= \frac{1}{\sqrt{V-1}}\sqrt{1+\frac{V^2}{(V-1)n}\left(\frac{2}{V}-1-\frac{1}{n}\right)} \\ &\geq \frac{1}{\sqrt{V-1}} - \frac{V}{(V-1)\sqrt{n}}\sqrt{\left(1+\frac{1}{n}-\frac{2}{V}\right)_+} \ . \end{aligned}$$

CHAPTER 11

Conclusions, open problems and prospects

> But my father's mind took unfortunately a wrong turn in the investigation; running, like the hypercritick's, altogether upon the ringing of the bell and the rap upon the door,–measuring their distance, and keeping his mind so intent upon the operation, as to have power to think of nothing else,– common-place infirmity of the greatest mathematicians! working with might and main at the demonstration, and so wasting all their strength upon it, that they have none left in them to draw the corollary, to do good with.
>
> *The Life and Opinions of Tristram Shandy, Gentleman*, Chapter 1, XXXV
> Laurence Sterne

Résumé. En guise de conclusion, nous revenons dans ce chapitre sur plusieurs des avancées principales de ce travail de thèse. Envisageant quatre points de vue distincts, nous tentons — pour ne pas suivre l'exemple de Walter Shandy — de dégager les principales conséquences des résultats des chapitres précédents. Nous proposons également une vingtaine de problèmes ouverts suggérés par ces mêmes résultats. Les quatre points de vues considérés sont les suivants : l'intérêt du rééchantillonnage en général et des diverses façons de rééchantillonner entre elles ; l'étude non-asymptotique de processus empiriques rééchantillonnés ; la calibration optimale de méthode de sélection de modèles par pénalisation ; les régions de confiance et les tests multiples.

In this chapter, we sum up some of the contributions of this thesis, and several open problems that are raised by our results. To do this, we propose four different approaches: resampling theory from the statistical (Sect. 11.1) and probabilistic (Sect. 11.2) viewpoints, accurate calibration of penalties in practice (Sect. 11.3) and confidence regions and multiple testing (Sect. 11.4). Notice that some of the open problems below are closely linked together.

11.1. Why should resampling be used?

11.1.1. A general purpose device. Discussing a paper of Wu [**Wu86**], Efron recalls the main difference between resampling and *ad hoc* procedures:

> "The jackknife and bootstrap are general-purpose devices, not specifically adapted to take advantage of a special model like
>
> $$y = X\beta + e, \qquad \mathrm{var}(e) = \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2) \ .$$
>
> Comparisons with specially adapted methods (...) are misleading if this is not made clear."

Indeed, the resampling heuristics is so general that it can be applied in most of the frameworks, and it often works (unfortunately, not always, the reason why a theoretical study of resampling

is useful). When comparing resampling procedures to other ones in a particular framework, one has to accept a small loss in performance as the price for generality.

For instance, for model selection in least-square regression, linear penalties like Mallows' C_p are much easier to compute and perform better for some "easy" problems (homoscedastic data with s smooth). The interest of Resampling and V-fold penalties (resp. RP and penVF) is that they are robust to heteroscedasticity, contrary to linear penalties (Chap. 4). Of course, one could design a special heteroscedastic penalty, adapted from Mallows' C_p (*e.g.* based upon an estimate of σ by a piecewise constant function on $\mathcal{X}$, with a bin size allowed to go to zero). This *ad hoc* procedure may outperform RP, but it is less general and robust.

For confidence regions and multiple testing, the resampling procedures defined in Chap. 10 need much longer computations than Bonferroni's ((10.11), which assumes that each coordinate is gaussian), but they are far less conservative when the data have correlated coordinates. Moreover, the computational simplicity of (10.11) comes from the gaussian assumption, whereas resampling (either used in each single test with a final Bonferroni correction, or as we do in Chap. 10) are more robust to non-gaussianity.

In both cases, resampling procedures seem to be more robust than *ad hoc* ones. Considering that real data may contain difficulties which have not been imagined by the statistician, our conclusion would be: use a well-calibrated *ad hoc* procedure if you trust your strong assumptions, prefer resampling procedures if you are more cautious.

11.1.2. A naturally adaptive device. Our results show that resampling naturally adapts to some features of the unknown distribution P. In least-square regression on histograms, penVF and RP are adaptive to the smoothness of s even if the noise is heteroscedastic (Chap. 5 and 6, in particular Thm. 6.2). Moreover, results about RP are proven under several quite general sets of assumptions (*cf.* Sect. 6.4.1 and 8.3). The main points remaining are the following:

> OPEN PROBLEM 1. V-fold and Resampling penalties are adaptive to the smoothness of s when it is α-hölder with $\alpha > 1$ (this would need oracle inequalities for piecewise polynomial models instead of histograms).

> OPEN PROBLEM 2. When s is α-hölder with $0 < \alpha < 1$, V-fold and Resampling penalties are adaptive to heteroscedasticity of the noise (*i.e.* extend **[GP05]** to $\alpha < 1$ and any dimension $k > 1$, since Thm. 6.2 already provides an upper bound on the risk).

In binary classification, Fromont's results **[Fro07]** show that global resampling penalties are minimax adaptive from the global viewpoint. Since penVF and RP are local penalties, we can conjecture the following.

> OPEN PROBLEM 3. In binary classification, V-fold and Resampling penalties are margin adaptive.

We provide partial arguments in this direction in Chap. 7, together with a detailed way towards a complete proof (in particular Sect. 7.3.3).

In Chap. 10, our simulation study suggests that the concentration and quantile thresholds adapt to the unknown correlation matrix of the data, at least when they are gaussian:

> OPEN PROBLEM 4. The confidence regions built in Chap. 10 are adaptive to the unknown correlation matrix Σ, *i.e.* they attain the minimax separation rate for every given Σ, without using the knowledge of Σ.

In Sect. 8.2, we show that adaptation of resampling procedures may come from structural reasons. Indeed, resampling estimates $F(P, P_n)$ by $\mathbb{E}_W\left[F(P_n, P_n^W)\right]$, which has automatically the same structure through F. We can wonder whether this is a more general phenomenon:

> OPEN PROBLEM 5. Is there a link between the structure of F and the properties to which resampling adapts when estimating $F(P, P_n)$?

11.1.3. Choice of the resampling scheme. For both RP (Chap. 6, and Chap. 5 for penVF) and concentration-based thresholds (Sect. 10.2), we provide a unified vision of many resampling weights, including all exchangeable weights and V-fold ones. In both frameworks, it appears that all the resampling schemes have the same estimation properties, *i.e.* there exists a multiplicative constant $C_{W,\infty}$ or B_W^{-1} (depending only on $\mathcal{L}(W)$) making them unbiased (at first order). In particular, Chap. 6 enlightens Shao's results [**Sha96**] about identification with Efron (m) penalties: when they are not multiplied by $C_{W,\infty}$, resampling penalties are consistent (for identification) when $C_{W,\infty} \to 0$. This should be compared to the ratio between AIC (asymptotically optimal for prediction) and BIC (consistent for identification). Our conclusion is that *one can* a priori *use any resampling scheme for either prediction or identification, up to an adapted choice of a multiplicative constant.*

As a consequence, there remains three ways for comparing resampling schemes:

- *computational complexity*: it is large with exchangeable weights (the leave-one-out being optimal), much smaller with V-fold weights or Monte-Carlo approximations with exchangeable weights.
- *variability of the estimates*: either measured by the variance of the resampling penalty, or by the remainder term in the concentration-based thresholds (*i.e.* $C_W B_W^{-1}$, see Sect. 10.2.4; see also Prop. 10.5 for the variability of Monte-Carlo approximations).
- *bias at second-order*: Sect. 6.6.1 shows that Efron's bootstrap penalties are slightly biased downwards, whereas Rad and Rho slightly biased upwards, and Loo is the more accurate.

The second point, which seems to be quite important according to our simulation studies, needs more theoretical results. We give in Sect. 10.2 some upper bounds for the concentration-based thresholds. However, in the case of penVF and RP, we are neither able to compare the variability of the resampling schemes nor to quantify the loss of variability induced by a Monte-Carlo approximation. We consider this as a major issue, since theory should at least be able to distinguish[1] between the hold-out (highly variable in practice) and the exchangeable resampling schemes (far less variable). Moreover, this would allow to quantify more precisely the accuracy-complexity trade-off involved for choosing V, as we do in Sect. 10.2.5 for concentration-based thresholds. Some results in that sense have also been proven by Celisse and Robin [**CR06**], in the density estimation framework.

> OPEN PROBLEM 6. Provide a sharp[2] non-asymptotic theoretical account for the variability of $\mathbb{E}_W\left[F(P_n, P_n^W)\right]$ according to the distribution of W (at least, in the case of RP and penVF, distinguish exchangeable, V-fold and deterministic weights).

The difficulty of this problem is highlighted by a negative result of Bengio and Grandvalet [**BG04**]: "there exists no universal (valid under all distributions) unbiased estimator of the variance of V-fold cross-validation".

[1] In the case of a particular randomized algorithm, Blum, Kalai and Langford [**BKL99**] proved that V-fold cross-validation is more efficient than hold-out. However, this does not apply to the prediction error *sensu stricto*.

[2] that is, prove both lower and upper bounds on the variability, since upper bounds alone can be misleading.

OPEN PROBLEM 7. Provide a non-asymptotic theoretical estimate of the loss of accuracy when using a Monte-Carlo approximation instead of an exact computation with exchangeable weights. For instance, generalize Prop. 10.5 to more general resampling estimates, including RP.

As a first step, we could focus on the least-square regression case. Then, keep in mind that the picture may be different in classification for instance, since empirical risk minimization algorithms can be less stable. In particular, the leave-one-out should appear more stable with least-square regression than it is with unstable algorithms.

11.2. Advances in the non-asymptotic theory of resampling

As we have noticed in Introduction, the non-asymptotic theory of resampling is still far less developped than the asymptotical one (in particular in Chap. 3.6 of the book of van der Vaart and Wellner [**vdVW96**]). Though, the non-asymptotic understanding of resampling is crucial: the resampling heuristics being non-asymptotic, it should give better procedures than the usual asymptotical approximations. This conjecture is both supported by Edgeworth expansions (Hall [**Hal92**]), large deviations (Hall [**Hal92**], Appendix 5) and simulation studies, but there is still a gap between non-asymptotic theory and practice.

In this thesis, our non-asymptotical approach allows us to deal with two particular statistical frameworks of interest:

- model selection when the family of models $\mathcal{M}_n$ depends on n, has a size possibly much greater than n, and contains models of dimension going to infinity with n (Chap. 2 to 9).
- confidence regions when the data belongs to $\mathbb{R}^K$ with $K \gg n$ (Chap. 10).

11.2.1. Concentration results. We prove several non-asymptotic concentration results on resampling in this thesis. According to the way they are proven, we can split them into four categories:

- exact computations in the histogram case, combined with moment inequalities (Boucheron, Bousquet, Lugosi and Massart [**BBLM05**]): Prop. 5.10 in Sect. 5.7.4, Prop. 6.8 in Sect. 6.8.7, Prop. 8.5 in Sect. 8.4.2.
- in a general framework, a moment inequality for p_2 by Boucheron and Massart [**BM04**], with a restriction to subsampling weights: Thm. 7.1 in Sect. 7.3.2.
- Gaussian concentration theorem (Thm. 8.1 in Sect. 8.5): Prop. 10.4 in Sect. 10.2.3.
- McDiarmid's inequality (Prop. 8.7 in Sect. 8.5): (10.8) and (10.9) in Thm. 10.1, Sect. 10.2.3. Prop. 1 and 2 of Fromont [**Fro07**] rely on the same inequality.

However, they are not completely satisfactory, because of their lack of generality (the first three assuming either a particular framework, a particular kind of resampling or a particular distribution for the noise), or because they are not accurate enough (McDiarmid's inequality is too rough to attain fast rates). This is why the following generalizations of our results would be quite interesting:

OPEN PROBLEM 8. Generalize Thm. 7.1 to exchangeable weights, or at least to Efron (m) weights.

OPEN PROBLEM 9. Prove a non-asymptotic concentration inequality for $\widehat{p}_1$ in a general framework (*e.g.* the one of Thm. 7.1).

In the case of subsampling, Problem 9 can be derived as Thm. 7.1 from a moment inequality on p_1 similar to the result of Boucheron and Massart [**BM04**] (see Problem 14). In the case of general weights, this may be related to Problem 8.

OPEN PROBLEM 10. Generalize Prop. 10.4 to sub-gaussian variables (or, more generally, under moment assumptions).

In order to obtain such general results, several techniques can be considered: moment inequalities (*e.g.* Boucheron, Bousquet, Lugosi and Massart [**BBLM05**]), optimal transport (*e.g.* Villani [**Vil03, Vil07a**]), coupling (*e.g.* Chen and Lo [**CL97**]), strong approximation (*e.g.* Berthet and Mason [**BM06b**] and references therein), to name but a few.

11.2.2. Expectations. The main lack of theory probably relies in the comparison between a quantity $F(P, P_n)$ and its resampling estimate in expectation, from the non-asymptotic viewpoint. In this thesis, we use three kinds of techniques for proving such results:

- explicit computations using the specificity of the histogram framework (Prop. 5.2 in Sect. 5.3.2, Lemma 5.7 in Sect. 5.7.2, and Sect. 6.3.3 for explicit computation of the constants for several particular resampling schemes).
- a gaussian assumption, when F has some homogeneity properties (Prop. 10.2 in Sect. 10.2.2)
- symmetrization-like inequalities, adapted from Fromont [**Fro07**], when the data is symmetric (Prop. 10.3 in Sect. 10.2.2). Notice that Fromont obtains less tight bounds without assumptions on the data.

It then seems that there is no general method for comparing $\mathbb{E}\left[F(P, P_n)\right]$ to $\mathbb{E}\left[F(P_n, P_n^W)\right]$ non-asymptotically (whereas Thm. 3.6.13 of van der Vaart and Wellner [**vdVW96**] provides an asymptotic comparison). However, in Sect. 8.4.1, we prove that it is sufficient to consider the case of exchangeable weights, when the data is exchangeable (Lemma 8.4). The expectation problem is thus reduced to the following:

OPEN PROBLEM 11. Prove non-asymptotic bounds on $\mathbb{E}\left[F(P, P_n)\right] / \mathbb{E}\left[F(P_n, P_n^W)\right]$ for general exchangeable weights, and quite general functions F.

In particular, this includes Problem 19 on global resampling complexities, and a similar question on concentration-based thresholds (see Sect. 11.4.3).

11.3. Optimal calibration of penalties

The main point developped in Chap. 2 is the need for theoretical results helping practical users to design optimal model selection procedures. Chapters 3 to 9 provide several answers to this question, with penalization procedures, while raising several new open problems.

11.3.1. Flexibility of penalization. A main argument in favour of penalization is that it is flexible. This is not new: AIC and BIC are the same penalty with a different multiplicative factor. Indeed, identification generally needs to overpenalize for consistency, or even only to control the probability of a type I error (Aerts, Claeskens and Hart [**ACH99**]). This is one of the drawbacks of V-fold cross-validation, which can not overpenalize more than within a factor $3/2$ (by taking $V = 2$, as suggested by Zhang [**Zha93**], Dietterich [**Die98**] and Alpaydin [**Alp99**]). Since we have defined resampling penalties that work in a more general framework than linear penalties for prediction, we can also expect to use penVF and RP for identification:

OPEN PROBLEM 12. V-fold and Resampling penalties, with a multiplicative factor $C = \ln(n) C_{W,\infty}/2$ (or, more generally, $C \gg C_{W,\infty}$), are consistent for

identification. In view of the assumption $m \to \infty$ in [**Sha96**], $C_{W,\infty} \gg n^{-1}$ may also have to be assumed.

This would enlighten the result of Shao [**Sha96**] about "m out of n" bootstrap, since $n \gg m \to \infty$ can also be written $C = 1 \gg C_{W,\infty} \sim m/n \gg n^{-1}$. Moreover, this would show that one can actually use (almost) any kind of resampling scheme for identification, contrary to what Shao's result seems to mean.

It is also known that when the bias of the models decay fast (*e.g.* exponentially fast in the number of parameters), BIC performs better than AIC for prediction. If Problem 12 holds true, it is likely that BIC-like penalties are the best one in such a case. When ideal penalties are linear, it is known that one can combine the strengths of AIC and BIC in several[3] situations (Yang [**Yan03**], van Erven, Grünwald et de Rooij [**vEGdR07**], and references quoted by Yang [**Yan05**]). It is then legitimate to ask whether one can *combine the strengths of AIC-like and BIC-like penalties* for prediction or estimation. From the non-asymptotic viewpoint, this is closely related to Problems 16 and 15.

11.3.2. Calibration of resampling penalties with the slope heuristics. In order to calibrate the constant C in front of RP or V-fold penalties with real data, the use of $C_{W,\infty}$ (which is justified theoretically in the histogram regression case, or when n goes to infinity) may seem hazardous. This is why we also propose to use the so-called slope heuristics. Whereas Birgé and Massart [**BM06c**] only consider penalties that are a function of the dimension (mainly linear, when $\mathcal{M}_n$ has a polynomial complexity), the results of Chap. 3 show that they are valid with general shapes of penalties, in the histogram regression case. This lead us to propose Algorithm 3.1 in Sect. 3.1, in which the shape of the penalty can be estimated from the data.

Moreover, according to Chap. 5 and 6, it appears that RP and penVF provide efficient estimates of the shape of the penalty. We then suggest the use of Algorithm 11.1 below.

ALGORITHM 11.1 (Resampling penalization with slope heuristics).

(1) Choose a resampling scheme, *i.e.* the law of a weight vector W.
(2) Compute the following resampling penalty for each $m \in \mathcal{M}_n$:
$$\mathrm{pen}_{\mathrm{shape}}(m) = \mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) - P_n^W \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) \right] \ .$$
(3) Compute the selected model $\widehat{m}(K)$ as a function of $K > 0$
$$\widehat{m}(K) \in \arg \min_{m \in \mathcal{M}_n} \left\{ P_n \gamma(\widehat{s}_m(P_n)) + K \, \mathrm{pen}_{\mathrm{shape}}(m) \right\} \ .$$
(4) Find $\widehat{K}_{\min} > 0$ such that $D_{\widehat{m}(K)}$ is too large for $K < \widehat{K}_{\min}$ and "reasonably small" for $K > \widehat{K}_{\min}$.
(5) Select the model $\widehat{m} = \widehat{m}\left(2\widehat{K}_{\min}\right)$.

The detailed construction of $\mathrm{pen}_{\mathrm{shape}}$ can be found in Chap. 5 and 6. Remarks about the exact definition of $\widehat{K}_{\min}$ and efficient ways of computing it can be found in Sect. 3.1.

The main drawback of Algorithm 11.1 is that its theoretical justification is restricted to the histogram regression case. First of all, one can wonder whether the slope heuristics itself stays valid in a general framework:

OPEN PROBLEM 13. Generalize the slope heuristics ((2.15), $p_1 \approx p_2$) and its consequences ((2.14), *i.e.* Thm. 3.2: existence of a minimal penalty $\mathrm{pen}_{\min}$,

[3] but not always, since there is a conflict between identification and adaptation, see *e.g.* Yang [**Yan05**].

dimension jump around $\mathrm{pen}_{\min}$; and Thm. 3.1: optimality of $2\,\mathrm{pen}_{\min}$) in a general framework.

Since Problem 13 may be quite hard to solve, we would like to insist on a major part of it, which is probably the most difficult one:

OPEN PROBLEM 14. In a general framework (more general than histogram regression), prove a lower bound on p_1 with high probability, that may be compared to p_2 or $\mathbb{E}[p_2]$.

It may seem strange that Problem 14 is difficult, whereas there exists some concentration results for p_2 (Prop. 7.2, which directly come from a moment inequality of Boucheron and Massart [**BM04**]). The main difference between p_1 and p_2 is that p_2 can be written as

$$p_2(m) = \sup_{t \in S_m} \{ P_n\gamma(s_m) - P_n\gamma(t) \} \quad ,$$

thus Talagrand's inequality (or moment inequalities from Boucheron, Bousquet, Lugosi and Massart [**BBLM05**]) can be directly applied. In the case of p_1, we can not get rid of the randomness of $\widehat{s}_m$ inside $P\gamma(\cdot)$ with the same trick. In the classification framework, there may be a starting point in Thm. 3.1 of Bartlett and Mendelson [**BM06a**]. We do not have found any other lower bound on p_1 in the literature, except the one we prove in the histogram regression case (Prop. 5.8 in Sect. 5.7.4). Notice that in the histogram regression framework, the concentration inequality for p_1 is also much more difficult to obtain that the one for p_2.

11.3.3. Need for overpenalization. We have already pointed out in Sect. 2.4.1 a major problem in practice: *overpenalization is necessary for the non-asymptotic optimality of a penalty*, when the goal is prediction. We observe this in the simulation studies (Sect. 5.4 and 6.5; see also Fig. 11.1), where we considered large signal-to-noise ratios. Moreover, the theoretical bound (2.16) suggests that the need for overpenalization is linked with the variability of $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$. This raises the following questions:

OPEN PROBLEM 15. How much shall we overpenalize? In particular, is the non-asymptotic optimal overpenalization factor related to the variability of $(\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(m)$ (at least for m among the models "likely to be selected")?

Since the variability of $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ may be related to the one of pen (when it is data-dependent, *e.g.* obtained by resampling), Problem 15 is also related to Problem 6.

Even if the above problem was solved, designing non-asymptotic optimal penalties in practice would still be an issue. When pen is resampling-based, we propose an answer, that is the following problem:

OPEN PROBLEM 16. Provide an accurate and practical overpenalization method. We make to proposals:

(1) Would an empirical $(1-\alpha)$ quantile of $P_n^W\left(\gamma(\widehat{s}_m) - \gamma\left(\widehat{s}_m^W\right)\right)$ be efficient in practice?
(2) Alternatively, these empirical quantiles may be used to derive a confidence region at level α on the prediction errors, and then a confidence set for the oracle $m^\star$. Would the "more parcimonious[4] model" in this set be a good choice for $\widehat{m}$?

In both cases, how to choose the "confidence level" α?

[4] *e.g.* defined as the one with the smallest penalty.

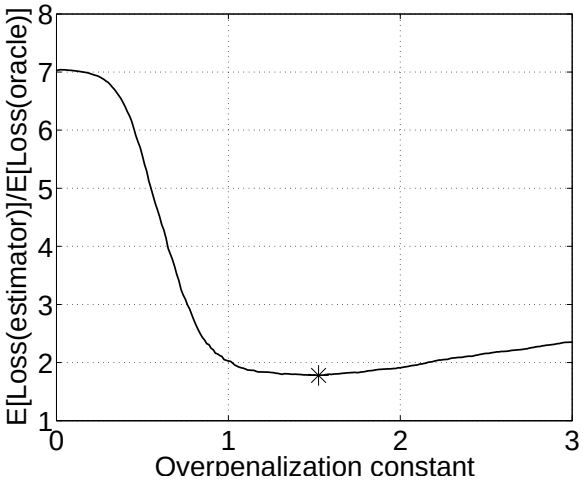


FIGURE 11.1. *How to choose the optimal overpenalization factor?* Performance of Mallows' C_p (measured by C_{or} defined by (4.3)) as a function of the overpenalization factor C_{ov}. Taking into account the uncertainty of estimation of C_{or} (standard deviations are smaller than 0.04), the optimal overpenalization factor belongs to $[1.25; 1.75]$ in this case. The data and the family of models are the one of experiment (S1) (Sect. 4.4.2). A similar behaviour is observed with experiments (S2), (HSd1) and (HSd2), and Resampling Penalties instead of Mallows' C_p.

Remark that for method (1), overpenalization may not necessarily match with the smaller values of α. The crucial point here is: "how does the variability of $\text{pen} - \text{pen}_{\text{id}}$ varies with D_m?". On the contrary, in method (2), since the confidence set for $m^\star$ is a non-increasing function of α, overpenalization corresponds to the smaller values of α.

The presence of a "confidence level" α in the two above proposals makes our suggest close to a testing procedure for making model selection. This may be related to the link between FDR control and model selection, by Abramovich, Benjamini, Donoho and Johnstone [**ABDJ06**]. See also Aerts, Claeskens and Hart [**ACH99**] which quantifies the "level" of AIC in an identification setting, and suggests to overpenalize in order to control this level. Another interesting reference for Problem 16 may be Birgé [**Bir06**], where a theoretical model selection procedure by testing is defined (as an alternative to penalization).

As an alternative to Problem 16, one could think of using (V-fold) cross-validation again for choosing the overpenalization factor C_{ov}. Apart from its prohibitive computational cost, this method would probably work well, at least at first order.

11.3.4. Larger families of models. All along this thesis, we only consider model selection among a polynomial family $\mathcal{M}_n$ (assumption (**P1**)). With our definition of RP and V-fold penalties, this is probably necessary, since too large families of models need larger penalties than polynomial families (Birgé and Massart [**BM01, BM06c**], Baraud [**Bar02**], Sauvé [**Sau06**]).

When there are much more models, for instance in the segmentation problem (*i.e.* classification with the entire family of histogram models) or in the change-points detection problem (*i.e.* regression with the entire family of histogram models, see Lebarbier [**Leb05**]), we have to make another proposal for using resampling penalties. Massart [**Mas07**] suggests to replace the rich family $(S_m)_{m\in\mathcal{M}_n}$ by the polynomial family $\left(\widetilde{S}_D\right)_{1\leq D\leq n}$, where $\widetilde{S}_D = \bigcup_{D_m=D} S_m$ and D_m is any complexity measure of the model m (for instance, its dimension as a vector space). In other words, we come back to the polynomial case with a different family of models. Notice that even

if each S_m is a model of histograms, neither are the $\widetilde{S}_D$, so that our results can not be applied to change points detection. In the classification case, proving a result on the use of resampling for segmentation would be quite interesting, since it is closely related to the use of resampling for stabilizing CART. In other words, we would like to prove:

> OPEN PROBLEM 17. Extend Thm. 6.1 to more general models, in particular for the change-point detection and segmentation problems.

When the noise is heteroscedastic, another problem arises, because the natural complexity measure (if there is one) depends on the unknown function $\sigma(\cdot)$. For instance, when $\sigma(x) = \mathbb{1}_{x \geq 1/2}$, the complexity of an histogram model m is measured by the number of jumps in $[1/2, 1]$, which may be much smaller than its dimension D_m. Since grouping the models according to D_m implicitly assumes that they should be penalized in the same way, the resulting procedure may then be suboptimal. We do not have yet suggestions for the following problem:

> OPEN PROBLEM 18. When the noise is heteroscedastic, how to group histogram regression models? More generally, is there a "natural" way of grouping the models, possibly a data-dependent one?

An alternative to grouping may come from an answer to Problem 15, since the need for larger penalties when $\mathrm{Card}(\mathcal{M}_n)$ is large is linked with the uniform fluctuations of $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$.

11.3.5. Global penalties. In Chap. 9, we highlighted the difficulty of calibration of global resampling penalties. There may be two answers for this, which are the two following problems:

> OPEN PROBLEM 19. With an additional assumption (for instance, that the best estimator in S_m has a risk larger than n^{-1}), prove tight theoretical bounds on the ratio $R_Z(\mathcal{F}_m)$, similar to the ones of Sect. 9.2.2.

> OPEN PROBLEM 20. Can we calibrate global penalties with a slope heuristics algorithm, since these penalties estimate $\mathrm{pen}_{\mathrm{id,g}}$ which may not have the shape of $\mathrm{pen}_{\mathrm{id}}$?

11.4. Confidence regions and multiple testing

In Chap. 10, we define several resampling-based confidence regions, with a non-asymptotic control of the level. A simulation study shows their performance when the coordinates are correlated: our thresholds seem to "adapt" to the unknown correlation matrix of the data. We can thus conjecture that the answer to Problem 4 is positive.

11.4.1. Quantiles without additional term. Moreover, the quantile thresholds defined in Sect. 10.3 seem to involve a too large additive term, maybe unnecessary in the gaussian framework. A control of the level of these thresholds without the additive term would result in a procedure uniformly more powerful than the "uncentered quantile":

> OPEN PROBLEM 21. The remainder term in Thm. 10.2 can be removed (or made much smaller), while keeping the level smaller than α (possibly with an additional assumption on the distribution of the data).
> Otherwise, is it possible to build a self-contained quantile threshold?

The interest of the last statement is that we would have a threshold valid for any symmetric variable, not only the gaussian or bounded symmetric ones.

11.4.2. Unknown noise-level. When applying our procedures to real data sets, the estimation of the noise-level σ can be an issue, as pointed out in Sect. 10.6.3. In neuroimaging, some independent data are often available, thanks to which we can estimate (at least roughly) the noise-level σ, and even local estimates of the variance, allowing some global heteroscedasticity. However, in general, σ has to be estimated with the same data. This raises a two-fold problem:

> OPEN PROBLEM 22. In the homoscedastic case, should the resampling-based thresholds take into account the estimation of σ, *i.e.* estimate a quantile of $\widehat{\sigma}^{-1}\phi\left(\overline{\mathbf{Y}}-\mu\right)$ instead of a quantile of $\sigma^{-1}\phi\left(\overline{\mathbf{Y}}-\mu\right)$? In the heteroscedastic case, what can be done?

This problem may in particular be related to the (single) test built by Baraud, Huet and Laurent [**BHL03**] upon concentration inequalities, with an unknown noise-level.

11.4.3. Non-gaussian or asymmetric data. In practice, gaussian or symmetry assumptions are often questionable. In our approach, we really use symmetry, for controlling expectations (in the concentration approach) and through the symmetrization trick (in the quantile approach). Results from Chap. 9 show that without symmetry, it may be difficult to obtain a general tight calibration of the concentration-based thresholds. However, the resampling heuristics is not linked with the symmetry of the noise, so that it seems somehow unnatural to restrict our results to this case.

For the concentration-based thresholds, solving Problem 10 would allow to consider sub-gaussian symmetric variables, instead of gaussian or bounded ones (recall that the thresholds in the bounded case are quite conservative, and make use of the bound A on the data). Sub-gaussian asymmetric variables would then require tight results in expectation (*e.g.* through Problem 19, which is part of Problem 11).

For the quantile approach, proving results without symmetrization trick seems quite hard: up to our best knowledge, all the non-asymptotic results on quantiles rely on symmetrization. Since asymptotical results show that general exchangeable weights can be used (Hall and Mammen [**HM94**]), as well as asymmetric data (as soon as it is not too far from gaussian), we can rightfully ask the following:

> OPEN PROBLEM 23. Control the level of the quantile thresholds with either asymmetric data or general exchangeable weights.

Bibliographie

[ABDJ06] Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2) :584–653, 2006.

[ABR07] Sylvain Arlot, Gilles Blanchard, and Étienne Roquain. Resampling-based confidence regions and multiple tests for a correlated random vector. In *COLT 2007*, volume 4539 of *Lecture Notes in Artificial Intelligence*, pages 127–141. Springer, Berlin, 2007.

[ACH99] Marc Aerts, Gerda Claeskens, and Jeffrey D. Hart. Testing the fit of a parametric function. *J. Amer. Statist. Assoc.*, 94(447) :869–879, 1999.

[AG92] Miguel A. Arcones and Evarist Giné. On the bootstrap of M-estimators and other statistical functionals. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 13–47. Wiley, New York, 1992.

[Aka70] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22 :203–217, 1970.

[Aka73] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[Ald85] David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[All74] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16 :125–127, 1974.

[Alp99] Ethem Alpaydin. Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. *Neur. Comp.*, 11(8) :1885–1892, 1999.

[BA02] Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference*. Springer-Verlag, New York, second edition, 2002. A practical information-theoretic approach.

[Bai98] Sylvain Baillet. *Vers une imagerie fonctionnelle de l'électrophysiologie corticale. Modélisation markovienne pour l'estimation des sources de la magnéto/électroencéphalographie et évaluations expérimentales.* PhD thesis, University Paris XI, July 1998. http ://cogimage.dsi.cnrs.fr/perso/sbaillet/PhD.html.

[Bar00] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.

[Bar02] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 :127–146 (electronic), 2002.

[Bar04] Yannick Baraud. Confidence balls in Gaussian regression. *Ann. Statist.*, 32(2) :528–551, 2004.

[Bau07] Jean-Patrick Baudry. Clustering through model selection criteria. , 2007. Poster session at One Day Statistical Workshop in Lisieux. `http ://www.math.u-psud.fr/~baudry`, June 2007.

[BB95] Philippe Barbe and Patrice Bertail. *The weighted bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995.

[BBL02] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48 :85–113, 2002.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : a survey of some recent advances. *ESAIM Probab. Stat.*, 9 :323–375 (electronic), 2005.

[BBLM05] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2) :514–560, 2005.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4) :1497–1537, 2005.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2(3) :499–526, 2002.

[Ber03] Rudolf Beran. The impact of the bootstrap on statistical algorithms and theory. *Statist. Sci.*, 18(2) :175–184, 2003. Silver anniversary of the bootstrap.

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[BG04] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.*, 5 :1089–1105 (electronic), 2004.

[BGH07] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with unknown variance. Preprint. Arxiv :math.ST/0701250, January 2007.

[BH75] H. G. Burchard and D. F. Hale. Piecewise polynomial approximation on optimal meshes. *J. Approximation Theory*, 14(2) :128–147, 1975.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) :289–300, 1995.

[BHL03] Yannick Baraud, Sylvie Huet, and Béatrice Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1) :225–251, 2003.

[BHL05] Yannick Baraud, Sylvie Huet, and Béatrice Laurent. Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.*, 33(1) :214–257, 2005.

[Bir06] Lucien Birgé. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3) :273–325, 2006.

[BKL99] Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out : bounds for K-fold and progressive cross-validation. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, pages 203–208 (electronic), New York, 1999. ACM.

[BM97] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities : risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory) :463–482, 2002.

[BM04] Stéphane Boucheron and Pascal Massart. Data-driven penalties : heuristics and results. Personal communication, February 2004.

[BM06a] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3) :311–334, 2006.

[BM06b] Philippe Berthet and David M. Mason. Revisiting two strong approximation results of dudley and philipp. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monograph Series*, pages 155–172. Inst. Math. Statist., 2006.

[BM06c] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields*, 134(3), 2006.

[BM06d] Gilles Blanchard and Pascal Massart. Discussion : "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6) :2664–2671, 2006.

[BMP04] Peter L. Bartlett, Shahar Mendelson, and Petra Philips. Local complexities for empirical risk minimization. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 270–284. Springer, Berlin, 2004.

[Boo03] Dennis D. Boos. Introduction to the bootstrap world. *Statist. Sci.*, 18(2) :168–174, 2003. Silver anniversary of the bootstrap.

[BR01] Peter J. Bickel and Jian-Jian Ren. The bootstrap in hypothesis testing. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 91–112. Inst. Math. Statist., Beachwood, OH, 2001.

[Bre83] Jean Bretagnolle. Lois limites du bootstrap de certaines fonctionnelles. *Ann. Inst. H. Poincaré Sect. B (N.S.)*, 19(3) :281–296, 1983.

[Bre92] Leo Breiman. The little bootstrap and other methods for dimensionality selection in regression : X-fixed prediction error. *J. Amer. Statist. Assoc.*, 87(419) :738–754, 1992.

[Bre96] Leo Breiman. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24(6) :2350–2383, 1996.

[BS92] Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60(3) :291–319, 1992.

[BS05] Peter Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles. *Statistica Sinica*, 2005. To appear.

[Bur89] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3) :503–514, 1989.

[Bur90] Prabir Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3) :314–345, 1990.

[Bur02] Prabir Burman. Estimation of equifrequency histograms. *Statist. Probab. Lett.*, 56(3) :227–238, 2002.

[BY02] Peter Bühlmann and Bin Yu. Analyzing bagging. *Ann. Statist.*, 30(4) :927–961, 2002.

[Cas03] George Casella, editor. *Silver anniversary of the bootstrap*. Institute of Mathematical Statistics, Bethesda, MD, 2003. Statist. Sci. **18** (2003), no. 2.

[Cha04] Sourav Chatterjee. *Concentration inequalities with exchangeable pairs*. PhD thesis, Stanford University, September 2004.

[CIS76] B. S. Cirel′son, I. A. Ibragimov, and V. N. Sudakov. Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975)*, pages 20–41. Lecture Notes in Math., Vol. 550, Berlin, 1976. Springer.

[CL97] Kani Chen and Shaw-Hwa Lo. On a mapping approach to investigating the bootstrap accuracy. *Probab. Theory Related Fields*, 107(2) :197–217, 1997.

[CR06] Alain Celisse and Stéphane Robin. Non-parametric density estimation by exact leave-p-out cross-validation. *C.S.D.A.*, 2006. To appear.

[CS97] Joseph E. Cavanaugh and Robert H. Shumway. A bootstrap variant of AIC for state-space model selection. *Statist. Sinica*, 7(2) :473–496, 1997.

[CW79] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4) :377–403, 1978/79.

[DBD07] Antoine DelCul, Sylvain Baillet, and Stanislas Dehaene. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, 5(10) :online, September 2007.

[DE96] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3) :189–228, 1996. With comments and a rejoinder by the authors.

[Die98] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neur. Comp.*, 10(7) :1895–1924, 1998.

[DJ95] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432) :1200–1224, 1995.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[DR98] Devdatt Dubhashi and Desh Ranjan. Balls and bins : a study in negative dependence. *Random Structures Algorithms*, 13(2) :99–124, 1998.

[DRP+05] Felix Darvas, M. Rautiainen, D. Pantazis, Sylvain Baillet, H. Benali, J.C. Mosher, L. Garnero, and R.M. Leahy. Investigations of dipole localization accuracy in MEG using the bootstrap. *NeuroImage*, 25(2) :355–368, April 2005.

[DSB03] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, 18(1) :71–103, 2003.

[DW77] Luc P. Devroye and Terry J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.*, 5(3) :536–540, 1977.

[DW78] Luc P. Devroye and Terry J. Wagner. Addendum to : "The strong uniform consistency of nearest neighbor density estimates" (Ann. Statist. **5** (1977), no. 3, 536–540). *Ann. Statist.*, 6(4) :935, 1978.

[Efr79] Bradley Efron. Bootstrap methods : another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 1979.

[Efr82] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[Efr83] Bradley Efron. Estimating the error rate of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382) :316–331, 1983.

[Efr86] Bradley Efron. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394) :461–470, 1986.

[Efr03] Bradley Efron. Second thoughts on the bootstrap. *Statist. Sci.*, 18(2) :135–140, 2003. Silver anniversary of the bootstrap.

[Efr04] Bradley Efron. The estimation of prediction error : covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467) :619–642, 2004. With comments and a rejoinder by the author.

[EP96] Sam Efromovich and Mark Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4) :925–942, 1996.

[ET93] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[ET97] Bradley Efron and Robert Tibshirani. Improvements on cross-validation : the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438) :548–560, 1997.

[FG94] Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4) :1947–1975, 1994.

[Fis35] Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[FR89] M. Falk and R.-D. Reiss. Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *Ann. Probab.*, 17(1) :362–371, 1989.

[Fro03] Magalie Fromont. *Quelques problèmes de sélection de modèles : construction de tests adaptatifs, ajustement de pénalités par des méthodes de bootstrap*. Thèse de doctorat, Université Paris-Sud XI, December 2003.

[Fro04] Magalie Fromont. Model selection by bootstrap penalization for classification. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 285–299. Springer, Berlin, 2004.

[Fro07] Magalie Fromont. Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3) :165–207, 2007.

[GDS03] Yongchao Ge, Sandrine Dudoit, and Terence P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1) :1–77, 2003. With comments and a rejoinder by the authors.

[Gei75] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70 :320–328, 1975.

[Gin97] Evarist Giné. Lectures on some aspects of the bootstrap. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, volume 1665 of *Lecture Notes in Math.*, pages 37–151. Springer, Berlin, 1997.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[GLZ00] Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

[GP05] Leonid Galtchouk and Sergey Pergamenshchikov. Efficient adaptive nonparametric estimation in heteroscedastic models. Université Louis Pasteur, IRMA, Preprint, 2005.

[GZ84] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4) :929–998, 1984. With discussion.

[GZ90] Evarist Giné and Joel Zinn. Bootstrapping general empirical measures. *Ann. Probab.*, 18(2) :851–869, 1990.

[Hal90] Peter Hall. Asymptotic properties of the bootstrap for heavy-tailed distributions. *Ann. Probab.*, 18(3) :1342–1360, 1990.

[Hal92] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[Hal03] Peter Hall. A short prehistory of the bootstrap. *Statist. Sci.*, 18(2) :158–167, 2003. Silver anniversary of the bootstrap.

[Har69] John A. Hartigan. Using subsample values as typical values. *J. Amer. Statist. Assoc.*, 64 :1303–1317, 1969.

[Har75] John A. Hartigan. Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.*, 3 :573–580, 1975.

[HDR89] Peter Hall, Thomas J. DiCiccio, and Joseph P. Romano. On smoothing and the bootstrap. *Ann. Statist.*, 17(2) :692–704, 1989.

[HJ93] Marie Hušková and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *Ann. Statist.*, 21(4) :1811–1823, 1993.

[HM88] Peter Hall and Michael A. Martin. Exact convergence rate of bootstrap quantile variance estimator. *Probab. Theory Related Fields*, 80(2) :261–268, 1988.

[HM89] Peter Hall and Michael A. Martin. A note on the accuracy of bootstrap percentile method confidence intervals for a quantile. *Statist. Probab. Lett.*, 8(3) :197–200, 1989.

[HM93] W. Härdle and E. Mammen. Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4) :1926–1947, 1993.

[HM94] Peter Hall and Enno Mammen. On general resampling algorithms and their performance in distribution estimation. *Ann. Statist.*, 22(4) :2011–2030, 1994.

[Hol79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 :65–70, 1979.

[HS05] Don Hush and Clint Scovel. Concentration of the hypergeometric distribution. *Statist. Probab. Lett.*, 75(2) :127–132, 2005.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[IS91] Makio Ishiguro and Yosiyuki Sakamoto. WIC : An estimation free criterion. *Research Memorandum of the Institute of Statistical Mathematics*, page 410, 1991.

[ISK97] Makio Ishiguro, Yosiyuki Sakamoto, and Genshiro Kitagawa. Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.*, 49(3) :411–434, 1997.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables, with applications. *Ann. Statist.*, 11(1) :286–295, 1983.

[JLN+07] Karim Jerbi, Jean-Philippe Lachaux, Karim N'Diaye, Dimitrios Pantazis, Richard M. Leahy, Line Garnero, and Sylvain Baillet. Coherent neural representation of hand speed in humans revealed by MEG imaging. *PNAS*, 104(18) :7676–7681, May 2007.

[JN00] Anatoli Juditsky and Arkadii Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3) :681–712, 2000.

[JP03] Arnold Janssen and Thorsten Pauls. How do bootstrap and permutation tests work? *Ann. Statist.*, 31(3) :768–806, 2003.

[JZ04] C. Matthew Jones and Anatoly A. Zhigljavsky. Approximating the negative moments of the Poisson distribution. *Statist. Probab. Lett.*, 66(2) :171–181, 2004.

[Kol01] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5) :1902–1914, 2001.

[Kol06] Vladimir Koltchinskii. 2004 IMS Medallion Lecture : Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6), 2006.

[Leb05] Émilie Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85 :717–736, 2005.

[Lec07a] Guillaume Lecué. *Méthodes d'agrégation : optimalité et vitesses rapides.* PhD thesis, LPMA, University Paris VII, May 2007.

[Lec07b] Guillaume Lecué. Suboptimality of penalized empirical risk minimization in classification. In *COLT 2007*, volume 4539 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, 2007.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[Lep02] Vincent Lepez. *Some estimation problems related to oil reserves.* PhD thesis, University Paris XI, 2002.

[Lew76] Robert A. Lew. Bounds on negative moments. *SIAM J. Appl. Math.*, 30(4) :728–731, 1976.

[Li87] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation : discrete index set. *Ann. Statist.*, 15(3) :958–975, 1987.

[Liu88] Regina Y. Liu. Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.*, 16(4) :1696–1708, 1988.

[Loz00] Fernando Lozano. Model selection using rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany.* ICSC Academic Press, 2000.

[LP05] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference : facts and fiction. *Econometric Theory*, 21(1) :21–59, 2005.

[LR05] Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses.* Springer Texts in Statistics. Springer, New York, third edition, 2005.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[Lug02] Gábor Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 1–56. Springer, Vienna, 2002.

[LW04] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4) :1679–1697, 2004.

[Mal73] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.

[Mam92] Enno Mammen. *When does bootstrap work ? Asymptotic results and simulations*, volume 77 of *Lecture Notes in Statistics*. Springer, 1992.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[McD89] Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[MM07] Cathy Maugis and Bertrand Michel. A nonasymptotic penalized criterion for gaussian mixture model selection. a variable selection and clustering problems. In preparation, September 2007.

[MN92] David M. Mason and Michael A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3) :1611–1624, 1992.

[MN06] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5) :2326–2366, 2006.

[MSP05] Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. Prediction error estimation : a comparison of resampling methods. *Bioinformatics*, 21(15) :3301–3307, 2005.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6) :1808–1829, 1999.

[Nem00] Arkadi Nemirovski. *Topics in non-parametric statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000.

[OW75] Anders Odén and Hans Wedel. Arguments for Fisher's permutation test. *Ann. Statist.*, 3 :518–520, 1975.

[PNBL05] Dimitrios Pantazis, Thomas E. Nichols, Sylvain Baillet, and Richard M. Leahy. A comparison of random field theory and permutation methods for statistical analysis of MEG data. *NeuroImage*, 25 :383–394, 2005.

[Pol03] Dimitris N. Politis. The impact of bootstrap methods on time series analysis. *Statist. Sci.*, 18(2) :219–230, 2003. Silver anniversary of the bootstrap.

[PPGVW04] Marco Perone Pacifico, Christopher R. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *J. Amer. Statist. Assoc.*, 99(468) :1002–1014, 2004.

[PRW99] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.

[PvdL03] Katherine S. Pollard and Mark J. van der Laan. Resampling-based multiple testing : Asymptotic control of type I error and applications to gene expression data. Working Paper Series Working Paper 121, U.C. Berkeley Division of Biostatistics, 2003. available at http ://www.bepress.com/ucbbiostat/paper121.

[PW93] Jens Præstgaard and Jon A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4) :2053–2086, 1993.

[Que49] Maurice H. Quenouille. Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B.*, 11 :68–84, 1949.

[Ris78] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, 1978.

[Rom89] Joseph P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, 17(1) :141–159, 1989.

[Rom90] Joseph P. Romano. On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.*, 85(411) :686–692, 1990.

[Roq07] Étienne Roquain. *Motifs exceptionnels dans des séquences hétérogènes. Contributions à la théorie et à la méthodologie des tests multiples.* PhD thesis, University Paris XI, October 2007.

[Rub81] Donald B. Rubin. The Bayesian bootstrap. *Ann. Statist.*, 9(1) :130–134, 1981.

[RW05] Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469) :94–108, 2005.

[RW07] Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4) :1378–1408, 2007.

[Sau06] Marie Sauvé. Histogram selection in non gaussian regression. Technical Report 5911, INRIA, may 2006.

[SBD05] Claire Sergent, Sylvain Baillet, and Stanislas Dehaene. Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, 8(10) :1391–1400, October 2005.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[Ser74] Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *Ann. Statist.*, 2 :39–48, 1974.

[Sha93] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422) :486–494, 1993.

[Sha96] Jun Shao. Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434) :655–665, 1996.

[Sha97] Jun Shao. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2) :221–264, 1997. With comments and a rejoinder by the author.

[Shi81] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1) :45–54, 1981.

[Shi86] Ritei Shibata. Selection of the number of regression variables ; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.*, 38(3) :459–474, 1986.

[Shi97] Ritei Shibata. Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica*, 7(2) :375–394, 1997.

[Sin81] Kesar Singh. On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, 9(6) :1187–1195, 1981.

[SS01] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[ST95] Jun Shao and Dong Sheng Tu. *The jackknife and bootstrap*. Springer Series in Statistics. Springer-Verlag, New York, 1995.

[ST06] Marie Sauvé and Christine Tuleau. Variable selection through cart. Technical report, INRIA, May 2006.

[Ste81] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6) :1135–1151, 1981.

[Sto74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36 :111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[Sto80] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6) :1348–1360, 1980.

[Sto85] Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

[SY87] B. W. Silverman and G. A. Young. The bootstrap : to smooth or not to smooth? *Biometrika*, 74(3) :469–479, 1987.

[Tal96] Michel Talagrand. A new look at independence. *Ann. Probab.*, 24(1) :1–34, 1996.

[THNC03] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, 18(1) :104–117, 2003.

[Tsy04] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1) :135–166, 2004.

[Tuk58] John Tukey. Bias and confidence in not-quite large samples. *Ann. Math. Statist.*, 29 :614, 1958. (abstract).

[Vap82] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[Vap98] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[VC74] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow, 1974. Theory of Pattern Recognition (In Russian).

[vdG06] Sara van de Geer. Discussion : "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6) :2688–2696, 2006.

[vdLDK04] Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.*, 3 :Art. 4, 27 pp. (electronic), 2004.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[vEGdR07] T. van Erven, P. D. Grünwald, and S. de Rooij. Catching up faster in bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, Cambridge, MA, 2007. MIT Press.

[Ver07] Nicolas Verzelen. Model selection for graphical models. In preparation, September 2007.

[Vil03] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[Vil07a] Cédric Villani. *Optimal transport, old and new*. Lecture Notes in Mathematics. Springer, 2007. Lectures from the 35th Summer School on Probability Theory held in Saint-Flour, July 2005.

[Vil07b] Fanny Villers. *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, University Paris XI, December 2007.

[Wen89] Chung-Sing Weng. On a second-order asymptotic property of the Bayesian bootstrap mean. *Ann. Statist.*, 17(2) :705–710, 1989.

[WGK+03] Till Waberski, R. Gobbele, W. Kawohl, C. Cordes, and H. Buchner. Immediate cortical reorganization after local anesthetic block of the thumb : source localization of somatosensory evoked potentials in human subjects. *Neurosci. Lett.*, 347 :151–154, 2003.

[Wu86] Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4) :1261–1350, 1986. With discussion and a rejoinder by the author.

[WY93] Peter H. Westfall and S. Stanley Young. *Resampling-based multiple testing : examples and methods for p-value adjustment*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. Wiley, New York, 1993. Examples and Methods for P- Value Adjustment.

[Yan03] Yuhong Yang. Regression with multiple candidate models : selecting or mixing ? *Statist. Sinica*, 13(3) :783–809, 2003.

[Yan04] Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1) :25–47, 2004.

[Yan05] Yuhong Yang. Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950, 2005.

[Yan06] Yuhong Yang. Comparing learning methods for classification. *Statist. Sinica*, 16(2) :635–657, 2006.

[Yan07] Yuhong Yang. Consistency of cross validation for comparing regression procedures. Accepted by Annals of Statistics, 2007.

[YB99] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference*, 82(1-2) :171–196, 1999. Multiple comparisons (Tel Aviv, 1996).

[You94] George Alastair Young. Bootstrap : more than a stab in the dark ? *Statist. Sci.*, 9(3) :382–415, 1994. With discussion and a rejoinder by the author.

[Zha93] Ping Zhang. Model selection via multifold cross validation. *Ann. Statist.*, 21(1) :299–313, 1993.

[Žni05] Marko Žnidarič. Asymptotic expansions for inverse moments of binomial and poisson distributions. arXiv :math.ST/0511226, November 2005.

Rééchantillonnage et Sélection de modèles

Résumé : Cette thèse s'inscrit dans les domaines de la statistique non-paramétrique et de la théorie statistique de l'apprentissage. Son objet est la compréhension fine de certaines méthodes de rééchantillonnage ou de sélection de modèles, du point de vue non-asymptotique.

La majeure partie de ce travail de thèse consiste dans la calibration précise de méthodes de sélection de modèles optimales en pratique, pour le problème de la prédiction. Nous étudions la validation croisée V-fold (très couramment utilisée, mais mal comprise en théorie, notamment pour ce qui est de choisir V) et plusieurs méthodes de pénalisation. Nous proposons des méthodes de calibration précise de pénalités, aussi bien pour ce qui est de leur forme générale que des constantes multiplicatives. L'utilisation du rééchantillonnage permet de résoudre des problèmes difficiles, notamment celui de la régression avec un niveau de bruit variable. Nous validons théoriquement ces méthodes du point de vue non-asymptotique, en prouvant des inégalités oracle et des propriétés d'adaptation. Ces résultats reposent entre autres sur des inégalités de concentration.

Un second problème que nous abordons est celui des régions de confiance et des tests multiples, lorsque l'on dispose d'observations de grande dimension, présentant des corrélations générales et inconnues. L'utilisation de méthodes de rééchantillonnage permet de s'affranchir du fléau de la dimension, et d'«apprendre» ces corrélations. Nous proposons principalement deux méthodes, et prouvons pour chacune un contrôle non-asymptotique de leur niveau.

Mots-clés : Statistique non-paramétrique, apprentissage statistique, rééchantillonnage, non-asymptotique, validation croisée V-fold, bootstrap, sélection de modèles, pénalisation, régression non-paramétrique, adaptation, hétéroscédastique, régions de confiance, tests multiples.

Resampling and Model Selection

Abstract:

This thesis takes place within the theories of non-parametric statistics and statistical learning. Its goal is to provide an accurate understanding of several resampling or model selection methods, from the non-asymptotic viewpoint.

The main advance in this thesis consists in the accurate calibration of model selection procedures, in order to make them optimal in practice for prediction. We study V-fold cross-validation (very commonly used, but badly known in theory, in particular for the question of choosing V) and several penalization procedures. We propose methods for calibrating accurately some penalties, for both their general shape and the multiplicative constants. The use of resampling allows to solve hard problems, in particular regression with a variable noise-level. We prove non-asymptotic theoretical results on these methods, such as oracle inequalities and adaptivity properties. These results rely in particular on some concentration inequalities.

We also consider the problem of confidence regions and multiple testing, when the data are high-dimensional, with general and unknown correlations. Using resampling methods, we can get rid of the curse of dimensionality, and «learn» these correlations. We mainly propose two procedures, and prove for both a non-asymptotic control of their level.

Keywords: Non-parametric statistics, statistical learning, resampling, non-asymptotic, V-fold cross-validation, bootstrap, model selection, penalization, nonparametric regression, adaptivity, heteroscedastic, confidence regions, multiple testing.

AMS Classification: 62G09, 62M20, 62G08, 62J02, 62G15, 62G10.

N° d'impression: 2831
Quatrième trimestre 2007