Apprentissage statistique multi-tâches

Matthieu Solnon

THÈSE

Présentée à

L'UNIVERSITÉ PARIS VI - PIERRE ET MARIE CURIE

ÉCOLE DOCTORALE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Par Matthieu SOLNON

POUR OBTENIR LE GRADE DE
DOCTEUR

Spécialité : Mathématiques

Apprentissage statistique multi-tâches

Directeurs de thèse : Sylvain ARLOT et Francis BACH
Rapporteurs : Vincent RIVOIRARD et Larry WASSERMAN

Soutenue le 25 novembre 2013 devant la commission d'examen formée de :

Sylvain ARLOT	CNRS - ENS	Directeur
Francis BACH	INRIA - ENS	Directeur
Gérard BIAU	UPMC - ENS	Examinateur
Arnak DALALYAN	ENSAE	Examinateur
Vincent RIVOIRARD	Université Paris Dauphine	Rapporteur
Jean-Philippe VERT	Mines ParisTech - Institut Curie	Examinateur

Département d'informatique
École normale supérieure
45 rue d'Ulm
F-75230 PARIS CEDEX 05

Remerciements

Je voudrais d'abord remercier Sylvain et Francis, qui m'ont accompagné et encadré ces années durant. Ce fut, pour moi, une chance de vous rencontrer, de pouvoir discuter avec vous et de profiter de vos points de vue, souvent différents mais toujours complémentaires ! Votre soutien ne fut pas que scientifique, et je ne saurais vous remercier suffisamment pour la chaleur de votre accueil.

Je tiens ensuite à remercier Vincent Rivoirard et Larry Wasserman d'avoir rédigé les rapports de cette thèse. La minutie, la précision et l'attention de votre lecture m'ont honoré et ont contribué à améliorer ce manuscrit. Un grand merci, aussi, à Gérard Biau, Arnak Dalalyan, Vincent Rivoirard et Jean-Philippe Vert pour leur participation à mon jury de soutenance.

J'ai eu le plaisir de travailler dans un cadre privilégié, d'abord au sein de l'équipe Willow, puis dans l'équipe Sierra. L'ambiance y a toujours été agréable ! Arriver dans un tel laboratoire, plus concerné par l'informatique que les mathématiques, a été dépaysant, mais aussi enrichissant, tout autant que travailler avec des *vision people* (doit-on les appeler des visionnaires ?). Les différentes retraites et conférences (Normandie, NIPS 2011 à Grenade, Bandol) ont été des moments très forts et qui m'ont beaucoup apporté. Tous mes remerciements vont donc aux membres, passés et présents, de l'équipe. Vous êtes trop nombreux pour être tous cités, mais j'adresse en particulier un grand remerciement à Édouard, mon inlassable compagnon du bureau 41. L'équipe administrative nous a bien facilité la tâche, merci, donc, à Joëlle, Cécile, Marine et Lindsay ! J'aimerais aussi remercier, pour leurs discussions mathématiques et leur disponibilité, Guillaume Obozinski (amateur éclairé de James-Stein !), Jean-Yves Audibert et Olivier Catoni (mes tentatives dans la direction PAC-Bayésienne n'ont malheureusement pas abouti, mais elles ont été très intéressantes). Ces discusssions n'ont pas toutes eu de débouché, mais toutes furent passionnantes ! Les colloques de Fréjus, auxquels j'ai eu le plaisir d'assister en 2010, 2011 et 2013, furent aussi un grand lieu de rencontres, et je ne peux qu'en remercier leurs organisateurs et leurs participants. Enfin, même si cela ne transparaît pas dans cette thèse, mon activité de monitorat à l'Université Paris 6 a été extrêmement enrichissante et m'a révélé le plaisir d'enseigner. Je ne peux que remercier mes collègues, notamment Claire David, Daniel Hoehener, Alexandre Guilbaud, Tabea Rebafka, Bertrand Michel et Patricia Conde-Céspedes, pour leur accueil.

Je souhaite vivement remercier toutes les personnes ayant contribué à mon parcours mathématique. J'ai eu la chance d'être aiguillé dans cette direction par Volny De Pascale (notamment lors de mémorables séjours à Manosque !), puis de rencontrer de talentueux professeurs, en particulier Pascal Galmiche et Jean-Jacques Técourt. À eux va toute ma

reconnaissance. Ma première rencontre avec les phénomènes aléatoires s'est déroulée lors de ma première année à l'École normale supérieure, sous la houlette de la *dream team* probabiliste, constituée de Jean Bertoin, Mathilde Weill, Wendelin Werner et Marie Théret. J'ai rarement tant appris qu'alors. Tombé sous le charme, j'ai ensuite eu la chance, en deuxième année, de suivre les enseignements de Patricia Reynaud-Bouret et de Vincent Rivoirard, qui m'ont convaincu de poursuivre mes études dans le domaine de la statistique mathématique. Enfin, je ne peux que remercier mes enseignants du M2 d'Orsay, et notamment Vincent Rivoirard, Cécile Durot, Pascal Massart, Gilles Stoltz, Sylvain Arlot et Francis Bach, de m'avoir amené à faire une thèse en statistique.

Je voudrais aussi remercier tous mes amis, qui m'ont soutenu durant ces années. Maud, Adrien, Nicolas, Clothilde et Ruben me supportent depuis de nombreuses années, et leur fidèle amitié m'a tant apporté que je ne saurais leur borner mes remerciements. Tous mes remerciements vont aussi aux *scrouickies*, Arthur, Pu, Stéphane, Nicolas et Marie, Manon et Rémy, Furcy, Igor, Oriane et Steve. Que serais-je sans vous ? Un grand merci à mes compagnons d'œnologie qui, entre expéditions viticoles, concours internationaux, salons et soirées de dégustations, se révèlent être de bons amis : Pierre et Cécile, Adrien, Guillaume et Anne-Sophie, Florian et Marie, Rémy. Enfin, je tiens à remercier mes camarades préparationnaires orcéennes : Oriane et Jehanne – mes camarades de M2 – Nicolas, l'expert ès jeux de plateaux – la bande d'irréductible rôlistes : Maud, Lucie, Victor et Guillaume – mes coéquipiers d'*esport* : Nicolas et Roland – Pierre, pour sa culture BD et ses talents culinaires – l'équipe du livrescolaire, notamment Raphaël, Émilie, Isabelle, Pénélope, Julie et Aurélie.

Toute ma reconnaissance va aussi à ma famille, qui m'a énormément soutenu, et, notamment, à mes parents et mes sœurs. Pierre m'accompagne tous les jours, j'aurais tant aimé qu'il assiste à cette soutenance et je pense particulièrement à lui en rédigeant ces dernières lignes.

Enfin, pour l'essentiel, Marion est toujours là, et je la remercie infiniment d'être présente à mes côtés. Construire notre vie ensemble est, pour moi, la plus belle des réalisations.

Résumé

Cette thèse a pour objet la construction, la calibration et l'étude d'estimateurs multi-tâches, dans un cadre fréquentiste non paramétrique et non asymptotique.

Nous nous plaçons dans le cadre de la régression *ridge* à noyau et y étendons les méthodes existantes de régression multi-tâches. La question clef est la calibration d'un paramètre de régularisation matriciel, qui encode la similarité entre les tâches. Nous proposons une méthode de calibration de ce paramètre, fondée sur l'estimation de la matrice de covariance du bruit entre les tâches. Nous donnons ensuite pour l'estimateur obtenu des garanties d'optimalité, via une inégalité oracle, puis vérifions son comportement sur des exemples simulés.

Nous obtenons par ailleurs un encadrement précis des risques des estimateurs oracles multi-tâches et mono-tâche dans certains cas. Cela nous permet de dégager plusieurs situations intéressantes, où l'oracle multi-tâches est plus efficace que l'oracle mono-tâche, ou *vice versa*. Cela nous permet aussi de nous assurer que l'inégalité oracle force l'estimateur multi-tâches à avoir un risque inférieur à l'estimateur mono-tâche dans les cas étudiés. Le comportement des oracles multi-tâches et mono-tâche est vérifié sur des exemples simulés.

Mots-clefs : Calibration de paramètres ; Inégalité oracle ; Méthodes à noyau ; Multi-tâches ; Régression *ridge* ; Statistique fréquentiste ; Statistique non asymptotique ; Statistique non paramétrique

Abstract

Multi-task statistical learning

This thesis aims at constructing, calibrating and studying multi-task estimators, in a frequentist non-parametric and non-asymptotic framework.

We consider here kernel ridge regression and extend the existing multi-task regression methods in this setting. The main question is the calibration of a matricial regularization parameter, which encodes the similarity between the tasks. We propose a method to calibrate this parameter, based on the estimation of the covariance matrix of the noise between tasks. We then show optimality guarantees for the estimator thus obtained, via an oracle inequality. We also check its behaviour on simulated examples.

We carefully bound the risks of both multi-task and single-task oracle estimators in some specific settings. This allows us to discern several interesting situations, whether the multi-task oracle outperforms the single-task one or not. This ensure the oracle inequality enforces the multi-task oracle to have a lower risk than the single-task one in the studied settings. Finally, we check the behaviour of the oracles on simulated examples.

Keywords : Frequentist statistics ; Kernel methods ; Multi-task ; Non-asymptotic statistics ; Non-parametric statistics ; Oracle inequality ; Parameter calibration ; Ridge regression

Table des matières

Chapitre 1

Introduction

Abstract. We present here, in French, a shortened version of the work contained in this thesis. The contributions of this work to our field of research can be found, in English, in Chapter 2. The results themselves can be found in the following chapters, which are also written in English. Chapter 3 corresponds to the work published by Solnon et al. [SAB12], while Chapter 4 corresponds to an article that is being submitted.

Nous présentons ici, en français, les travaux contenus dans cette thèse. Le chapitre 2 contient un résumé des contributions de cette thèse à notre domaine de recherche, écrit en anglais. Les chapitres suivants, aussi écrits en anglais, contiennent les résultats proprement dits. Le chapitre 3 reprend les travaux publiés par Solnon et al. [SAB12], tandis que le chapitre 4 correspond à un article en cours de soumission.

1.1 Présentation du domaine

Le lecteur érudit en statistique pourra se référer directement aux parties suivantes, cette partie-ci présentant principalement le domaine statistique de manière historique et peu technique. Pour le lecteur non mathématicien, cela risque, hélas, d'être la seule partie aisément compréhensible [1]. Nous essaierons donc de retarder le plus possible l'apparition des détails techniques et de les limiter autant que faire se peut !

1.1.1 Introduction à la statistique

Le mot « statistique » est attribué, par exemple par Littré, à Gottfried Achenwall [2]. Il est employé aujourd'hui en de nombreux sens, qui ne correspondent pas tous, loin s'en faut, au sujet de cette thèse. Le dictionnaire de l'Académie française, que ce soit dans sa

1. Les parties les plus accessibles au lecteur non mathématicien sont la partie 1.1.1, le début de la partie 1.1.2 ainsi que la partie 1.1.4

2. Gottfried Achenwall : économiste allemand (1719-1772)

1.1. PRÉSENTATION DU DOMAINE

sixième (1835) [3] ou dans sa huitième (1932-1935) [4] édition, en donne un sens (compilation de données numériques concernant un État) bien éloigné de sa signification mathématique actuelle, tandis que les éditions précédentes ne le mentionnent pas, et pour cause : le mot « statistique » n'existait pas encore ! Littré donne, lui, plusieurs sens à ce mot [5], dont le second nous intéressera plus particulièrement : « science des dénombrements et de leurs conséquences ». Enfin, le Trésor de la Langue Française donne au nom « statistique » trois sens [6], dont le second semble encore pouvoir caractériser notre domaine : « Branche des mathématiques ayant pour objet l'analyse (généralement non exhaustive) et l'interprétation de données quantifiables ».

Nous dirons donc « statistique » pour parler de statistique mathématique. Il s'agit, à partir d'observations d'un processus, d'inférer certaines de ses propriétés afin d'aider à son étude ou bien pour permettre de prédire les futures réalisations de ce processus. Le processus en question pourra être une construction mathématique abstraite, dans ce cas le statisticien s'attachera à développer et analyser des méthodes générales capables de s'appliquer à de nombreux cas concrets. Il pourra aussi provenir d'une des nombreuses sciences qui utilisent des outils statistiques, le statisticien devra alors utiliser et adapter ces méthodes aux particularités du problème étudié. La réalité du travail d'un statisticien se situe d'ailleurs bien souvent entre ces deux points extrêmes, et cela constitue la richesse des statistiques que de mettre en contact de si nombreux domaines. On verra par exemple que le modèle que nous utiliserons, la régression *ridge*, trouve son origine dans une publication d'un journal d'ingénierie chimique !

3. STATISTIQUE. s. f. Science qui apprend à connaître un État sous les rapports de son étendue, de sa population, de son agriculture, de son industrie, de son commerce, etc.
Il signifie aussi, description détaillée d'un pays relativement à son étendue, à sa population, à ses ressources agricoles et industrielles, etc. [...]
Il s'emploie aussi adjectivement ; et alors il est des deux genres. [...]

4. STATISTIQUE. n. f. T. didactique. Science qui a pour objet de recueillir et de dénombrer les divers faits de la vie sociale. [...]
Il désigne encore la description détaillée d'un pays relativement à son étendue, à sa population, à ses ressources agricoles et industrielles, etc. [...]
Il s'emploie aussi comme adjectif des deux genres. [...]

5. STATISTIQUE s. f. (sta-ti-sti-k')

i Science qui a pour but de faire connaître l'étendue, la population, les ressources agricoles et industrielles d'un État. Achenwall, qui vivait vers la fin du milieu du XVIIIe siècle, est généralement considéré comme le premier écrivain systématique sur la statistique, et on dit que c'est lui qui lui a donné son nom actuel.

ii Plus généralement, science des dénombrements et de leurs conséquences. [...]

iii Description d'un pays relativement à son étendue, à sa population, à ses ressources agricoles et industrielles, etc. [...]

6. STATISTIQUE, subst. fém. et adj.

I Subst. fém.

A Recueil de données numériques concernant des faits économiques et sociaux. [...]

B Branche des mathématiques ayant pour objet l'analyse (généralement non exhaustive) et l'interprétation de données quantifiables. [...]

C Ensemble de données numériques (généralement analysées et interprétées) concernant une catégorie de faits. [...]

[...]

Nous pouvons prendre pour exemple un problème classique en vision artificielle : une collection d'images étant disponible (c'est *l'échantillon*), on doit apprendre à un ordinateur à reconnaître un type d'objet particulier (par exemple, une voiture) dans ces images (décrites, par exemple, pixel par pixel). Le mot « reconnaître » est ici flou et peut prendre plusieurs significations : dire si une voiture est présente ou non dans une image (on parlera alors de classification), donner les coordonnées d'un rectangle contenant une voiture, donner un ensemble de pixels formant la voiture, etc. Apprendre n'est, bien sûr, pas à prendre au premier degré, mais signifie que l'on doit construire un algorithme pouvant prendre en entrée n'importe quelle collection d'images (ces images peuvent contenir les informations recherchées ou non) afin de répondre à la question posée sur les images données ou bien sur des futures images. Cet algorithme, ou *estimateur*, doit donc pouvoir s'adapter à de nouvelles données et doit être le plus précis possible (on cherchera, par exemple, à réduire le nombre d'erreurs de classification, ou bien à réduire les portions de voiture non détectées). Pour effectuer sa tâche le statisticien doit décider d'un cadre théorique, le *modèle*, dans lequel il élaborera ses estimateurs, dont le choix se fait en fonction des contraintes apportées par le problème étudié et par le but recherché.

Nous introduirons dans la partie suivante quelques modèles, en partant des plus simples pour aboutir à ceux qui nous intéressent ici.

1.1.2 Quelques modèles de régression

Nous nous intéressons ici à un type de problèmes statistiques bien particulier : la régression. On observe ici des couples de points, que l'on notera (X_1, Y_1), (X_2, Y_2), etc. Les points X_i sont les covariables et sont censées décrire le phénomène que l'on étudie. Les points Y_i, eux, représentent la quantité qui nous intéressent dans un problème donné. On suppose alors que les données Y_i peuvent être expliquées par les covariables, c'est-à-dire qu'il existe une fonction f telle que $f(X_i)$ est proche de Y_i. C'est ce que l'on appelle un modèle génératif, et l'on suppose bien souvent que la fonction f possède certaines propriétés (par exemple, c'est une fonction linéaire, polynomiale, etc.). La différence entre les valeurs $f(X_i)$ et les observations Y_i est alors appelée « bruit », et peut avoir plusieurs origines :

- des erreurs ou des imprécisions ont été faites dans la mesure de Y_i ;
- des facteurs expliquant les observations Y_i ont été oubliés dans le groupe des covariables ou n'ont pas été observés ;
- l'hypothèse faite sur la fonction f est fausse.

On peut voir dans l'exemple de la figure 1.1 le cas affine : les points (X_i, Y_i) sont dans le plan $\mathbb{R}^2$, et l'on suppose que la fonction f est affine.

Nous verrons ensuite comment approcher cette fonction f dans plusieurs cas.

Régression par moindres carrés

Comme nous l'avons vu précédemment, le statisticien élabore ses estimateurs à partir d'un modèle, qu'il fixe en fonction du problème étudié. L'un des modèles les plus simples est peut-être celui de la régression linéaire. Prenons comme exemple celui de la dimension un, qui a l'avantage d'être plus aisément représentable (voir par exemple la figure 1.1) : sont donnés des couples de nombres réels (X_i, Y_i) (c'est l'échantillon) et l'on essaie d'y ajuster

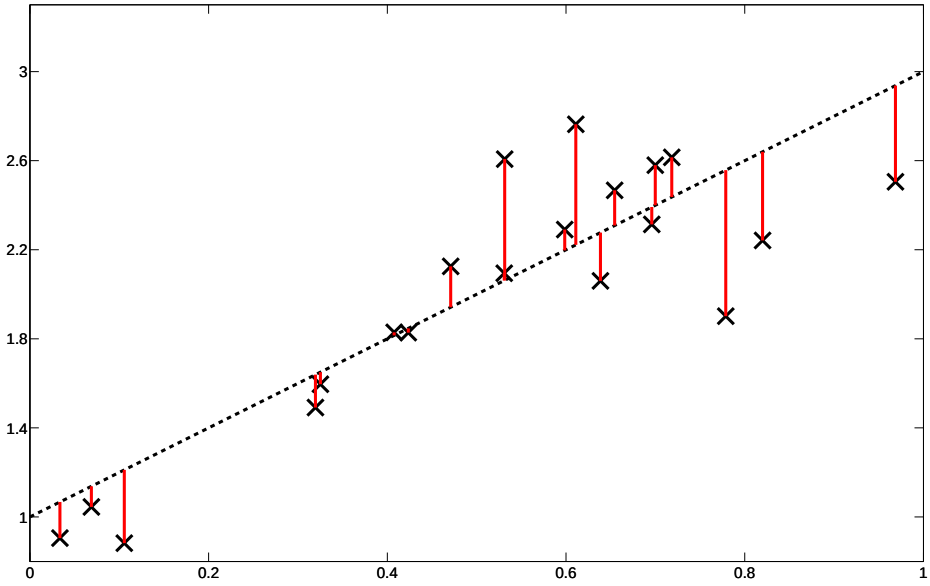

FIGURE 1.1 – Modèle génératif : l'échantillon est représenté par les points, la fonction f est représentée par la droite en pointillés. Le bruit est alors représenté par les segments rouges verticaux.

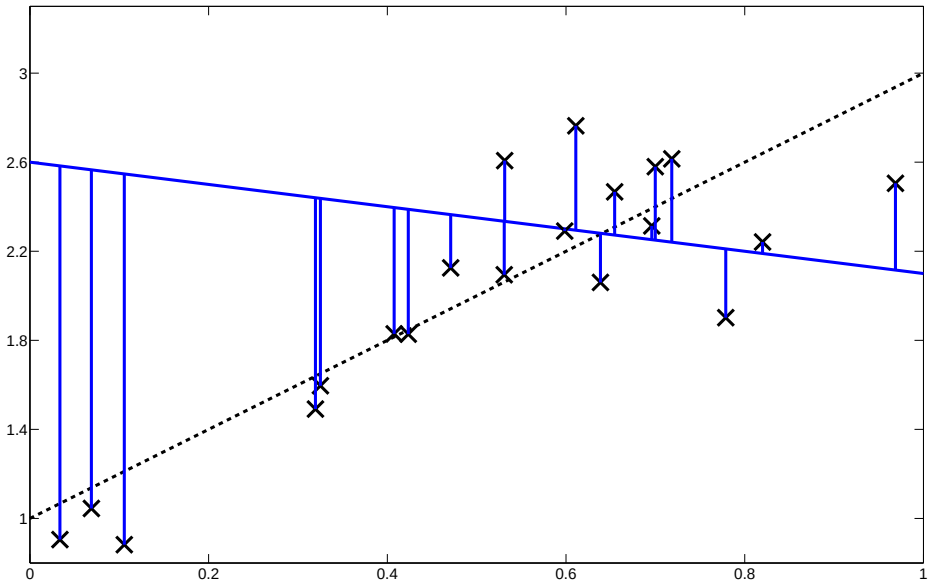

FIGURE 1.2 – Illustration de la méthode des moindres carrés : l'échantillon est représenté par les points, la droite ayant généré les données est la droite en pointillés. La droite de régression est la droite bleue et les erreurs sont les segments bleus verticaux. On cherche donc à minimiser la somme des carrés des longueurs des segments bleus.

une droite [7] au nuage de points (X_i, Y_i) (c'est le modèle). Quelle droite peut-on alors choisir, qui s'ajustera au mieux sur le nuage de points ? Une réponse classique sera donnée par la méthode des moindres carrés, introduite par Legendre [Leg05] mais dont la paternité est aussi attribuée à Gauss [Gau09]. Il s'agit de considérer, pour une droite quelconque, la somme des carrés des différences entre les observations Y_i et les projections verticales de ces observations sur la droite. On considère ensuite la droite qui minimise la somme de ces carrés, d'où le nom de la méthode. On dit alors que l'on a effectué une régression. Une illustration de cette méthode est donnée dans la Figure 1.2. Un avantage de cette méthode est de donner un estimateur qui s'exprime explicitement et simplement et qui possède de surcroît une interprétation simple dans le cadre des modèles linéaires gaussiens.

Généralisons l'exemple précédent. Supposons que l'on observe n points X_i, chaque point ayant k coordonnées (avec $k < n$). Pour chaque point X_i, on observe un nombre réel Y_i. On peut donc former la matrice X, de taille $n \times k$, dont la ligne i est constituée des coordonnées de X_i ainsi que le vecteur Y dont la $i$ème coordonnée est Y_i. Le problème de régression se pose donc ainsi : trouver un vecteur β (de taille k) tel que le vecteur $X\beta$ approche au mieux Y, par exemple en termes de distance euclidienne. La méthode des moindres carrés

7. Droite, car nous sommes en dimension 1. En plus grande dimension, c'est-à-dire si les X_i sont des k-uplets, on cherchera un hyperplan dans l'espace de dimension $k + 1$.

vise alors à chercher une solution qui minimise la quantité

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-(X\beta)_i)^2=\frac{1}{n}(Y-X\beta)^\top(Y-X\beta)\ . \tag{1.1}$$

La solution à ce problème de minimisation se trouve aisément [8] et l'on obtient comme valeur de β :

$$\widehat{\beta}=\left(X^\top X\right)^{-1}X^\top Y\ .$$

On remarquera que la lettre β est chapeautée, notation traditionnelle en statistique pour signaler qu'il s'agit de l'estimateur du paramètre. Cet estimateur est aussi l'estimateur du maximum de vraisemblance dans le modèle linéaire gaussien, mais nous ne rentrerons pas dans ces détails. Remarquons aussi que la quantité (1.1), que l'on minimise en effectuant cette méthode, porte souvent le nom de « risque empirique ». Le risque (quadratique) étant l'espérance du carré de la différence entre l'estimateur et la fonction recherchée, la somme des carrés des erreurs est donc bien, en effet, la version « empirique » du risque. Cette quantité apparaîtra de nombreuses fois par la suite, chaque fois sous cette dénomination.

L'estimateur que nous avons présenté est l'un des plus simples et des plus connus que l'on puisse trouver en statistique. S'il fonctionne raisonnablement bien sur des problèmes simples, il possède de nombreux défauts. Certains de ces défauts ne nous intéressent pas directement ici, comme par exemple la limitation $k < n$ (une part importante des publications de ces vingt dernières années a concerné le développement de méthodes de régression en haute dimension, souvent en utilisant des outils de parcimonie).

En ce qui nous concerne, un premier défaut est l'utilisation de l'inverse de la matrice $X^\top X$. Quand cette matrice possède de très petites valeurs propres (on dira qu'elle est mal conditionnée, et l'on notera $\lambda_{\min}$ la plus petite de ses valeurs propres), deux phénomènes apparaissent.

- Numériquement, l'inversion de la matrice $X^\top X$ est instable, le calcul de $\widehat{\beta}$ par un ordinateur peut donc donner des solutions aberrantes.
- Sous une hypothèse de normalité [9] du bruit, le risque de l'estimateur $\widehat{\beta}$, défini par $\mathbb{E}\left[\left((\widehat{\beta}-\beta_0)^\top(\widehat{\beta}-\beta_0)\right)^2\right]$, vaut $\sigma^2\operatorname{tr}(X^\top X)^{-1}$, dont une minoration est $\sigma^2\frac{n}{\lambda_{min}}$. Si $\lambda_{\min}$ est trop petit, l'estimation de β_0 par $\widehat{\beta}$ risque d'être très mauvaise. Notamment, les coordonnées de $\widehat{b}$ risquent d'être beaucoup trop grandes.

La méthode de régression *ridge* fut développée pour pallier ces limitations, comme nous l'expliquerons plus tard.

Un deuxième défaut est la limitation du modèle aux relations linéaires. En poursuivant notre exemple, tous les nuages de points ne sont pas alignés sur des droites ! Nous verrons ensuite que les outils de régression à noyau permettent d'obtenir des modèles moins contraignants.

8. On suppose pour simplifier que $XX^\top$ est inversible.

9. Cela signifie que l'on suppose qu'il existe un vecteur β_0 tel que le vecteur des erreurs $Y - X\beta_0$, c'est-à-dire le bruit, suit une loi gaussienne — ou normale — $\mathcal{N}(0, \sigma^2 I_n)$, où σ^2 est la variance du bruit.

Régression *ridge*

La régression *ridge*[10] vise à répondre au premier problème soulevé sur la méthode des moindres carrés, lié à l'inversion de $X^\top X$. Comme l'explique R. Hoerl [Hoe85][11], sur lequel nous nous appuyons fortement, l'analyse *ridge* fut développée par A. Hoerl [Hoe59] afin de permettre l'étude géométrique graphique des surfaces quadratiques dépendant d'un grand nombre de variables, ce qui répondait à un besoin industriel concret. Rappelons brièvement de quoi relève cette analyse. On étudie les propriétés de la surface définie par l'équation

$$Y = b_0 + b^\top x + \frac{1}{2} x^\top B x \ , \tag{1.2}$$

où b_0 est un nombre réel, b et x des vecteurs de taille k et B une matrice symétrique de taille k, sous la contrainte $x^\top x \leq C^2$ (les variables ont été au préalable recentrées et renormalisées, d'où la contrainte sphérique). On étudie alors le problème de la maximisation (ou de la minimisation) de l'équation (1.2) contrainte par $x^\top x \leq R^2$, où $\mathbb{R}^2 \in [0, C^2]$. L'ensemble des solutions x de ce problème de maximisation (respectivement, de minimisation) est appelée la crête maximale (respectivement, minimale), car il est un lieu d'optimums locaux. En introduisant le multiplicateur de Lagrange λ de la contrainte, on obtient alors

$$x = -(B - \lambda I_k)^{-1} b \ ,$$

quand cette expression est bien définie, c'est-à-dire hors des valeurs propres de B. En fonction de la répartition des valeurs propres de B, on obtient donc un certain nombre de courbes, paramétrées par λ. Ces courbes sont aussi appelées crêtes, la crête maximale (respectivement, minimale) s'obtenant pour les paramètres supérieurs (respectivement, inférieurs) à la plus grande valeur (respectivement, la plus petite) propre de B. Nous renvoyons à l'étude de R. Hoerl [Hoe85] pour un exemple d'utilisation de ces crêtes, qui permettent d'obtenir facilement des informations qualitatives sur la surface étudiée.

Revenons maintenant à notre problème de régression. Peu après avoir construit l'analyse *ridge*, A. Hoerl [Hoe62] remarqua que la somme des carrés résiduels de la méthode des moindres carrés était une forme quadratique du coefficient de la régression, β. En appliquant sa méthode d'analyse à ce cas-là, il put suivre les crêtes alors construites pour obtenir des valeurs de paramètres plus stables. Il restait un problème, épineux : quel point choisir sur la crête ? A. Hoerl et Kennard [HK70], dans un article qui est souvent cité comme source de la régression *ridge*, analysent ce problème et montrent qu'un choix existe toujours, et qu'il est meilleur que la solution initiale, l'estimateur des moindres carrés, en terme de risque quadratique (c'est-à-dire, en considérant la somme des carrés des résidus). On peut alors obtenir la forme moderne de l'estimateur *ridge* :

$$\widehat{\beta}_{ridge} = (X^\top X + \lambda I_k)^{-1} X^\top Y \ ,$$

le paramètre λ étant à choisir. Cet estimateur minimise la quantité suivante :

$$(Y - X\beta)^\top (Y - X\beta) + \lambda \beta^\top \beta \ . \tag{1.3}$$

10. Suivant l'usage, nous garderons l'expression anglaise et ne traduirons pas *ridge* par son équivalent français : crête.

11. Note au lecteur : deux auteurs se partagent ici le patronyme de Hoerl : Arthur E. Hoerl, à qui correspondent les articles de 1959, 1962 et 1970 et Roger W. Hoerl, à qui correspond l'article de 1985.

On voit donc apparaître une des motivations de cet estimateur, qui a fait sa célébrité : l'utiliser revient à pénaliser le risque empirique (c'est $n^{-1}(Y-X\beta)^\top(Y-X\beta)$) par le carré de la norme de β ce qui, contrairement au cas des moindres carrés, devrait empêcher l'estimateur résultant d'avoir une trop grande norme. Mentionnons finalement une version populaire de cet estimateur, dans le cadre du *design*[12] fixe. On cherche alors à retrouver les valeurs de sortie sur les points X_i, et non plus à comprendre le processus dans sa globalité. Cela revient donc à essayer d'enlever le bruit sur ces observations. On considère donc l'estimateur à *design* fixe $X\widehat{\beta}_{ridge}$, qui prend la forme suivante, grâce à un tour de passe-passe d'algèbre linéaire :

$$X\widehat{\beta}_{ridge} = X(X^\top X + \lambda I_k)^{-1}X^\top Y = XX^\top(XX^\top + \lambda I_n)^{-1}Y \ .$$

Régression *ridge* à noyau

Nous nous intéressons maintenant à la seconde limitation de l'estimateur des moindres carrés que nous avions citée, et nous verrons que les solutions proposées ont un lien fort avec la régression *ridge*. Il s'agit donc de s'affranchir de la contrainte de linéarité imposée par l'estimateur des moindres carrés. On pourrait, par exemple, vouloir ajuster sur le nuage de points, non pas une droite ou un hyperplan, mais une fonction lisse[13]. Tel est l'objectif du lissage par les *splines*[14]. On trouvera de nombreux détails sur ces *splines* dans le livre de Wahba [Wah90] (dans lequel nous avons puisé la plupart de nos références), le livre de Gu [Gu02] étant aussi une excellente référence. Étant donné une subdivision $\sigma = \{x_0, \dots, x_n\}$ d'un intervalle $[a, b]$, des valeurs $y_0, \dots, y_n$ et une régularité $m \in \mathbb{N}$, une *spline* est la fonction qui interpole les points de σ aux valeurs y_i, en respectant des conditions de régularité globale et des conditions de régularité aux raccordements. C'est aussi la fonction f qui minimise $\int_a^b (f^{(m)}(x))^2 dx$ sur l'espace de Sobolev W_m (ce sont les fonctions $m-1$ fois continûment dérivables, et dont la dérivée $m$ième est de carré intégrable) avec la contrainte que, pour tout i, $f(x_i) = y_i$. On se déporte alors de l'interpolation vers la régularisation comme ceci : on cherche une fonction dans un espace régulier (par exemple, dans un espace de Sobolev) qui possède à la fois de bonnes propriétés de régularité et qui soit toujours proche des observations. Cela nous amène naturellement à considérer la fonction qui minimise sur l'espace considéré la quantité suivante :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda \int_a^b (f^{(m)}(x))^2 dx \ , \tag{1.4}$$

au prix de l'introduction du paramètre λ, ce dont nous parlerons plus tard. La ressemblance avec la formulation de l'estimateur *ridge* de l'équation (1.3) n'est bien entendu pas fortuite, nous en reparlerons là aussi plus tard.

Une question peut alors légitimement se poser : l'estimateur que l'on obtient en optimisant l'équation (1.4) est-il aisément calculable ? Fort heureusement, les espaces $\mathcal{F}$ considérés

12. On se pliera ici à l'usage en gardant le mot anglais *design*, que l'on pourrait traduire par « plan d'expérience ».

13. On entendra par lisse, très classiquement, une fonction possédant de fortes propriétés de régularité, comme le vérifient par exemple les fonctions des espaces de Sobolev W_m, de faible norme.

14. On gardera ici le mot anglais. Une *spline* est une baguette souple, utilisée en construction navale pour interpoler plusieurs points à la main. On fixait les baguettes aux points désirés, puis on reliait les baguettes, ce qui crée une interpolation naturelle.

avec nos *splines* possèdent tous une structure particulière : ce sont des espaces de Hilbert, possédant une fonction de description $\Phi : \mathbb{R} \to \mathcal{F}$ vérifiant la propriété

$$\forall f \in \mathcal{F},\ \forall x \in \mathbb{R}, f(x) = \langle \Phi(x), f \rangle \tag{1.5}$$

(d'où, d'ailleurs, le nom de représentation). Les espaces possédant ce type de structure ont été étudiés par Aronszajn [Aro50] et l'on peut en donner une définition ainsi que quelques propriétés simples. Soit $\mathcal{X}$ un ensemble quelconque et $\mathcal{F}$ un espace de Hilbert de fonctions sur $\mathcal{X}$ à valeurs réelles. On dira que $\mathcal{F}$ est un espace de Hilbert à noyau auto-reproduisant (RKHS [15]) si, pour tout élément x de $\mathcal{X}$, l'application $f \mapsto f(x)$ est continue. On peut alors montrer qu'il existe une fonction Φ vérifiant la propriété (1.5), que l'on appelle fonction de description [16]. De cette fonction de description, on peut ensuite construire le noyau k (au sens de Mercer) par $\forall (x, y) \in \mathcal{X}^2, k(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Remarquons que l'ordre de construction est arbitraire, et que l'on peut déduire la fonction de description du noyau, par la propriété $\Phi(x)(y) = k(x, y)$. On voit bien, alors, la notion de reproductibilité : en notant $\Phi(x) = k(x, \cdot)$, on a

$$\forall (x, y) \in \mathcal{X}^2,\ k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle\ . \tag{1.6}$$

Le fameux *kernel trick* [17] vient de cette propriété : on pourra calculer, dans $\mathcal{F}$, les produits scalaires entre les observations $\Phi(X_i)$ via les $k(X_i, X_j)$ uniquement, de là la popularité de ces outils.

Une fois ces outils analysés, on peut donc retourner le problème de départ. On ne cherche plus une *spline*, mais un élément d'un espace de Hilbert caractérisé par son noyau. On étend alors le problème de minimisation (1.4) au suivant : trouver dans le RKHS donné une fonction minimisant

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \|f\|^2\ ,$$

la norme ici écrite étant celle du RKHS. Un simple raisonnement, connu sous le nom de théorème du représentant, montre qu'une solution de ce problème est atteinte sur l'espace vectoriel engendré par les fonctions $\Phi(X_i)$. On pourra donc aisément exprimer un tel estimateur, les calculs se faisant simplement grâce à la propriété (1.6). Enfin, l'estimateur de *design* fixe s'exprime aussi très simplement ici. En notant K la matrice du noyau, dont les coefficients sont $k(X_i, X_j)$, on obtient l'estimateur

$$K(K + \lambda I_n)^{-1} Y\ .$$

L'analogie avec le cas de l'estimateur *ridge* est donc complète, l'un pénalisant le risque empirique par la norme des vecteurs pour obtenir un estimateur plus petit, l'autre pénalisant le risque empirique par la norme de la fonction de régression pour obtenir un estimateur plus lisse. Dans le premier cas, l'estimateur en *design* fixe ne dépend que des $\langle X_i, X_j \rangle$, via $XX^\top$, tandis que, dans le second, il ne dépend que des $k(X_i, X_j)$. C'est donc pour cela que le second estimateur porte le nom d'estimateur *ridge* à noyau, qui nous semble plus répandu aujourd'hui que celui de *spline*. Dans les deux cas, on appellera régularisation l'action qui

15. De l'anglais reproducing kernel Hilbert space.
16. En anglais, *feature map*.
17. On conservera le terme anglais, qui signifie « astuce du noyau ».

consiste à pénaliser le risque empirique par le carré de la norme de l'estimateur. Enfin, il faut bien remarquer qu'aucun tour de magie n'a été effectué ici. Bien que ces estimateurs aient des propriétés agréables que ne possède pas l'estimateur des moindres carrés, cela se fait au prix de l'introduction d'un paramètre, ici noté λ, que l'on appelle souvent « paramètre de régularisation ». Le choix de ce paramètre, on parle de calibration, incombe donc au statisticien, et le choix arbitraire ou « au doigt mouillé » n'est guère satisfaisant. On a donc résolu un problème en en posant un autre, aussi ardu. Nous discutons dans la partie suivante des méthodes possibles pour réaliser cette tâche.

1.1.3 Choisir un modèle, ou calibrer son estimateur

Nous avons vu dans la partie précédente que les modèles introduits pour pallier les défauts de la régression par moindres carrés introduisent chacun un nouveau paramètre, le paramètre de régularisation. Le statisticien doit alors calibrer ce paramètre, et donc comparer, en un certain sens, les estimateurs obtenus pour chaque paramètre. Si l'on considère qu'un paramètre définit un modèle, il faut ici choisir parmi une famille de modèles : c'est ce que l'on appelle la « sélection de modèles ». Or, pour pouvoir comparer deux estimateurs, il faut s'accorder sur une mesure commune, qui marquera l'efficacité de ces estimateurs. Nous rappelons que ces estimateurs sont censés approcher des quantités qui sont, implicitement, inhérentes au processus étudié. Nous allons maintenant expliciter ces quantités. Il faut bien noter que l'étape que nous allons effectuer, qui est le premier pas de la modélisation mathématique d'un problème statistique, est une abstraction de ce problème, et peut donc être assez éloignée des données réelles étudiées.

Commençons par décrire ce processus de manière non technique. On suppose que les points X_i sont tirés indépendamment, selon un certain processus (comme si ce processus avait été recopié, sans tenir compte des autres copies), et que les points Y_i sont les images des X_i par une fonction f, perturbées par un bruit. La modélisation mathématique de cela s'énonce comme suit.

> Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, $\mathcal{X}$ un ensemble et $(X_i)_{i=1}^n$ une suite de variables indépendantes et de même loi $\mathcal{P}$. On suppose alors qu'il existe une fonction $f : \mathcal{X} \mapsto \mathbb{R}$, dans un ensemble $\mathcal{F}$, telle qu'il existe une suite $(\varepsilon_i)_{i=1}^n$ de variables aléatoires indépendantes et centrées, indépendante de (X_i), telle que
>
> $$\forall i \in \{1, \dots, n\},\ Y_i = f(X_i) + \varepsilon_i \ .$$

L'objectif est maintenant clair : retrouver f par les observations (X_i, Y_i). La qualité d'un estimateur $\widehat{f}$ sera donc mesurée par sa distance avec f, que l'on peut définir de plusieurs façons, que l'on appellera à chaque fois « risque » :

Prédiction en *design* fixe : Le risque est $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \widehat{f}(X_i))^2 \middle| X_1, \dots, X_n\right]$. Ici, on cherche donc à retrouver les valeurs de $f(X_i)$, c'est-à-dire à enlever le bruit aux observations Y_i.

Prédiction en *design* aléatoire : Si X_{n+1} est une variable aléatoire de loi $\mathcal{P}$, indépendante de $(X_i)_{i=1}^n$, on considère le risque $\mathbb{E}\left[(f(X_{n+1}) - \widehat{f}(X_{n+1}))^2\right]$. Ici, on cherche à prédire ce que donnerait une nouvelle observation du processus.

Ces risques mènent bien entendu à des analyses différentes ! Plusieurs remarques s'imposent d'emblée.

- On peut considérer ces risques sans espérance, on parle alors de perte. Dans ce cas, on cherchera à contrôler cette perte sur un ensemble de haute probabilité.
- Le choix du carré dans chaque risque (on parle de perte quadratique) résulte de plusieurs nécessités (simplicité de calcul, d'interprétation, etc.). On aurait pu choisir une autre perte, ce qui mène alors à d'autres estimateurs.
- Nous n'avons cité que des risques de prédiction, mais on peut s'intéresser aussi à ceux d'estimation, qui visent à connaître certaines propriétés de f, sans forcément de rapport avec la loi de l'échantillon.
- On remarquera que l'on se situe ici dans un cadre fréquentiste [18] très classique. On ne parlera pas (ou fort peu), ici, de méthodes bayésiennes [19].

Le but est, maintenant, de trouver parmi une collection d'estimateurs $\left\{\widehat{f}_\lambda, \lambda \in \Lambda\right\}$ un estimateur minimisant le risque choisi.

Dans certains cas, avec une perte quadratique, on peut commencer par étudier la forme de ce risque en réalisant une « décomposition biais-variance ». Prenons comme exemple le cas de la régression *ridge* à noyau en *design* fixe. On suppose, pour simplifier, que la suite $(X_1, \dots, X_n)$ est déterministe et l'on veut donc estimer le vecteur $f = (f(X_1), \dots, f(X_n))^\top$. On notera, pour un vecteur u de taille n, $\|u\|_n^2 = n^{-1}\sum_{i=1}^n u_i^2$ et le vecteur de bruit sera noté $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. L'estimateur est alors $\widehat{f}_\lambda = A_\lambda Y = K(K + \lambda I_n)^{-1}Y$. On a alors décomposer le risque de la façon suivante :

$$\mathbb{E}\left[\left\|\widehat{f}_\lambda - f\right\|^2\right] = \mathbb{E}\left[\|A_\lambda(f+\varepsilon) - f\|^2\right] = \underbrace{\mathbb{E}\left[\|(A_\lambda - I_n)f\|^2\right]}_{\text{Biais}} + \underbrace{\mathbb{E}\left[\|A_\lambda \varepsilon\|^2\right]}_{\text{Variance}} \ .$$

Le biais représente ici la proximité entre l'estimateur et la fonction estimée, tandis que la variance représente la variabilité apportée par cet estimateur. Un bon choix d'estimateur consistera donc à réaliser un compromis entre ces deux quantités. On remarquera ensuite que, dans le cas $\lambda = 0$, le biais est nul. Régulariser l'estimateur revient donc à biaiser celui-ci, en espérant réduire suffisamment la variance afin de compenser ce biais.

Le risque, ainsi que sa décomposition biais-variance quand elle existe, ne nous sont pas connus, car nous n'avons accès qu'à l'échantillon $(X_i, Y_i)_{i=1}^n$. La quantité s'en rapprochant le plus, à première vue, est le risque empirique,

$$\frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{f}_\lambda(X_i))^2 \ ,$$

que nous avions introduit lors de l'étude de la régression par les moindres carrés. Il peut alors sembler légitime de chercher le modèle (ou, c'est équivalent, le paramètre) qui minimise ce

18. Fréquentiste, car les résultats que l'on montre sont assurés de se réaliser avec une fréquence proche de 1.

19. En analyse bayésienne, on suppose que l'on a une connaissance *a priori* de f, qui peut être subjective, donnée par une distribution de probabilité. On cherche ensuite à calculer la distribution *a posteriori*, une fois l'échantillon observé. L'analyse fréquentiste n'est pas forcément plus objective que l'analyse bayésienne, car le choix de $\mathcal{F}$, le modèle, est toujours subjectif. Il est toutefois souvent difficile de concilier les deux approches, ou d'interpréter l'une en fonction de l'autre.

risque empirique (dans le cas de la régression *ridge*, cela revient, par construction, à prendre $\lambda = 0$). C'est, en fait, une mauvaise idée. En effet, les estimateurs $\widehat{f}_\lambda$ sont construits avec les échantillons $(X_i, Y_i)_{i=1}^n$. Il n'est pas raisonnable de réutiliser directement ces échantillons pour évaluer les performances de ces estimateurs. Cela pourrait mener à une évaluation trop optimiste du critère et à une mauvaise sélection du paramètre. C'est ce que l'on appelle le phénomène de sur-apprentissage [20]. On pourra dire, plus légèrement, qu'il n'est pas raisonnable qu'un échantillon soit à la fois juge et partie ! De manière plus technique, on peut aussi remarquer que, comme les estimateurs ne sont pas indépendants de l'échantillon, il n'y a pas de raison que l'espérance du risque empirique soit proche du risque de prédiction avec un *design* aléatoire. Nous présentons maintenant deux façons de remédier à ce problème.

Validation croisée

Nous l'avons dit, l'échantillon ne doit pas être juge et partie dans la construction et l'évaluation des estimateurs. Il peut alors sembler naturel de court-circuiter ce problème par la stratégie suivante : on sépare (aléatoirement [21]) l'échantillon en deux parties. On considère la première comme un nouvel échantillon, et l'on construit la famille d'estimateurs dessus : c'est « l'échantillon d'entraînement ». Sur la deuxième, on calcule l'erreur faite par chaque estimateur (avec le critère choisi) : c'est « l'échantillon de test ». En notant $\widehat{f}_{I,\lambda}$ l'estimateur construit sur l'échantillon $I \subset \{1, \dots, n\}$ avec le paramètre λ, I_{ent} l'échantillon d'entraînement et I_{test} l'échantillon de test (avec $I_{ent} \sqcup I_{test} = \{1, \dots, n\}$), cela revient à choisir

$$\widehat{\lambda} \in \operatorname*{argmin}_{\lambda \in \Lambda} \left\{ \frac{1}{|I_{test}|} \sum_{i \in I_{test}} \left(\widehat{f}_{I_{ent},\lambda}(X_i) - Y_i \right)^2 \right\} \ .$$

On dit alors que l'on a réalisé une validation simple, et l'on pourra se référer à l'introduction (en français) de la thèse de Arlot [Arl07], ou bien à l'étude (en anglais) de Arlot et Celisse [AC10], pour une description plus poussée de ces méthodes. Les estimateurs sont donc maintenant indépendants des données sur lesquelles on les évalue. On se contentera de dire que l'étude théorique de la validation simple est aisée, qu'elle permet de montrer des résultats théoriques puissants, mais qu'elle fonctionne mal en pratique. Une des raisons de cet échec pratique est que l'on considère un seul échantillon de taille moitié moindre que celle de l'échantillon initial.

La validation croisée *V-fold* permet d'étendre cette démarche en produisant un choix efficace en pratique, mais beaucoup plus difficile à étudier théoriquement. Pour la réaliser, on découpe l'échantillon initial en V blocs de taille égale, aléatoirement. On prend un bloc comme échantillon de test, et le reste des blocs comme échantillon d'entraînement. Comme précédemment, on calcule les estimateurs sur l'échantillon d'entraînement et on les évalue sur l'échantillon de test. Il ne reste plus qu'à réaliser cela en prenant à chaque fois un bloc différent comme échantillon de test. On a donc évalué V fois chaque estimateur (une fois par bloc) avec le critère choisi. Finalement, on choisit le paramètre qui minimise la moyenne

20. Connu en anglais comme *overfitting*.

21. L'hypothèse sur les variables aléatoires X_i est ici cruciale. Si on ne peut pas échanger deux observations, il est absurde d'essayer d'utiliser une méthode de validation. Il est donc important que les deux parties sélectionnées soient indépendantes.

de ces V évaluations. En notant $I_1, \dots, I_V$ la partition de $\{1, \dots, n\}$ en V blocs, et en reprennant les notations précédentes, on choisit donc le paramètre

$$\widehat{\lambda} \in \operatorname*{argmin}_{\lambda \in \Lambda} \left\{ \frac{1}{V} \sum_{k \in \{1,\dots,V\}} \frac{1}{|I_k|} \sum_{i \in I_k} \left(\widehat{f}_{\{1,\dots,V\} \setminus I_k, \lambda}(X_i) - Y_i \right)^2 \right\} \ .$$

Nous ne nous étendrons pas plus sur les méthodes de validation croisée, qui constituent un domaine de recherche très riche. Elles constitueront principalement un point de comparaison ultérieur pour nos algorithmes.

Sélection de modèle par pénalisation

Nous présentons maintenant une procédure moins générale de sélection de modèle. Volontairement, nous n'entrerons pas dans tous les détails de ce domaine en nous limitant au cas qui nous intéresse ici. Nous renvoyons le lecteur avide de détails au livre de Massart [Mas07]. Comme nous l'avions expliqué précédemment, nous avons une famille d'estimateurs $(\widehat{f}_\lambda)_{\lambda \in \mathbb{R}_+}$ et nous voulons choisir l'estimateur qui possède le plus petit risque, par exemple, en reprenant le critère utilisé pour le *design* fixe, $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i) - \widehat{f}(X_i))^2 | X_1, \dots, X_n\right]$. Or, comme nous l'avons vu, la quantité que nous pouvons calculer et qui se rapproche le plus de ce critère, le risque empirique, le sous-estime. On peut donc pénaliser ce risque empirique, par une pénalité qui dépendra, entre autres, de λ, ce qui donne la quantité

$$\mathrm{crit}(\lambda) = \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{f}_\lambda(X_i))^2 + \mathrm{pen}(\lambda) \ .$$

Mais comment choisir cette pénalité $\mathrm{pen}(\lambda)$? Nous suivrons ici de nombreux auteurs, à commencer par Akaike [Aka70], en considérant qu'une bonne pénalité doit donner un critère sans biais, c'est-à-dire que l'espérance de $\mathrm{crit}(\lambda)$ vaut le risque recherché. Dans le cadre de la régression en *design* fixe, cela donne

$$\mathbb{E}\left[\mathrm{crit}(\lambda) | X_1, \dots, X_n\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (f(X_i) - \widehat{f}(X_i))^2 \middle| X_1, \dots, X_n \right] \ .$$

Un choix possible pour une pénalité donnant un tel critère est donc naturellement

$$\begin{aligned} \mathrm{pen}_{\mathrm{id}}(\lambda) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (f(X_i) - \widehat{f}_\lambda(X_i))^2 \middle| X_1, \dots, X_n \right] \\ - \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{f}_\lambda(X_i))^2 \middle| X_1, \dots, X_n \right] \ , \end{aligned} \tag{1.7}$$

pénalité que l'on appellera « pénalité idéale ». Dans le cas de la régression *ridge* à noyau, cela donne, par exemple,

$$\mathrm{pen}_{\mathrm{id}}(\lambda) = \frac{2\sigma^2}{n} \operatorname{tr}(A_\lambda) \ .$$

Cette pénalité dépend malheureusement des données du problèmes auxquelles nous n'avons pas accès (par exemple, ici, on voit la variance du bruit σ^2 intervenir), mais il est parfois possible (et ce sera notre cas) de construire une pénalité $\widehat{\text{pen}}$, s'en approchant suffisamment. On considèrera alors le paramètre $\widehat{\lambda}$ minimisant le critère

$$\widehat{\text{crit}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}_\lambda(X_i))^2 + \widehat{\text{pen}}(\lambda) \ . \tag{1.8}$$

Il est en fait assez aisé de montrer qu'un choix judicieux de $\widehat{\text{pen}}$ peut mener à une bonne sélection de $\widehat{\lambda}$. Une fois n'est pas coutume, nous allons développer un court calcul, car il nous semble que ledit calcul est à la base de la plupart des résultats en sélection de modèle et qu'il n'utilise que les définitions (1.7) et (1.8). Pour simplifier les notations, nous adopterons les conventions suivantes : pour tout paramètre λ, on note $\left\|\widehat{f}_\lambda - f\right\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda(X_i) - f(X_i))^2$ et $\left\|\widehat{f}_\lambda - Y\right\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda(X_i) - Y_i)^2$ tandis que l'on suppose que la suite $(X_1, \ldots, X_n)$ est déterministe. On a alors, pour tout paramètre λ,

$$\begin{aligned}
\widehat{\text{crit}}(\widehat{\lambda}) &= \left\|Y - \widehat{f}_{\widehat{\lambda}}\right\|_n^2 + \widehat{\text{pen}}(\widehat{\lambda}) \\
&= \mathbb{E}\left[\left\|Y - \widehat{f}_{\widehat{\lambda}}\right\|_n^2\right] + \mathbb{E}\left[\widehat{\text{pen}}(\widehat{\lambda})\right] + \left\|Y - \widehat{f}_{\widehat{\lambda}}\right\|_n^2 - \mathbb{E}\left[\left\|Y - \widehat{f}_{\widehat{\lambda}}\right\|_n^2\right] + \widehat{\text{pen}}(\widehat{\lambda}) - \mathbb{E}\left[\widehat{\text{pen}}(\widehat{\lambda})\right]
\end{aligned}$$

Notons $\Delta(\lambda) = \left\|Y - \widehat{f}_\lambda\right\|_n^2 - \mathbb{E}\left[\left\|Y - \widehat{f}_\lambda\right\|_n^2\right] + \widehat{\text{pen}}(\lambda) - \mathbb{E}\left[\widehat{\text{pen}}(\lambda)\right]$. On a alors

$$\begin{aligned}
\widehat{\text{crit}}(\widehat{\lambda}) &= \mathbb{E}\left[\left\|Y - \widehat{f}_{\widehat{\lambda}}\right\|_n^2\right] + \mathbb{E}\left[\widehat{\text{pen}}(\widehat{\lambda})\right] + \Delta(\widehat{\lambda}) \\
&= \mathbb{E}\left[\left\|\widehat{f}_{\widehat{\lambda}} - f\right\|_n^2\right] + \mathbb{E}\left[\widehat{\text{pen}}(\widehat{\lambda})\right] - \text{pen}_{\text{id}}(\widehat{\lambda}) + \Delta(\widehat{\lambda}) \\
&\leq \mathbb{E}\left[\left\|\widehat{f}_\lambda - f\right\|_n^2\right] + \mathbb{E}\left[\widehat{\text{pen}}(\lambda)\right] - \text{pen}_{\text{id}}(\lambda) + \Delta(\lambda)
\end{aligned}$$

Il en découle donc que

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f}_{\widehat{\lambda}} - f\right\|_n^2\right] + \mathbb{E}\left[\widehat{\text{pen}}(\widehat{\lambda})\right] - \text{pen}_{\text{id}}(\widehat{\lambda}) + \Delta(\widehat{\lambda}) \\
&\qquad\qquad \leq \inf_\lambda \left\{\mathbb{E}\left[\left\|\widehat{f}_\lambda - f\right\|_n^2\right] + \mathbb{E}\left[\widehat{\text{pen}}(\lambda)\right] - \text{pen}_{\text{id}}(\lambda) + \Delta(\lambda)\right\} \ .
\end{aligned}$$

Il nous manque donc quelques conditions pour pouvoir contrôler la qualité de notre estimateur :

1. uniformément en λ, $|\Delta(\lambda)|$ doit être petit ;
2. uniformément en λ, $\mathbb{E}\left[\widehat{\text{pen}}(\lambda)\right] \geq \text{pen}_{\text{id}}(\lambda)$;
3. uniformément en λ, $\mathbb{E}\left[\widehat{\text{pen}}(\lambda)\right] - \text{pen}_{\text{id}}(\lambda)$ doit être petit.

Si ces conditions sont vérifiées, on aboutira alors à une « inégalité oracle », c'est-à-dire du type

$$\mathbb{E}\left[\left\|\widehat{f}_{\widehat{\lambda}} - f\right\|^2\right] \leq C \inf_{\lambda}\left\{\mathbb{E}\left[\left\|\widehat{f}_{\lambda} - f\right\|^2\right] + R\right\} \ .$$

Cela certifie que la calibration de λ est bien faite : on ne peut trouver un paramètre dont l'estimateur associé est bien meilleur que celui que nous obtenons.

Comment peut-on s'assurer que ces conditions seront bien vérifiées ? Pour cela, on utilisera des outils de concentration de la mesure, que l'on peut par exemple trouver dans le livre de Massart [Mas07]. La concentration de la mesure est un phénomène qui apparaît quand des quantités aléatoires se concentrent, avec grande probabilité, autour de leur espérance. Un tel phénomène apparaît dans nos problèmes et permet directement de certifier que la première hypothèse est vérifiée. Il suffit alors de prescrire les contraintes en espérance (c'est ce que l'on fait en choisissant $\widehat{\mathrm{pen}}(\lambda)$) afin de montrer que les autres hypothèses sont réalisées [22]. On remarquera qu'une des difficultés de ces résultats est que nos contraintes 1., 2. et 3. doivent être vraies simultanément pour tous les paramètres λ, ce qui ajoute quelques complications.

Enfin, nous pouvons remarquer une chose : nos résultats sont toujours énoncés avec n étant fixé, et non comme en considérant une limite quand n tend vers $+\infty$. Le réflexe classique du statisticien théoricien est de considérer que la taille de l'échantillon, n, est grande. Cela a souvent été traduit, au début, en considérant la limite des quantités considérées quand n tend vers l'infini : c'est le cadre asymptotique. Cela permet d'utiliser des théorèmes de convergence, comme le théorème central limite, et d'obtenir simplement bon nombre de résultats importants. Mais cette approche a de nombreux défauts : elle ne donne pas, ou peu, d'indications sur les cas pratiques et suppose que les autres quantités ne dépendent pas de n. C'est tout le contraire de ce que l'on voudrait faire ! On peut, au contraire, exprimer les résultats en fixant n, quitte à considérer que n est plus grand qu'une certaine constante [23]. Cela complique passablement les résultats, mais leur donne aussi une finesse et une précision qui sont indispensables. Tous nos résultats seront écrits dans ce cadre-là, que l'on nomme cadre non-asymptotique.

1.1.4 Où l'on voit poindre le multi-tâches

Dans les parties précédentes, nous avons vu comment construire des modèles plus intéressants que celui de la régression par moindres carrés et comment calibrer le paramètre de régularisation de ces modèles. On peut alors se demander s'il existe des estimateurs qui sont meilleurs que cela. On peut aussi se poser la question suivante : pour un problème donné, quelle est la meilleure qualité d'un estimateur ? Bien sûr, une fois le modèle spécifié par

$$\forall i \in \{1, \dots, n\},\ Y_i = f(X_i) + \varepsilon_i \ .$$

il existe un choix d'estimateur parfait : c'est f. Mais ce choix n'est bon que pour cet exemple-là, pas pour les autres ! On se placera donc dans un cadre pessimiste, et l'on considèrera la pire situation possible pour un estimateur. En notant $\ell(\widehat{f}, f)$ la perte qui nous intéresse (par

22. Nous ne cachons pas que l'on puisse rencontrer quelques « petits » calculs en chemin.

23. Il nous semble que, en statistique, l'on finisse toujours par supposer que n est grand. Un de nos théorèmes demande, par exemple, que $2\ln(n) \geq 1027$!

exemple, la perte de régression par *design* fixe ou celle de régression par *design* aléatoire), on étudie alors le pire risque qui puisse intervenir, c'est-à-dire, en considérant un ensemble $\mathcal{F}$ n'étant pas trop riche,

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\ell(\widehat{f},f)\right]\right\} \quad .$$

Un « bon » estimateur sera donc un estimateur qui minimisera cette quantité, c'est-à-dire dont le risque sera proche de

$$\inf_{\widehat{f}}\left\{\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\ell(\widehat{f},f)\right]\right\}\right\} \quad .$$

Cette dernière quantité s'appelle le « risque minimax [24] ». Un estimateur dont le risque s'approche du risque minimax aura cette qualité qu'il ne pourra pas être amélioré uniformément sur $\mathcal{F}$, en tout cas pas de beaucoup. L'analyse de ces risques minimax a fructifié lors de ces dernières années, et l'on trouvera bon nombre d'articles détaillant tel ou tel risque dans un contexte précis. On pourra citer comme références conséquentes l'article de Johnstone [Joh94] et les livres de Massart [Mas07] et de Tsybakov [Tsy08]. Ces risques minimax sont donc bien connus, dans les cas qui nous intéressent. En régression linéaire gaussienne de dimension k (c'est le cadre de la régression des moindres carrés), le risque minimax est $\frac{k}{n}$., que l'estimateur des moindres carrés n'atteint pas. En régression non linéaire, si l'on suppose que f se trouve dans une boule centrée en 0 d'un espace de Sobolev de régularité α, alors le risque minimax sera typiquement de la forme $n^{-2a/(2\alpha+1)}$. Sous certaines hypothèses de régularité du noyau utilisé, un estimateur *ridge* à noyau, avec une bonne sélection de paramètre, peut avoir un risque proche de ce risque minimax.

Il y a donc des limitations, que l'on ne peut dépasser en utilisant une approche statistique classique. Que peut-on alors faire si la précision des estimateurs que l'on utilise n'est pas suffisante, et que l'on ne peut pas augmenter la taille de l'échantillon ? Le statisticien peut alors avoir accès, simultanément, à des problèmes connexes, qui sont reliés à son problème d'estimation initial. Il dispose donc, en fait, d'une source d'informations sur ce problème, vu que ces problèmes annexes lui sont reliés. L'exploitation de cette similarité entre plusieurs problèmes a donné lieu à ce que nous appelons les méthodes « multi-tâches », qui traitent donc simultanément plusieurs problèmes statistiques, que l'on opposera aux méthodes « mono-tâche », que constituent en fait l'ensemble des méthodes classiques existantes.

Pour expliciter cela, nous pouvons reprendre l'exemple [25] de la détection de voitures dans des images, que nous avions développé dans la partie 1.1.1. Rappelons de quoi il était question. Notre scientifique, ici un spécialiste de vision artificielle, a construit un estimateur qui, à partir d'images, détecte des voitures dans ces dernières. Il est confiant dans l'excellence de son algorithme et a peu d'espoirs de pouvoir améliorer sa méthode de ce côté [26]. Les performances de l'estimateur ne conviennent cependant pas à notre scientifique. Celui-ci ne peut

24. Minimax, car on a ici un inf suivi d'un sup, que de nombreux auteurs écrivent, avec un abus de notation, $\min_{\widehat{f}}\max_{f\in\mathcal{F}}$.

25. Rappelons-le, cet exemple est totalement fictif. On pourra cependant trouver de nombreuses applications similaires dans la littérature, comme par exemple dans l'article de Lim et al. [LST11].

26. Il s'agit, bien entendu, une situation imaginaire. Le domaine de la vision artificielle, dans lequel, il est vrai, nous n'avons que peu de compétences, évolue très rapidement. Dans ce cas-là, il serait abusif de dire d'un estimateur qu'il est, en pratique, inaméliorable.

malheureusement pas obtenir plus d'images, car il faut, en plus de les obtenir, payer une personne qui devra annoter, à la main, l'échantillon, en disant pour chaque nouvelle image si celle-ci contient ou non une voiture. Or, heureuse coïncidence, ce scientifique travaille simultanément sur un projet semblable : il s'agit de reconnaître des minibus dans des images. Il serait donc tentant d'essayer de mêler les deux approches : apprendre à reconnaître une voiture devrait pouvoir aider à reconnaître un minibus, et vice-versa. De nombreuses questions se posent alors : comment modéliser cette ressemblance ? peut-on toujours exploiter une ressemblance ? est-ce utile ? Nous discuterons plus loin de tout cela.

1.2 Petit historique du multi-tâches

Le but de cette thèse est d'analyser une méthode multi-tâches bien particulière. Nous dressons un portrait général de ces méthodes dans ce chapitre, avant de détailler ensuite les contributions de cette thèse à notre domaine, la régression multi-tâches.

1.2.1 Le paradoxe de Stein

Le premier résultat que l'on peut inclure dans le domaine du multi-tâches est peut-être le paradoxe de Stein. C'est un exemple bien connu en statistique, nous rappelons brièvement ici de quoi il retourne. Ce paradoxe fut trouvé à la fin des années 50, à un moment ou l'on pensait que l'estimateur du maximum de vraisemblance était « optimal ». Voici le modèle : on suppose que l'on observe un vecteur

$$Y \sim \mathcal{N}(\theta, \sigma^2 I_p),\ \theta \in \mathbb{R}^p\ .$$

On veut estimer[27] θ, en minimisant le risque quadratique $\mathbb{E}\left[\left\|\widehat{\theta} - \theta\right\|^2\right]$. Dans ce modèle, l'estimateur du maximum de vraisemblance est $\widehat{\theta}_{\mathrm{MV}} = Y$. Peut-on améliorer cet estimateur ? La réponse, donnée par Stein [Ste56] dans un célèbre article, est oui. Il montra l'existence d'un estimateur dont le risque est inférieur, pour tous les paramètres, à celui de $\widehat{\theta}_{\mathrm{MV}}$, si $p \geq 3$. James et Stein [JS61] précisèrent ensuite cet estimateur, qui porte leur nom, dont voici la forme :

$$\widehat{\theta}_{\mathrm{JS}} = \left(1 - \frac{(p-2)\sigma^2}{Y^\top Y}\right) Y\ .$$

Tâchons de comprendre comment fonctionne cet estimateur : il corrige l'estimateur du maximum de vraisemblance en le déplaçant[28] vers 0. Si ce dernier est très loin de 0 (en supposant que σ^2 est fixe), le facteur de correction est proche de 1 : le déplacement est imperceptible. Mais, dans le cas contraire, le facteur de correction sera très petit. Nous pouvons faire ici deux remarques.

– On peut déplacer l'estimateur vers n'importe quel point ordinaire, et non pas uniquement vers 0, ainsi que vers la moyenne $\overline{Y}$.

27. On estime ici p quantités différentes. Ces quantités peuvent être, par exemple, des moyennes empiriques.
28. Nous traduisons ainsi l'anglais *shrink*

– On peut s'assurer que les coordonnées ne changent pas de signe lors de ce déplacement, en considérant la partie positive de l'estimateur de James-Stein,

$$\widehat{\theta}_{\mathrm{JS}+} = \left(1 - \frac{(p-2)\sigma^2}{Y^\top Y}\right)_+ Y \ ,$$

dont le risque est inférieur à celui de l'estimateur de James-Stein pour tous les paramètres. C'est d'ailleurs ce dernier estimateur qui sera le plus utilisé.

Les premières preuves données par James et Stein [JS61] sont très techniques. Stein [Ste81] en donna une preuve plus générale, qui mena à l'élaboration de la méthode SURE [29]. On peut enfin trouver une approche bayésienne empirique [30] dans plusieurs articles, notamment celui de Efron et Morris [EM73].

On peut remarquer que $\widehat{\theta}_{\mathrm{MV}}$ et $\widehat{\theta}_{\mathrm{JS}}$ ont des risques maximums égaux. En effet, tandis que le risque de l'estimateur du maximum de vraisemblance est constant et égal à $p\sigma^2$, celui de l'estimateur de James-Stein varie en fonction de $\|\theta\|$. Il converge vers $p\sigma^2$ lorsque $\|\theta\|$ tend vers $+\infty$, mais vaut $2\sigma^2$ lorsque $\theta = 0$. Ainsi, d'un point de vue minimax, ces deux estimateurs sont équivalents. Mais l'estimateur de James-Stein conduit à un gain important quand θ est petit, c'est-à-dire lorsque l'hypothèse faite par le statisticien — θ est petit — est valide.

On remarquera aussi que l'estimateur de James-Stein est fortement lié à l'estimateur de régression *ridge*. Les deux reviennent en effet à déplacer les coordonnées de l'estimateur initial, qui est dans les deux cas l'estimateur du maximum de vraisemblance, vers 0. L'estimateur de James-Stein peut donc être vu comme un cas particulier de régression *ridge* en *design* aléatoire, avec l'avantage de fournir une sélection du paramètre de régularisation — via $(p-2)\sigma^2$ — arbitraire et efficace. On pourra par exemple lire le rapport de Draper et Van Nostrand [DVN79] à ce sujet.

On peut interprêter l'estimateur de James-Stein comme étant un estimateur multi-tâches dans un cadre bien particulier : l'estimation de moyennes. Pour chaque tâche, on doit estimer l'espérance d'une distribution. On calcule alors la moyenne empirique des échantillons pour chaque tâche. Si l'on peut faire quelques hypothèses sur les variances de ces moyennes empiriques, on peut alors considérer l'estimateur de James-Stein (ou plutôt, sa partie positive), en mettant dans chaque coordonnée la moyenne obtenue pour une tâche. On rapprochera alors les moyennes les unes des autres et l'on gagnera beaucoup si toutes ces espérances se révèlent être proches.

L'estimateur de James-Stein a donc toutes les propriétés que l'on attend d'un estimateur multi-tâches. Détaillons ces propriétés. Son utilisation demande d'abord à ce que l'on s'intéresse à la somme des erreurs quadratiques sur les différentes tâches, et non à une erreur en particulier. De plus, on obtient une garantie d'efficacité globale, et non tâche par tâche. L'estimateur multi-tâches ainsi créé pourra être moins bon pour une tâche particulière que l'estimateur mono-tâche associé mais, s'il est suffisamment meilleur sur les autres tâches, nous en serons satisfaits. Enfin, l'estimateur multi-tâches apporte une amélioration significative quand l'hypothèse qu'a faite le statisticien est correcte (ici, vers où diriger les coordonnées). Dans le cas contraire, il risque de n'y avoir que peu à gagner, et l'estimateur multi-tâches pourrait même être moins bon que l'estimateur mono-tâche.

29. SURE : Stein's Unbiased Risk Estimation.
30. De l'anglais *Empirical Bayes*

1.2.2 Quelques modèles multi-tâches

Nous avons déjà présenté l'estimateur de James-Stein, qui peut être vu comme un estimateur de moyennes multi-tâches. D'autres modèles ont ensuite été développés.

Régression *ridge* multivariée

On peut ensuite citer la méthode de régression *ridge* multivariée de Brown et Zidek [BZ80]. Ici, le mot multivarié a une signification proche de celle de multi-tâches : on observe plusieurs processus, mais on ne suppose pas nécessairement qu'ils se ressemblent. Nous verrons que ces tâches partagent quand même un peu d'informations. Précisons ce modèle : on observe

$$Y = X\beta + \varepsilon \ ,$$

où Y est une matrice de taille $n \times p$, X une matrice de taille $n \times k$, β une matrice de taille $k \times p$ et ε une matrice de taille $n \times p$. On suppose que l'on observe Y et X et que $\mathbb{E}[\varepsilon] = 0$. Si on note A^j la colonne j d'une matrice A, on a ici p modèles de régression, du type $Y^j = X\beta^j + \varepsilon^j$. Ces modèles partagent deux choses :

- ils ont les mêmes covariables, données par X ;
- leurs bruits ne sont pas nécessairement indépendants, car on demande que la condition suivante soit vraie : $\forall (i,j) \in \{1,\dots,p\}^2, \exists \gamma_{i,j},\ \mathrm{Cov}\,(\varepsilon^i, \varepsilon^j) = \gamma_{i,j} I_n$.

On construit alors un estimateur *ridge* en *design* aléatoire, dépendant d'une matrice de régularisation M, comme suit (on note par $\otimes$ le produit de Kronecker) :

$$\widehat{\beta}(M) = (X^\top X \otimes I_p + I_k \otimes M)^{-1}(X^\top X \otimes I_q)Y \ .$$

La dépendance de cet estimateur M permet alors de s'adapter au fait que ces p tâches ne sont pas indépendantes, et des méthodes de choix de M adaptées à des cas bien précis sont détaillées dans l'article cité. Cet exemple est important à nos yeux, car il ouvre la voie à une analyse multi-tâches de la régression *ridge*. Les estimateurs développés plus tard en seront proches, mais devront en plus s'adapter à une hypothèse supplémentaire : les différentes fonctions de régression sont censées se « ressembler ».

Régressions multi-tâches

Nous arrivons, enfin, au sujet qui nous intéresse ici : les modèles de régression multi-tâches. On commence donc par supposer un modèle de régression multi-tâches. On fixe un ensemble $\mathcal{X}$ (pour simplifier, ce sera $\mathbb{R}^d$), ainsi que p fonctions $f^1, \dots, f^p$, à variables dans $\mathcal{X}$ et à valeurs dans $\mathbb{R}$. On tire ensuite p échantillons [31] $((X_i^1, Y_i^1)_{i=1}^n, \dots, (X_i^p, Y_i^p)_{i=1}^n)$ et l'on suppose qu'il existe des variables aléatoires centrées ε_i^j telles que

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\},\ Y_i^j = f^j(X_i^j) + \varepsilon_i^j \ .$$

On suppose aussi, bien entendu, que, si $i \neq i'$, les variables ε_i^j et $\varepsilon_{i'}^{j'}$ sont indépendantes, et ce pour tout couple (j, j') [32]. Nous ne demandons maintenant qu'à exprimer une idée :

31. Pour chaque échantillon, la loi diffère d'observation en observation.

32. On se tiendra autant que possible à la convention suivante : les indices indiquent la position dans l'échantillon, dans $\{1,\dots,n\}$, les exposants indiquent la tâche, dans $\{1,\dots,p\}$.

les fonctions $f^1, \dots, f^p$ se ressemblent. Mais que cela signifie-t-il ? On peut trouver, dans la littérature, deux types de catégories dans lesquelles on peut classer ces modèles de régression.

Faible dimension : On peut supposer que toutes les fonctions sont linéaires, de la forme $f^j(x) = x^\top \beta^j$, et que tous les vecteurs $(\beta^1, \dots, \beta^p)$ appartiennent à un même sous-espace vectoriel, de préférence de faible dimension. Pour résumer : les p fonctions peuvent être décrites par un petit nombre de descripteurs.

Similarité euclidienne : On suppose que $\mathcal{X}$ est muni d'une structure hilbertienne (le plus souvent, ce sera un RKHS), et on suppose alors que toutes les fonctions $f^1, \dots, f^p$) se trouvent dans une boule de $\mathcal{X}$, de préférence de petit rayon.

On le voit, ce sont des hypothèses très différentes, qui traduisent des conceptions assez éloignées du concept de « similarité ». On utilisera pourtant, dans les deux cas, des méthodes de pénalisation du risque empirique.

Similarité euclidienne

Nous touchons ici au cœur de notre sujet, car c'est dans ce cadre que nous placerons notre étude. Ce dernier a été moins étudié que le cadre de la faible dimension, que nous venons de décrire. La principale référence est un article de Evgeniou et al. [EMP05], sur lequel nous nous fonderons pour introduire le modèle. Nous décrivons maintenant le cadre de cet article. On considère que l'on observe des couples (Y_i^j, X_i^j), définis par

$$Y = X^\top \beta + \varepsilon \ .$$

On étend alors, et nous l'expliquerons, le critère *ridge* à ceci :

$$\frac{1}{np} \|Y - X\beta\|^2 + \sum_{j,k} M_{j,k} (\beta^j)^\top \beta^k \ .$$

Le second terme est, comme précédemment, un terme de régularisation, la matrice M étant alors l'analogue du paramètre de régularisation. Un avantage de cette formulation est qu'il est aisé de voir qu'il prolonge la régression *ridge* en conservant certains de ses avantages, notamment la facilité de calcul des estimateurs recontrés. Mais quelle régularisation effectue-t-on alors ? Cela dépend de la matrice M. Une matrice M diagonale fera que ce critère se découplera en p problèmes indépendants, ce qui revient à considérer les régressions mono-tâches. Avec $M = (\lambda + p\mu) I_p - \mu \mathbf{1}\mathbf{1}^\top$, on pourra aussi créer un terme de régularisation de la forme $\lambda \sum_j \left\|\beta^j\right\|^2 + \mu \sum_{j,k} \left\|\beta^j - \beta^k\right\|^2$, forçant ainsi les p estimateurs de tâches à être proches. Le cadre que nous avons cité est donc très souple, trop peut-être, et permet d'exprimer plusieurs hypothèses sur la répartition des tâches.

Faible dimension, parcimonie et norme nucléaire

Les méthodes de régression multi-tâches utilisant cette hypothèse de faible dimension ont été assez largement étudiées. Peut-être est-ce dû à la popularité des méthodes de régression linéaire mono-tâche qui utilisent des hypothèses de faible dimension, comme le Lasso par exemple. Commençons par une situation simple de parcimonie : le cas où il existe un ensemble

$B^\star \subset \{1, \dots, p\}$ tel que, pour chaque tâche j, le support de β^j est inclus dans $B^\star$. Le modèle est ici simplifié, car on sait alors que le sous-espace vectoriel sur lequel on recherche nos estimateurs est bien particulier. Ce problème est alors souvent traité en pénalisant le risque empirique de cette manière :

$$\frac{1}{np} \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{i=1}^{n} (\beta_i^j)^2} \ .$$

On appelle cela le lasso groupé et les vecteurs de paramètres ainsi obtenus auront effectivement tendance à avoir un support, commun, de petite taille, grâce notamment aux propriétés du Lasso (qui correspond à la régularisation L_1). L'estimateur qui minimise cette quantité est aussi aisément calculable. Ce cas particulier fut par exemple étudié par Obozinski et al. [OWJ11] ainsi que par Lounici et al. [LPTvdG09]. On notera l'application du Lasso groupé à des problèmes de détection de rupture par Bleakley et Vert [BV11] Plusieurs extensions sont possibles, par exemple en considérant plusieurs groupes de variables, où chaque groupe doit avoir un support restreint [LPTvdG11], en étendant cela à la régression à noyau [KY08] ou bien à l'apprentissage par noyaux multiples [KY10].

Traiter le cas où l'on suppose juste que les descripteurs des différentes tâches appartiennent à un sous-espace de petite dimension est plus délicat. Les pénalisations qui peuvent être alors utilisées, comme on peut le voir chez Argyriou et al. [AEP08], ne mènent alors pas à des estimateurs que l'on peut facilement calculer. On s'en sortira souvent en considérant le problème de minimisation suivant

$$\frac{1}{np} \|Y - X\beta\|^2 + \lambda \|\beta\|_\star \ ,$$

où $\|\beta\|_\star = \operatorname{tr} \sqrt{\beta^\top \beta}$ est la norme nucléaire de β, c'est-à-dire la somme de ses valeurs singulières. On peut aussi citer l'article de Jacob et al. [JBV08], mêlant cette approche à un problème de *clustering*. Rohde et Tsybakov [RT11] ont analysé ce cas-là et on montré qu'il menait à de bonnes performances de l'estimateur multi-tâches. On pourra aussi remarquer le travail de Giraud [Gir11], qui mène à une pénalisation légèrement différente.

Enfin, même si ces estimateurs peuvent sembler séduisants, la calibration du (ou des) paramètres de régularisation peut être problématique. Il n'existe pas aujourd'hui, à notre connaissance, de méthode complètement adaptative et que l'on sache analyser qui permette une telle sélection, sauf peut-être, d'un point de vue théorique, la validation simple avec un échantillon de test de taille $n/\ln n$.

Des validations expérimentales

Nous n'entrerons pas ici dans les détails, mais nous nous bornerons juste à dire que des méthodes multi-tâches ont été expérimentées dans de nombreux cadres. On citera principalement le travail de thèse de Caruana [Car97], qui fait la part belle aux réseaux de neurones et teste ses méthodes sur de nombreux jeux de données, réels ou artificiels. On pourra citer aussi quelques applications en robotique, par exemple, comme dans l'article de Thrun et O'Sullivan [TO96]. Enfin, on pourra aussi trouver de nombreuses occurrences de l'expression *transfer learning*, notamment en vision artificielle. Il s'agit alors d'utiliser dans une tâche

une partie de l'échantillon d'une autre tâche, quitte à sélectionner ou déformer cette partie de l'échantillon empruntée pour la rendre compatible à la tâche étudiée.

Des modèles multi-tâches éloignés de notre sujet

Plusieurs modèles ont été développés pour étudier théoriquement le multi-tâches, pour lesquels nous n'entrerons pas, là non plus, dans les détails. Baxter [Bax00] a développé un cadre d'étude général, concernant l'apprentissage d'hypothèses par minimisation du risque empirique. Il montre des bornes concernant l'apprentissage de plusieurs tâches tirées aléatoirement selon un processus commun. Dans un cadre de classification, Ben-David et Schuller [BDS03] définissent ce qu'est une similarité entre deux tâches bien précisément. Soit $\mathcal{X}$ un ensemble, $\mathcal{P}_1$ et $\mathcal{P}_2$ deux mesures de probabilité (que l'on doit apprendre) sur $\mathcal{X} \times \{0,1\}$ et $\mathcal{F}$ un groupe de permutations de $\mathcal{X}$. Alors $\mathcal{P}_1$ et $\mathcal{P}_2$ sont dites $\mathcal{F}$-semblables si, pour toute partie T de $\mathcal{X} \times \{0,1\}$ $\mathcal{P}_1$ mesurable, $f(T)$ est $\mathcal{P}_2$ mesurable et $\mathcal{P}_1(T) = \mathcal{P}_2(f(T))$. Les auteurs de cet article montrent alors que l'apprentissage de classifieurs de plusieurs distributions $\mathcal{F}$-semblables est possible, en utilisant les outils introduits par l'article précédent. C'est une approche intéressante, car elle modélise précisément ce qu'est la similarité entre les tâches, même si cette modélisation semble très contraignante et difficilement utilisable en pratique.

1.2.3 Quelles questions se pose-t-on ici ?

Dans cette thèse, nous étudierons le modèle de régression multi-tâches, en nous plaçant dans le cadre de similarité euclidienne et en utilisant des outils de régression à noyau. Nous nous tenterons alors de répondre à plusieurs interrogations :

1. Quels types de similarités peut-on exprimer avec notre modèle ?
2. Comment ces similarités s'expriment-elles dans notre modèle ?
3. Peut-on calibrer les estimateurs que l'on obtient ?
4. Le cas échéant, l'estimateur ainsi calibré a-t-il de bonnes qualités ? Notamment, vérifie-t-il une inégalité oracle ?
5. L'estimateur multi-tâches ainsi obtenu est-il plus efficace que l'estimateur mono-tâche [33] ?
6. Y a-t-il des situations intrinsèquement favorables, ou défavorables, à une estimation multi-tâches ?

1.3 Contributions de la thèse

Nous expliquons dans cette partie les réponses que cette thèse apporte à ces questions.

1.3.1 Cadre et modèle

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. On suppose que l'on observe l'échantillon $\mathcal{D}_n = (X_i, Y_i^1, \ldots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$. Pour chaque tâche $j \in \{1, \ldots, p\}$, $\mathcal{D}_n^j = (X_i, Y_i^j)_{i=1}^n$ est un n échantillon de loi $\mathcal{P}^j$, dont la première loi marginale est $\mathcal{P}$. On cherche à résoudre un

33. Autrement dit, notre travail sert-il à quelque chose ?

problème de régression pour chaque tâche. Détaillons maintenant le modèle. Nous supposons d'abord qu'il existe $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$ et des vecteurs $(\varepsilon_i^j)_{j=1}^p$ indépendants et de même loi normale $\mathcal{N}(0,\Sigma)$. On suppose aussi que pour tout $j \in \{1,\dots,p\}$, il existe $F^j \in L^2(\mathbb{P})$ tel que

$$\forall i \in \{1,\dots,n\}, \forall j \in \{1,\dots,p\},\ Y_i^j = F^j(X_i) + \varepsilon_i^j \ .$$

Remarque 1.1. *On suppose donc, ici, que toutes les tâches ont le même design. Cela facilite notamment l'étude théorique, qui peut se faire sans cette hypothèse, au prix de quelques suppositions supplémentaires.*

Remarque 1.2. *La matrice Σ est la matrice de covariance du bruit entre les tâches, qui ne sont donc pas nécessairement indépendantes conditionellement à $(X_i)_{i=1}^n$. L'estimation de cette matrice se révèlera être très importante.*

On se place maintenant dans un cadre de régression en *design* fixe, le risque qui nous intéresse est donc, pour un estimateur $\widehat{F}$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (F(X_i) - \widehat{F}(X_i))^2 \middle| X_1,\dots,X_n\right] \ .$$

Toutes les espérances qui suivent sont implicitement prises conditionnellement à $(X_1,\dots,X_n)$, afin de garder des notations concises. Nous noterons aussi

$$f = \mathrm{vec}\left((F^j(X_i))_{i,j}\right),\ f^j = \mathrm{vec}\left((F^j(X_i))_{i=1}^n\right) \quad \text{et} \quad y = \mathrm{vec}\left((Y_i^j)_{i,j}\right) \ ,$$

et prenons des notations similaires pour les estimateurs. Avec de telles notations, les éléments sont regroupés tâche par tâche dans des vecteurs, en commençant par ceux liés à la première tâche pour finir par ceux liés à la dernière. Nous travaillons donc avec une perte quadratique, notée $\left\|\widehat{f} - f\right\|^2$, et le risque quadratique associé, $\mathbb{E}\left[\left\|\widehat{f} - f\right\|^2\right]$.

Nous nous plaçons ensuite dans un cadre de régression *ridge* à noyau. On se donne donc un RKHS $\mathcal{F} \subset L^2(\mathbb{P})$, dont le noyau associé est k et la fonction de description est $\Phi : \mathcal{X} \to \mathcal{F}$. Cela nous donne la matrice de noyau $K = (k(X_i, X_\ell))_{1\le i,\ell\le n} \in \mathcal{S}_n^+(\mathbb{R})$. Nous cherchons donc à construire des estimateurs dans $\mathcal{F}$. Pour cela, dans la droite ligne des travaux de Brown et Zidek [BZ80] ainsi que de Evgeniou et al. [EMP05], nous considérons l'estimateur solution du problème de minimisation, dépendant d'un paramètre de régularisation $M \in \mathcal{S}_p^+(\mathbb{R})$,

$$\widehat{F}_M \in \operatorname*{argmin}_{G\in\mathcal{F}^p} \left\{ \underbrace{\frac{1}{np}\sum_{i=1}^n\sum_{j=1}^p (Y_i^j - G^j(X_i))^2}_{\text{Risque empirique}} + \underbrace{\sum_{j=1}^p\sum_{\ell=1}^p M_{j,\ell}\langle G^j, G^\ell\rangle_{\mathcal{F}}}_{\text{Terme de régularisation}} \right\} \ .$$

On peut alors construire un RKHS dépendant de M qui permet d'utiliser le théorème du représentant et d'obtenir l'estimateur à *design* fixe

$$\widehat{f}_M = A_M y = \widetilde{K}_M(\widetilde{K}_M + npI_{np})^{-1}y = (M^{-1}\otimes K)\left((M^{-1}\otimes K) + npI_{np}\right)^{-1} y \ .$$

La matrice M permet de représenter la similarité entre les tâches. Calibrer cette matrice devrait donc permettre de s'adapter, au moins en partie, à cette similarité. Finalement, voici deux exemples de type de matrice M que nous utiliserons souvent par la suite.

Exemple 1.1. *Si l'on veut traiter les p tâches séparément, et donc obtenir les estimateurs mono-tâche, on peut alors prendre* $M = M_{\text{ind}}(\lambda) := \frac{1}{p}\operatorname{Diag}(\lambda_1, \dots, \lambda_p)$ *pour tout* $\lambda \in \mathbb{R}^p$. *Cela mène à la régularisation*

$$\frac{1}{p}\sum_{j=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^j - G^j(X_i))^2 + \lambda_j \|G^j\|_{\mathcal{F}}^2\right] \quad ,$$

qui se découple bien.

Exemple 1.2. *On peut suivre Evgeniou et al. [EMP05] et définir, pour tout* $(\lambda, \mu) \in (0, +\infty)^2$,

$$M_{\text{SD}}(\lambda, \mu) := (\lambda + p\mu)I_p - \mu \mathbf{1}\mathbf{1}^\top = \begin{pmatrix} \lambda + (p-1)\mu & & -\mu \\ & \ddots & \\ -\mu & & \lambda + (p-1)\mu \end{pmatrix} \quad .$$

Avec $M = M_{\text{SD}}(\lambda, \mu)$, *on obtient la régularisation*

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_i^j - G^j(X_i))^2 + \lambda \sum_{j=1}^{p}\left\|G^j\right\|_{\mathcal{F}}^2 + \frac{\mu}{2}\sum_{j=1}^{p}\sum_{k=1}^{p}\left\|G^j - G^k\right\|_{\mathcal{F}}^2 \quad .$$

Cela permet donc de régulariser à la fois les normes des fonctions G^j *et celles de leurs différences,* $G^j - G^k$. *Ainsi, les matrices* $M_{\text{SD}}(\lambda, \mu)$ *peuvent être utilisées lorsqu'on suppose que les fonctions* F^j *sont proches dans* $\mathcal{F}$.

1.3.2 Calibration d'un estimateur multi-tâches

Nous cherchons d'abord à sélectionner une matrice M, à partir d'un ensemble $\mathcal{M}$, afin que l'estimateur associé $\widehat{f}_M$ ait un faible risque. Pour cela, nous allons mettre en œuvre la méthode de pénalisation du risque empirique par la pénalité idéale, que nous avions décrite auparavant.

Pénalisation idéale du risque empirique

On cherche donc une pénalité dépendant uniquement des données et qui approche au mieux

$$\operatorname{pen}_{\text{id}}(M) := \mathbb{E}\left[\frac{1}{np}\|\widehat{f}_M - f\|_2^2\right] - \mathbb{E}\left[\frac{1}{np}\left\|y - \widehat{f}_M\right\|_2^2\right] \quad .$$

Un simple calcul montre que cela vaut, nonobstant un terme ne dépendant pas de M,

$$\operatorname{pen}_{\text{id}}(M) = \frac{2\operatorname{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} \quad .$$

Or, cela dépend de Σ, que l'on ne connaît pas. La première étape de ce travail est donc d'estimer Σ et de montrer que cette estimation est suffisamment précise pour que la pénalité

$$\widehat{\operatorname{pen}}(M) = \frac{2\operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n)\right)}{np} \tag{1.9}$$

approche suffisamment bien la pénalité idéale.

Estimation de Σ

Notre estimateur de la matrice de covariance Σ est fondé sur le concept de pénalité minimale. Nous ne discuterons pas en détails ici de ces pénalités, le lecteur intéressé pourra lire l'article de Arlot et Bach [AB11] ou bien consulter le court résumé se trouvant partie 3.3, page 53. Nous nous bornerons à dire que cela permet, dans notre cadre, d'estimer la variance du bruit dans un problème de régression mono-tâche.

Notre stratégie d'estimation est donc la suivante :

1. Sélectionner un ensemble de directions dans $\mathbb{R}^p$, où chaque coordonnée représente une tâche.
2. Pour chaque direction, considérer le problème mono-tâche correspondant à projection multi-tâches selon la direction choisie et estimer la variance du bruit dans cette direction.
3. Construire $\widehat{\Sigma}$ à partir de ces estimations uni-dimensionnelles.

Nous formulons cela de manière plus précise. Pour tout $z \in \mathbb{R}^p$ on considère le problème de régression mono-tâche

$$Y_z := Y \cdot z = F \cdot z + E \cdot z = F_z + \varepsilon_z \ . \tag{P_z}$$

On peut alors noter par $a(z)$ l'estimateur de la variance du problème (P_z) et $(e_1, \dots, e_p)$ la base canonique de $\mathbb{R}^p$. On voit alors que $a(e_i)$ estime $\Sigma_{i,i}$ et que $a(e_i + e_j)$ estime $\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}$. Ainsi, $\Sigma_{i,j}$ peut être estimé par $(a(e_i + e_j) - a(e_i) - a(e_j))/2$.

On introduit donc la fonction $J : \mathbb{R}^{p(p+1)/2} \mapsto \mathcal{S}_p$, que l'on définit par

$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,i} = a_i \text{ si } 1 \leq i \leq p \ ,$$
$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,j} = \frac{a_{i,j} - a_i - a_j}{2} \text{ si } 1 \leq i < j \leq p \ .$$

On peut donc voir que

$$\Sigma = J(\Sigma_{1,1}, \dots, \Sigma_{p,p}, \Sigma_{1,1} + \Sigma_{2,2} + 2\Sigma_{1,2}, \dots)$$

et l'on pose

$$\widehat{\Sigma} := J\left(a(e_1), \dots, a(e_p), a(e_1 + e_2), \dots, a(e_1 + e_p), \dots, a(e_{p-1} + e_p)\right) \ . \tag{1.10}$$

On peut alors montrer le résultat suivant, en notant $c(\Sigma)$ le conditionnement de Σ, $\preceq$ la relation d'ordre définie par $A \preceq B$ si $B - A$ est symétrique positive et en introduisant une hypothèse sur le biais du modèle :

$$\left.\begin{array}{l} \forall j \in \{1, \dots, p\}\,, \exists \lambda_{0,j} \in (0, +\infty)\,, \\ \qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{et} \quad \dfrac{1}{n} \left\| (A_{\lambda_{0,j}} - I_n) F_{e_j} \right\|_2^2 \leq \Sigma_{j,j} \sqrt{\dfrac{\ln n}{n}} \end{array}\right\} \tag{\textbf{Hdf}}$$

Théorème 1.1. *Soit $\widehat{\Sigma}$ l'estimateur défini dans l'équation* (1.10) *et supposons que* (**Hdf**) *soit vérifiée. Pour tout $\delta \geq 2$, il existe une constante $n_0(\delta)$, une constante $L_1 > 0$ ainsi qu'un événement $\widetilde{\Omega}$, vérifiant $\mathbb{P}(\widetilde{\Omega}) \geq 1 - p(p+1)/2 \times n^{-\delta}$, tels que, si $n \geq n_0(\delta)$, sur $\widetilde{\Omega}$,*

$$(1 - \eta)\Sigma \preceq \widehat{\Sigma} \preceq (1 + \eta)\Sigma \tag{1.11}$$
$$\textit{avec} \qquad \eta := L_1(2 + \delta) p \sqrt{\frac{\ln(n)}{n}} c(\Sigma)^2 \ .$$

Notre estimateur de Σ converge donc bien vers Σ, avec une vitesse précisée ici.

Le cas le plus souvent étudié en estimation de matrice de covariance est le cas où f est constante ou nulle, et l'on cherche alors à améliorer la matrice de covariance empirique. Bickel et Levina [BL08] ou Cai et al. [CZZ10], par exemple, utilisent des méthodes de seuillage pour obtenir, dans le second cas, des taux de convergence minimax. D'autres supposent une hypothèse de parcimonie et utilisent ensuite des méthodes de seuillage, comme Karoui [Kar08], ou bien des méthodes de régularisation, comme chez Lam et Fan [LF09].

Notre cadre est assez éloigné de ces situations-là, car notre échantillon $(Y_i^1, \dots, Y_i^p)_{i=1}^n$ n'est pas centré, ni même homoscédastique. Former une matrice de covariance empirique n'a pas, ici, de sens.

Inégalité oracle

Nous pouvons maintenant garantir l'efficacité de l'estimateur $\widehat{f}_{\widehat{M}}$ calibré par la méthode de pénalisation (1.9), via une inégalité oracle. Notre résultat recouvre deux types de situation :

1. une situation où l'ensemble des matrices est discret, cela permet de recouvrir certaines situations où l'on a peu d'informations *a priori* sur la répartition des tâches ;
2. une situation où l'ensemble des matrices est codiagonalisable en base orthonormée, ce qui arrive dans plusieurs situations où l'on a beaucoup d'informations *a priori* sur la répartition des tâches, par exemple quand toutes les fonctions sont regroupées dans plusieurs clusters.

Dans la première situation, on suppose que l'on a un ensemble $\mathcal{M}$ vérifiant

$$\exists (C, \alpha_{\mathcal{M}}) \in (0, +\infty)^2, \quad \operatorname{card}(\mathcal{M}) < C n^{\alpha_{\mathcal{M}}} \ . \tag{1.12}$$

On peut alors définir le paramètre sélectionné par

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2 \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\} \ .$$

L'inégalité oracle recherchée s'énonce alors comme suit, en nommant par $\sigma_{\max}$ la plus grande valeur propre de Σ.

Théorème 1.2. *Soit $\alpha = \max(\alpha_{\mathcal{M}}, 2)$, $\delta \geq 2$ et supposons que les hypothèses* (**Hdf**) *et* (1.12) *sont vérifiées. Il existe alors des constantes $L_2, \kappa' > 0$, une constante $n_1(\delta)$ ainsi qu'un événement $\widetilde{\Omega}$, vérifiant $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa' p(p + C) n^{-\delta}$, tels que, si $n \geq n_1(\delta)$, sur $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 c(\Sigma)^4 \sigma_{\max} (\alpha + \delta)^2 \frac{p^4 \ln(n)^3}{np} \ .$$

Ce résultat est une version simplifiée du théorème 3.3, page 57. Le résultat original possède aussi une variante énoncée en espérance.

Dans la deuxième situation, on suppose que l'on a un ensemble $\mathcal{M}$ vérifiant

$$\exists P \in O_p(\mathbb{R}), \quad \mathcal{M} \subseteq \left\{ P^\top \operatorname{Diag}(d_1, \dots, d_p) P \, , \, (d_i)_{i=1}^p \in (0, +\infty)^p \right\} \ . \tag{HM}$$

En définissant $(u_i)_{i=1}^p$ par $\forall j \in \{1,\dots,p\},\ u_j = P^\top e_j$, on estime alors Σ par

$$\widehat{\Sigma}_{\mathrm{HM}} = P\,\mathrm{Diag}(a(u_1),\dots,a(u_p))P^\top \ ,$$

La paramètre sélectionné est alors

$$\widehat{M}_{\mathrm{HM}} \in \underset{M\in\mathcal{M}}{\mathrm{argmin}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2\,\mathrm{tr}\left(A_M \cdot (\widehat{\Sigma}_{\mathrm{HM}} \otimes I_n) \right) \right\} \ . \tag{1.13}$$

On peut maintenant énoncer l'inégalité oracle concernant ce cas-là

Théorème 1.3. *Soit $\alpha = 2$, $\delta \geq 2$ et supposons que les hypothèses* (**Hdf**) *et* (**HM**) *sont vérifiées. Il existe alors des constantes $L_2 > 0$, κ'', une constante $n_1(\delta)$ ainsi qu'un événement $\widetilde{\Omega}$, vérifiant $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa'' p n^{-\delta}$ tels, si $n \geq n_1(\delta)$, sur $\widetilde{\Omega}$,*

$$\frac{1}{np}\left\| \widehat{f}_{\widehat{M}_{HM}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \inf_{M\in\mathcal{M}} \left\{ \frac{1}{np}\left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 \sigma_{\max} (2+\delta)^2 \frac{p\ln(n)^3}{np} \ .$$

Ce résultat est une version simplifiée du théorème 3.4, page 58. Le résultat original possède aussi une variante énoncée en espérance.

Nous pouvons remarquer plusieurs choses.

- L'obtention d'informations supplémentaires sur la répartition des tâches permet d'obtenir un algorithme simplifié possédant une plus forte garantie de convergence.
- Pour que les inégalités oracles contraignent le risque de l'estimateur, c'est-à-dire que le terme de droite additif soit négligeable devant l'infimum, n et p doivent être contraints, ce qui exclut les cas du type $p \gg n$. Ces contraintes sont discutées dans la remarque 3.13, page 59.
- Des simulations viennent confirmer le bon comportement de notre estimateur multi-tâches dans des cas qui ne le contraignent pas via l'inégalité oracle (partie 3.6, page 60).

1.3.3 Le multi-tâche fonctionne-t-il ?

Nous disposons maintenant d'un estimateur multi-tâches pouvant s'adapter à une famille de paramètres $\mathcal{M}$, c'est-à-dire choisir un paramètre $\widehat{M} \in \mathcal{M}$ dont le risque est proche du meilleur risque possible sur $\mathcal{M}$. On peut alors se poser la question suivante : l'estimateur ainsi obtenu est-il plus performant que l'estimateur mono-tâche associé ? Au vu de ce que nous avons montré, il suffit d'étudier les risques des estimateurs oracles. Par souci de simplicité, nous supposons que $\Sigma = \sigma^2 I_p$.

Décomposition du risque

Nous allons étudier le cas où les p fonctions de régression appartenant à chaque tâche sont censées être proches. Nous savons qu'il existe des matrices permettant de régulariser ces fonctions ainsi que leurs différences, une rapide étude du risque permet cependant de se rendre compte qu'il est plus judicieux de régulariser la moyenne des fonctions ainsi que leur variance. On utilise donc l'ensemble de matrices

$$\mathcal{M}_{\mathrm{AV}} = \left\{ M_{\mathrm{AV}}(\lambda,\mu), (\lambda,\mu) \in \mathbb{R}^2 \right\} \ ,$$

avec

$$M_{\mathrm{AV}}(\lambda,\mu) := \frac{\lambda}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) \quad .$$

Cela mène au critère suivant :

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - G^j(X_i))^2 + \lambda\left\|\frac{\sum_{j=1}^p G^j}{p}\right\|_{\mathcal{F}}^2 + \mu\left[\frac{\sum_{j=1}^p \left\|G^j\right\|_{\mathcal{F}}^2}{p} - \left\|\frac{\sum_{j=1}^p G^j}{p}\right\|_{\mathcal{F}}^2\right] \quad .$$

Le risque oracle est donc

$$\mathfrak{R}^\star_{\mathrm{ST}} = \inf_{(\lambda^1,\ldots,\lambda^p)\in\mathbb{R}_+^p}\left\{\frac{1}{np}\mathbb{E}\left[\sum_{j=1}^{p}\left\|\widehat{f}^j_{\lambda^j} - f^j\right\|_2^2\right]\right\}$$

pour le mono-tâche et

$$\mathfrak{R}^\star_{\mathrm{MT}} = \inf_{(\lambda,\mu)\in\mathbb{R}_+^2}\left\{\frac{1}{np}\mathbb{E}\left[\left\|\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)} - f\right\|_2^2\right]\right\}$$

pour le multi-tâches.

On peut tout d'abord remarquer que le risque oracle mono-tâche $\mathfrak{R}^\star_{\mathrm{ST}}$ est un infimum sur p paramètres, alors que le risque oracle multi-tâches $\mathfrak{R}^\star_{\mathrm{MT}}$ est un infimum sur uniquement deux paramètres. Le mono-tâche possède donc plus de degrés de liberté que le multi-tâches, mais ne peut pas utiliser les données de tâches différentes simultanément. Il n'est donc pas évident de pouvoir obtenir la simple garantie « $\mathfrak{R}^\star_{\mathrm{MT}} \leq \mathfrak{R}^\star_{\mathrm{ST}}$ ».

Notons $(\gamma_i)_{i=1}^n$ les valeurs propres de K et, pour tout $j \in \{1,\ldots,p\}$, $(h_i^j)_{i=1}^n$ les coordonnées de f^j sur la base orthonormée qui diagonalise K. Notons aussi la moyenne de h_i^j

$$\mu_i = \nu_i^1 = \frac{h_i^1 + \cdots + h_i^p}{\sqrt{p}}$$

et la « variance » inter-tâches [34]

$$\varsigma_i^2 = \frac{\sum_{j=1}^p (h_i^j)^2}{p} - \left(\frac{\sum_{j=1}^p h_i^j}{p}\right)^2 = \frac{1}{p}\sum_{j=1}^{p}\left(h_i^j - \frac{\sum_{j=1}^p h_i^j}{p}\right)^2 \quad .$$

On peut alors écrire le risque de l'estimateur $\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)}$, grâce à une décomposition biais-variance, comme

$$n\lambda^2\sum_{i=1}^{n}\frac{\frac{\mu_i^2}{p}}{(\gamma_i+n\lambda)^2} + \frac{\sigma^2}{np}\sum_{i=1}^{n}\left(\frac{\gamma_i}{\gamma_i+n\lambda}\right)^2 + n\mu^2\sum_{i=1}^{n}\frac{\varsigma_i^2}{(\gamma_i+n\mu)^2} + \frac{(p-1)\sigma^2}{np}\sum_{i=1}^{n}\left(\frac{\gamma_i}{\gamma_i+n\mu}\right)^2 \quad .$$

On peut donc étudier chaque partie séparément pour obtenir un contrôle du risque oracle.

34. Cette variance inter-tâches n'est pas une variance de variables aléatoires. Pour rappeler son interprétation comme une variance, et pour la différencier de σ, nous noterons toujours cette quantité par la variante de la lettre sigma, quand cette dernière se trouve en fin de mot : ς.

Contrôle du risque oracle multi-tâches

Nous avons besoin d'hypothèses afin de contrôler la décroissance des suites (γ_i), (μ_i) et (ς_i). Ces hypothèses, très classiques, sont par exemple vérifiées dans le cas où le RKHS est un espace de Sobolev W_m et où les fonctions de régression sont suffisamment régulières. Voici ces hypothèses :

$$1 < 2\delta < 4\beta + 1 \ . \tag{$\mathbf{H_M}(\beta,\delta)$}$$

$$\forall i \in \{1,\dots,n\},\ \gamma_i = n i^{-2\beta} \ . \tag{$\mathbf{H_K}(\beta)$}$$

$$\forall i \in \{1,\dots,n\},\ \left\{ \begin{array}{rcl} \frac{\mu_i^2}{p} & = & C_1 n i^{-2\delta} \\ \varsigma_i^2 & = & C_2 n i^{-2\delta} \end{array} \right. . \tag{$\mathbf{H_{AV}}(\delta,C_1,C_2)$}$$

Sous ces hypothèses, le risque minimax est connu, vaut $(n/\sigma^2)^{1/2\delta-1}$, et peut être atteint par des estimateurs *ridge* à noyau. On pourra pour cela consulter l'article de Johnstone [Joh94], ou les livres de Wasserman [Was06] et de Massart [Mas07].

On peut alors étudier le risque oracle, ce qui permet d'aboutir au résultat suivant, en notant $\kappa(\beta,\delta)$ une constante ne dépendant que de β et δ.

Théorème 1.4. *Pour tout n, p, C_1, C_2, σ^2, β et δ tels que l'hypothèse $(\mathbf{H_M}(\beta,\delta))$ est vérifiée, on a*

$$\mathfrak{R}^\star_{\mathrm{MT}} \leq 2^{1/(2\delta)} \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta,\delta) \left[C_1^{1/(2\delta)} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta} \right] \ .$$

De plus, il existe des constantes N et $\alpha \in (0,1)$ telles que, si $n \geq N$, $p/\sigma^2 \leq n$ et $2 < 2\delta < 4\beta$, on a

$$\mathfrak{R}^\star_{\mathrm{MT}} \geq \alpha \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta,\delta) \left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta} \right] \ .$$

Contrôle du risque oracle mono-tâche

On cherche maintenant à faire de même pour le risque oracle mono-tâche. Malheureusement, les hypothèses précédentes ne correspondent pas à une seule répartition des tâches. On spécifie donc maintenant deux types de répartition des tâches, qui représentent notre hypothèse : les tâches sont groupées ensemble.

– Hypothèse « 2 points » : supposons, pour simplifier, que p est pair et que

$$f^1 = \cdots = f^{p/2} \ \text{ et } \ f^{p/2+1} = \cdots = f^p \ . \tag{2Points}$$

– Hypothèse « 1 *outlier* » :

$$f^1 = \cdots = f^{p-1} \ . \tag{1Out}$$

Ces hypothèses supposent, respectivement, que les fonctions sont toutes également réparties sur deux points, ou toutes sur un point avec une fonction en dehors. Elles sont très restrictives, mais nous mènerons des simulations dans des cas plus intéressants.

Sous ces hypothèses, nous pouvons maintenant étudier le risque oracle mono-tâche.

Corollaire 1.1. *Pour tout n, p, C_1, C_2, σ^2, β et δ tels que $2 < 2\delta < 4\beta$ et $n\sigma^2 > 1$, si les hypothèses* (2Points)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ et $(\mathbf{H_K}(\beta))$ sont vérifiées, alors*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \frac{\kappa(\beta,\delta)}{2} \times p^{1-1/2\delta} \left[\left(\sqrt{C_1}+\sqrt{C_2}\right)^{1/\delta} + \left|\sqrt{C_1}-\sqrt{C_2}\right|^{1/\delta}\right] \ .$$

Corollaire 1.2. *Pour tout n, p, C_1, C_2, σ^2, β et δ tels que $2 < 2\delta < 4\beta$ et $n\sigma^2 > 1$, si les hypothèses* (1Out)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ et $(\mathbf{H_K}(\beta))$ sont vérifiées, alors*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta,\delta) \times p^{1-1/2\delta} \left[\frac{p-1}{p}\left(\sqrt{C_1}+\sqrt{\frac{C_2}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|\sqrt{C_1}-\sqrt{(p-1)C_2}\right|^{1/\delta}\right] \ .$$

On remarquera, et c'était attendu, que les estimateurs mono-tâche ont un risque proche (à une constante près) du risque minimax. Ces estimations sont aussi suffisamment précises pour assurer que l'estimateur multi-tâches introduit précédemment a un risque quadratique négligeable par rapport à celui de l'oracle mono-tâche.

Comparaison entre les oracles mono-tâche et multi-tâches

On peut donc maintenant comparer les résultats précédents. On s'intéresse à la quantité

$$\rho = \frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} \ ,$$

c'est-à-dire au rapport entre les risques des oracles multi-tâches et mono-tâche, et on l'exprimera en fonction de

$$r = \frac{C_2}{C_1} \ .$$

Le paramètre r contrôle la part de signal qui est contenue dans la moyenne des tâches. Si r est petit, toutes les fonctions de régression sont très proches de leur moyenne et l'oracle multi-tâche devrait bien mieux fonctionner que l'oracle mono-tâche. Si r est grand, le contraire devrait arriver. On a alors les résultat suivants.

Corollaire 1.3. *Pour tout n, p, C_1, C_2, σ^2, β et δ tels que $2 < 2\delta < 4\beta$ et $n\sigma^2 > 1$, si les hypothèses* (2Points)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ et $(\mathbf{H_K}(\beta))$ sont vérifiées, alors*

$$\rho \asymp \frac{p^{1/(2\delta)-1} + (\frac{p-1}{p})^{1-(1/2\delta)} r^{1/2\delta}}{(1+\sqrt{r})^{1/\delta} + |1-\sqrt{r}|^{1/\delta}} \ .$$

Corollaire 1.4. *Pour tout n, p, C_1, C_2, σ^2, β et δ tels que $2 < 2\delta < 4\beta$ et $n\sigma^2 > 1$, si les hypothèses* (1Out)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ et $(\mathbf{H_K}(\beta))$ sont vérifiées, alors*

$$\rho \asymp \frac{p^{1/(2\delta)-1} + \left(\frac{p-1}{p}\right)^{1-(1/2\delta)} r^{1/2\delta}}{\frac{p-1}{p}\left(1+\sqrt{\frac{r}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|1-\sqrt{r(p-1)}\right|^{1/\delta}} \ .$$

Ces deux cas sont très différents :

- Quand r est petit, c'est-à-dire quand les tâches sont très similaires, dans les deux cas, ρ tend vers $Cst \times p^{1/2\delta-1}$, c'est-à-dire que l'oracle multi-tâches a la même efficacité que l'oracle mono-tâche ayant un échantillon p fois plus grand.
- Au contraire, quand r est grand, les deux situations diffèrent. D'un côté, sous l'hypothèse (2Points), ρ reste borné : l'oracle multi-tâche ne peut pas faire arbitrairement moins bien que le mono-tâche. De l'autre, sous (1Out), ρ tend vers $+\infty$: l'oracle multi-tâche fait arbitrairement moins bien que le mono-tâche.

Ces comportements se confirment sur des simulations (partie 4.8, page 106) dans un cadre plus étendu.

Le cas positif pour le multi-tâches n'est guère surprenant. Cependant, le cas de la situation défavorable au multi-tâches étudiée ici est moins clair : il est fort probable que, plus qu'une impossibilité d'utiliser toute méthode multi-tâches dans ce cas-là, c'est l'inadéquation du modèle $\mathcal{M}_{\mathrm{AV}}$ utilisé ici qui induit ce comportement. Faire une erreur de modélisation en incluant une tâche à tort dans un groupe peut donc être fort dommageable aux résultats d'une telle méthode multi-tâches, qui n'est donc pas robuste à de telles erreurs. D'où la nécessité de développer des méthodes qui puissent mieux et davantage s'adapter aux données !

Notations

We recall here some notations used throughout the manuscript.

Abbreviations.

a.k.a.	also known as
e.g.	*exempli gratia*
Eq.	Equation
et al.	*et alii*
etc.	*et cetera*
i.e.	*id est*
i.i.d.	independent and identically distributed
p.	page
resp.	respectively
RKHS	reproducing kernel Hilbert space
Sect.	Section

General mathematical notations.

$\mathbb{P}$, $\mathbb{E}$, Var	probability, expectation, variance
vec	operator which stacks the columns of a matrix into a vector
$\mathcal{M}_n(\mathbb{R})$	set of all real matrices of size n
$\mathcal{S}_p(\mathbb{R})$	set of symmetric matrices of size p
$\mathcal{S}_p^+(\mathbb{R})$	set of symmetric positive-semidefinite matrices of size p.
$\mathcal{S}_p^{++}(\mathbb{R})$	set of symmetric positive-definite matrices of size p
$O_p(\mathbb{R})$	set of orthogonal matrices of size p
$(e_1, \ldots, e_p)$	canonical basis of $\mathbb{R}^p$
$\preceq$	partial ordering on $\mathcal{S}_p(\mathbb{R})$ defined by: $A \preceq B$ if and only if $B - A \in \mathcal{S}_p^+(\mathbb{R})$
$\mathbf{1}$	vector of size p whose components are all equal to 1
$\|\cdot\|_2$	usual Euclidean norm on $\mathbb{R}^k$ for any $k \in \mathbb{N}$: $\forall u \in \mathbb{R}^k$, $\|u\|_2^2 := \sum_{i=1}^k u_i^2$

NOTATIONS

$\mathcal{N}(0,\Sigma)$	normal multivariate distribution, with mean 0 and covariance matrix Σ
$\widehat{\Sigma}$, $\widehat{C}$, etc.	estimators
I^c	complementary of the set I
$A \otimes B$	Kronecker product of matrices A and B

Notations used both in Chapter 3 and in Chapter 4.

n	sample size
p	number of tasks
$(X_i, Y_i)_{i=1}^n$	sample in the single-task setting
$(X_i, Y_i^1, \ldots, Y_i^p)_{i=1}^n$	sample in the multi-task setting (Sect 3.2.1, p. 48 and Sect 4.2.1, p. 88)
M	$p \times p$ matricial hyper-parameter
$\mathcal{X}$	input space
$\mathcal{F}$	set of target functions
$(F^1, \ldots, F^p)$	target functions (Eq. (3.1), p. 48 and Eq. (4.1), p. 88)
$(f^1, \ldots, f^p)$, f	target vectors in the fixed-design setting (Sect. 3.2.1, p. 49 and Sect. 4.2.1, p. 88
k	kernel of the RKHS (Sect. 3.2.1, p. 48 and Sect 4.2.1, p. 88)
K	kernel matrix (Sect. 3.2.1, p. 48 and Sect 4.2.1, p. 88)
Φ	kernel feature map (Sect. 3.2.1, p. 48) and Sect 4.2.1, p. 88
$\widehat{F}_M$, $\widehat{f}_M$	ridge multi-task estimator with regularization hyper-parameter M (Eq (3.2), p. 49 and Eq. (4.2), p. 89)
$M_{\mathrm{SD}}(\lambda,\mu)$, $M_{\mathrm{AV}}(\lambda,\mu)$	particular matricial hyper-parameters, suited for the multi-task hypothesis usually made here (Sect. 3.2.1, p. 50 and Sect. 4.2.2, p. 89)
A_M	ridge matrix for the multi-task estimator with regularization hyper-parameter M (Sect. 3.2.1, p. 51 and Sect. 4.2.1, p. 89)
$\mathcal{M}$, $\mathcal{M}_{\mathrm{SD}}$, $\mathcal{M}_{\mathrm{ind}}$, $\mathcal{M}_{\mathrm{clus}}$, $\mathcal{M}_{\mathrm{interval}}$	multi-task model, subset of $\mathcal{S}_p^+(\mathbb{R})$, on which the multi-task ridge estimator has to be calibrated
Hdf	assumption on the bias (Eq. (**Hdf**), p. 55 and Eq. (**Hdf**), p. 104)
$\mathbb{H}_0$, $\mathbb{H}_1$	null and alternative hypothesis
N	number of replications in the simulation experiments

Notations for Chapter 3.

$\mu_{\min}(\Sigma)$ (resp. $\mu_{\max}(\Sigma)$)	smallest (resp. largest) eigenvalue of Σ
$c(\Sigma)$	condition number of Σ, that is, $\mu_{\max}(\Sigma)/\mu_{\min}(\Sigma)$

$M_{\mathrm{ind}}(\lambda^1, \ldots, \lambda^p)$	matricial hyper-parameters that gives the single-task estimator (Sect. 3.2.1, p. 49)
$M_I(\lambda, \mu, \nu)$	matricial hyper-parameters that clusters the tasks in two clusters, indicated by the sets I and I^c (Sect. 3.2.1, p. 50)
$(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$	RKHS for the multi-task ridge regression (Sect. 3.2.1, p. 50)
$M^\star$	oracle hyper-parameter, with respect to $\mathcal{M}$ (Sect. 3.2.2, p. 51)
$\mathrm{pen}_{\mathrm{id}}(M)$	ideal penalty to calibrate the matrix M (Eq. (3.7), p. 53)
$\mathrm{df}(\lambda)$	degrees of freedom of A_λ (that is, its trace), the ridge matrix with hyper-parameter λ (Sect. 3.3, p. 53)
$\mathrm{pen}_{\mathrm{min}}(\lambda)$	minimal penalty, used to estimate the noise variance (Sect. 3.3, p. 53)
$\mathcal{Z}$	set of points alongside which the noise covariance matrix is estimated (Sect. 3.4, p. 55)
J	maps that builds a symmetric matrix out of its outputs on the canonical basis (Eq. (3.4), p. 55)
Ω	high probability event on which the oracle inequalities are shown to hold.
HM	structural assumption on the multi-task model which allows a simpler estimation (Eq. (**HM**), p. 58)
δ, m	fixed quantities during the simulation experiments (Sect. 3.6, p. 60)
α, z	quantities randomly drawn during the simulation experiments (Sect. 3.6, p. 60)
L_1, L_2, L_3, L_4, $\widetilde{L_4}$	fixed but uncalculated constants

Notations for Chapter 4.

σ^2	noise intensity ($\Sigma = \sigma^2 I_p$) (Sect. 4.2.1, p. 88)
λ, μ, λ_1, …, λ_p	regularization parameters
$M_{\mathrm{AV}}(\lambda, \mu)$	particular matricial hyper-parameters, suited for the multi-task hypothesis usually made here (Sect. 4.2.2, p. 90)
$\mathfrak{R}^\star_{\mathrm{ST}}$ (resp. $\mathfrak{R}^\star_{\mathrm{MT}}$)	single-task (resp. multi-task) oracle risk (Sect. 4.3, p. 90)
$(\gamma_i)_{i=1}^n$	eigenvalues of K
$(h_i^j)_{i=1}^p$	coordinated of the jth regression function on the orthonormal that diagonalizes K (Sect. 4.3.1, p. 91)
$(\mu_i/\sqrt{p})_{i=1}^n$ (resp. $(\varsigma_i)_{i=1}^n$)	coordinates of the mean (resp. variance) of the p regression tasks in the orthonormal basis that diagonalizes K (Sect. 4.3.3, p. 93)
β, δ	regularity parameters of the kernel and of the signal

NOTATIONS

$\mathbf{H_M}(\beta, \delta)$	assumption that links β and δ and ensures the ridge estimator is minimax optimal (Eq. ($\mathbf{H_M}(\beta, \delta)$), p. 94)
$\mathbf{H_K}(\beta)$	assumption on the regularity of the kernel (Eq. ($\mathbf{H_K}(\beta)$), p. 94)
$\mathbf{H_{AV}}(\delta, C_1, C_2)$	assumption on the regularity of the signal (Eq. ($\mathbf{H_{AV}}(\delta, C_1, C_2)$), p. 94)
C_1 (resp. C_2)	parameter that controls the strength of the mean (resp. variance) of the p regression tasks
$R(n, p, \sigma^2, \cdot, \beta, \delta, C)$	risk of a ridge estimator, depending on its regularization parameter (Eq. (4.8), p. 95)
$R^\star(n, p, \sigma^2, \beta, \delta, C)$	infimum of $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$ (Sect. 4.4.1, p. 96)
$\kappa(\beta, \delta)$	constant that only depends on β and δ (Eq. (4.23), p. 120)
$\lambda_R^\star$	parameter which minimizes $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$ (Sect. 4.4.1, p. 97)
$\lambda^\star$, $\mu^\star$	oracle regularization hyper-parameters (Sect. 4.4.2, p. 98)
2Points, 1Out	assumptions that control the repartitions of the p regression functions (Sect. 4.5, p. 99)
ρ	ratio between the multi-task and single-task oracle risks (Sect. 4.6, p. 101)
r	ratio between C_1 and C_2 (Sect. 4.6, p. 101)

Throughout the manuscript, we will also try to keep the following two conventions.

– Concerning observations, letter i refers to the position of the observation in the sample (between 1 and n) and is subscripted, while letter j refers to the index of the task (between 1 and p) and is superscripted. Thus, X_i^j refers to the ith observation for the jth task, X_i refers to the ith observation and does not depend on which task is considered and f^j is an object which is related to the jth task.

– Concerning signals, the function itself is capitalized (for instance, G) while the corresponding vector whose coordinates are the value of this function on the observation points is lowercased (for instance, g).

Chapitre 2

Main contributions of the thesis

RÉSUMÉ. Nous détaillons ici, en anglais, les contributions principales que cette thèse apporte à notre domaine. Le contenu de ce chapitre correspond à la partie 1.3, écrite en français.

We detail here the main contributions this thesis brings to our field. This chapter corresponds to Section 1.3, which is written in French.

2.1 Framework and model

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We suppose that, for a given $n \in \mathbb{N}$, we observe the sample $\mathcal{D}_n = (X_i, Y_i^1, \dots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$. For every task $j \in \{1, \dots, p\}$, $\mathcal{D}_n^j = (X_i, Y_i^j)_{i=1}^n$ is an n-sample of distribution $\mathcal{P}^j$, whose first marginal is $\mathcal{P}$. We seek to solve a regression problem for each task. We now detail our model. We first suppose there exists $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$ and i.i.d. vectors $(\varepsilon_i^j)_{j=1}^p$ following a $\mathcal{N}(0, \Sigma)$ distribution. We also suppose that for every $j \in \{1, \dots, p\}$, there exists $F^j \in L^2(\mathbb{P})$ such that

$$\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, p\},\ Y_i^j = F^j(X_i) + \varepsilon_i^j \ .$$

Remark 2.1. *We suppose here that all the tasks have the same design. This facilitates the theoretical study, which could be done without it, at the price of added assumptions.*

Remark 2.2. *The matrix Σ is the covariance matrix of the noise between the tasks, which are not necessarily independent conditionally on $(X_i)_{i=1}^n$. The estimation of this matrix is of major importance here.*

We now consider a fixed design regression setting, the risk of interest being here, for an estimator $\widehat{F}$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (F(X_i) - \widehat{F}(X_i))^2 \middle| X_1, \dots, X_n\right] \ .$$

All the expectations are implicitly noted conditionally on $(X_1, \dots, X_n)$, so that the notations are kept simple. We shall also note

$$f = \mathrm{vec}\left((F^j(X_i))_{i,j}\right),\ f^j = \mathrm{vec}\left((F^j(X_i))_{i=1}^n\right) \quad \text{et} \quad y = \mathrm{vec}\left((Y_i^j)_{i,j}\right) \ ,$$

and consider similar notations for the estimators. With such notations, the elements are grouped task by task in vectors, beginning by those related to the first task and finishing by those related to the last. We shall use the quadratic loss, denoted by $\left\|\widehat{f}-f\right\|^2$, and the associated quadratic risk, $\mathbb{E}\left[\left\|\widehat{f}-f\right\|^2\right]$.

We consider a kernel ridge regression setting. Given a RKHS $\mathcal{F}\subset L^2(\mathbb{P})$, whose associated kernel is k and whose feature function is $\Phi:\mathcal{X}\to\mathcal{F}$. This gives us the kernel matrix $K=(k(X_i,X_\ell))_{1\leq i,\ell\leq n}\in\mathcal{S}_n^+(\mathbb{R})$. We now seek to build estimators in $\mathcal{F}$. To do this, following the works of Brown and Zidek [BZ80] and of Evgeniou et al. [EMP05], we consider the estimator which is solution of the minimization problem, depending of the regularization parameter $M\in\mathcal{S}_p^+(\mathbb{R})$,

$$\widehat{F}_M\in\operatorname*{argmin}_{G\in\mathcal{F}^p}\left\{\underbrace{\frac{1}{np}\sum_{i=1}^n\sum_{j=1}^p(Y_i^j-G^j(X_i))^2}_{\text{Empirical risk}}+\underbrace{\sum_{j=1}^p\sum_{\ell=1}^p M_{j,l}\langle G^j,G^\ell\rangle_{\mathcal{F}}}_{\text{Regularization term}}\right\}\ .$$

We can then build a RKHS, which depends on M, which enables us to use the Representer's Theorem and obtain the fixed-design estimator

$$\widehat{f}_M=A_My=\widetilde{K}_M(\widetilde{K}_M+npI_{np})^{-1}y=(M^{-1}\otimes K)\left((M^{-1}\otimes K)+npI_{np}\right)^{-1}y\ .$$

The matrix M encodes the similarity between the tasks. Calibrating this matrix should allow us to be adapted, at least partly, to this similarity. Finally, here are two examples of such a matrix M, which we will often use hereafter.

Example 2.1. *If we want to treat the p tasks separately, thus obtaining the single-task estimator, we can consider $M=M_{\mathrm{ind}}(\lambda):=\frac{1}{p}\operatorname{Diag}(\lambda_1,\dots,\lambda_p)$ for every $\lambda\in\mathbb{R}^p$. This leads to the regularization term*

$$\frac{1}{p}\sum_{j=1}^p\left[\frac{1}{n}\sum_{i=1}^n(Y_i^j-G^j(X_i))^2+\lambda_j\|G^j\|_{\mathcal{F}}^2\right]\ ,$$

which decouples along the tasks

Example 2.2. *We can follow Evgeniou et al. [EMP05] and define, for every $(\lambda,\mu)\in(0,+\infty)^2$,*

$$M_{\mathrm{SD}}(\lambda,\mu):=(\lambda+p\mu)I_p-\mu\mathbf{1}\mathbf{1}^\top=\begin{pmatrix}\lambda+(p-1)\mu & & -\mu\\ & \ddots & \\ -\mu & & \lambda+(p-1)\mu\end{pmatrix}\ .$$

With $M=M_{\mathrm{SD}}(\lambda,\mu)$, we obtain the regularizer

$$\frac{1}{np}\sum_{i=1}^n\sum_{j=1}^p(Y_i^j-G^j(X_i))^2+\lambda\sum_{j=1}^p\left\|G^j\right\|_{\mathcal{F}}^2+\frac{\mu}{2}\sum_{j=1}^p\sum_{k=1}^p\left\|G^j-G^k\right\|_{\mathcal{F}}^2\ .$$

This allows us to regularize both the norms of the functions G^j and of their differences G^j-G^k. Hence, the matrices $M_{\mathrm{SD}}(\lambda,\mu)$ can be used when the functions F^j are close in $\mathcal{F}$.

2.2 Calibration of a multi-task estimator

We want to select a matrix M from a set $\mathcal{M}$, so that the risk of the associated estimator $\widehat{f}_M$ has a low risk. In order to do this, we will penalize the empirical risk by an "ideal penalty".

2.2.1 Ideal penalization of the empirical risk

We look for an entirely data-driven penalty, which best mimics the ideal penalty

$$\mathrm{pen}_{\mathrm{id}}(M) := \mathbb{E}\left[\frac{1}{np}\|\widehat{f}_M - f\|_2^2\right] - \mathbb{E}\left[\frac{1}{np}\left\|y - \widehat{f}_M\right\|_2^2\right] \ .$$

A simple calculation shows that, up to a term that does not depend on M,

$$\mathrm{pen}_{\mathrm{id}}(M) = \frac{2\,\mathrm{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} \ .$$

However, this depends on Σ, which we do not know. The first stage of this work is thus to estimate Σ and to show that this estimate is precise enough so that the plug-in penalty

$$\widehat{\mathrm{pen}}(M) = \frac{2\,\mathrm{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n)\right)}{np} \tag{2.1}$$

approaches sufficiently well the ideal penalty.

2.2.2 Estimation de Σ

Our estimator of the covariance matrix Σ is based on the concept of minimal penalty. We shall not discuss these penalties in details here, and refer the interested reader to the article of Arlot and Bach [AB11] or to the short summary in section 3.3, page 53. We will just say that they allow, in our framework, to estimate the variance of the noise in a single-task framework.

Our estimation strategy is the following:

1. Select a set of directions in $\mathbb{R}^p$, where each coordinate represents a task.
2. For each direction, consider the single-task problem corresponding to the multi-task projection along the chosen dimension and estimate the variance of the noise along this dimension.
3. Build $\widehat{\Sigma}$ back from those one-dimensional estimations.

We precise this here. For every $z \in \mathbb{R}^p$, we consider the single-task regression problem

$$Y_z := Y \cdot z = F \cdot z + E \cdot z = F_z + \varepsilon_z \ . \tag{P_z}$$

We can denote by $a(z)$ the estimator of the variance of the noise of problem (P_z) and $(e_1, \ldots, e_p)$ the canonical basis of $\mathbb{R}^p$. Then, we see that $a(e_i)$ estimates $\Sigma_{i,i}$ and that $a(e_i{+}e_j)$ estimates $\Sigma_{i,i}{+}\Sigma_{j,j}{+}2\Sigma_{i,j}$. So, $\Sigma_{i,j}$ can be estimated by $(a(e_i{+}e_j){-}a(e_i){-}a(e_j))/2$.

Therefore, we introduce the function $J : \mathbb{R}^{p(p+1)/2} \mapsto \mathcal{S}_p$, defined by

$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,i} = a_i \text{ si } 1 \leq i \leq p \ ,$$
$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,j} = \frac{a_{i,j} - a_i - a_j}{2} \text{ si } 1 \leq i < j \leq p \ .$$

We can see that

$$\Sigma = J(\Sigma_{1,1}, \dots, \Sigma_{p,p}, \Sigma_{1,1} + \Sigma_{2,2} + 2\Sigma_{1,2}, \dots)$$

and we denote

$$\widehat{\Sigma} := J\left(a(e_1), \dots, a(e_p), a(e_1 + e_2), \dots, a(e_1 + e_p), \dots, a(e_{p-1} + e_p)\right) \ . \tag{2.2}$$

We can then show the following result, denoting by $c(\Sigma)$ the condition number of Σ, $\preceq$ the order relation defined by $A \preceq B$ if $B - A$ is symetric positive semi-definite and by introducing an assumption on the bias of the model:

$$\left.\begin{array}{l} \forall j \in \{1, \dots, p\}\,, \exists \lambda_{0,j} \in (0, +\infty)\,, \\ \qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{and} \quad \frac{1}{n}\left\|(A_{\lambda_{0,j}} - I_n)F_{e_j}\right\|_2^2 \leq \Sigma_{j,j}\sqrt{\frac{\ln n}{n}} \end{array}\right\} \tag{\textbf{Hdf}}$$

Theorem 2.1. *Let $\widehat{\Sigma}$ be the estimator defined by Eq.* (1.10) *and suppose that Eq.* (**Hdf**) *holds. For every $\delta \geq 2$, there exists a constant $n_0(\delta)$, a constant $L_1 > 0$ and an event $\widetilde{\Omega}$, verifying $\mathbb{P}(\widetilde{\Omega}) \geq 1 - p(p+1)/2 \times n^{-\delta}$, such that, if $n \geq n_0(\delta)$, on $\widetilde{\Omega}$,*

$$(1 - \eta)\Sigma \preceq \widehat{\Sigma} \preceq (1 + \eta)\Sigma \tag{2.3}$$
$$\text{with} \qquad \eta := L_1(2 + \delta)p\sqrt{\frac{\ln(n)}{n}}c(\Sigma)^2 \ .$$

Consequently, our estimator Σ is consistent and converges to Σ with a rate that is given here.

The most often studied case in covariance matrix estimation is the one where f is constant or null, one may then seek to improve the performance of the empirical covariance matrix. Bickel and Levina [BL08] or Cai et al. [CZZ10], for instance, use thresholding methods to obtain, in the second article, minimax convergence rates. Others formulate a sparsity assumption and then use thresholding methods, like Karoui [Kar08], or regularization methods, as in Lam and Fan [LF09].

Our framework is quite far away from those methods, since our sample $(Y_i^1, \dots, Y_i^p)_{i=1}^n$ is nor centered nor homoscedastic. Constructing an empirical covariance matrix makes no sense here

2.2.3 Oracle inequality

We can now guarantee the efficiency of the estimator $\widehat{f}_{\widehat{M}}$, calibrated by the penalization scheme (2.1), by showing an oracle inequality. Our results cover two kinds of situations:

1. a situation where the matrix set $\mathcal{M}$ is discrete, which can help to deal with setting where few *a priori* knowledge on the repartition of the tasks is available ;

2. a situation where the matrix set $\mathcal{M}$ is jointly diagonalizable in an orthonormal basis, which happens in settings where strong *a priori* knowledge on the repartition of the tasks is available, for instance when all the tasks are known to be grouped in several clusters.

In the first situation, we suppose $\mathcal{M}$ verifies

$$\exists (C, \alpha_{\mathcal{M}}) \in (0, +\infty)^2, \quad \mathrm{card}(\mathcal{M}) < C n^{\alpha_{\mathcal{M}}} \ . \tag{2.4}$$

We can then define the selected parameter by

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2 \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\} \ .$$

We can then enunciate the oracle inequality as follows, denoting by $\sigma_{\max}$ the largest eigenvalues of Σ.

Theorem 2.2. *Let $\alpha = \max(\alpha_{\mathcal{M}}, 2)$, $\delta \geq 2$ and suppose Assumptions* (**Hdf**) *and* (2.4) *hold. Then, there exists constants $L_2, \kappa' > 0$, a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$, verifying $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa' p(p + C) n^{-\delta}$, such that, if $n \geq n_1(\delta)$, on $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 c(\Sigma)^4 \sigma_{\max} (\alpha + \delta)^2 \frac{p^4 \ln(n)^3}{np} \ .$$

This is a simplified version of Theorem 3.3, page 57. The original result also has a version that is stated in expectation.

In the second situation, we suppose $\mathcal{M}$ verifies

$$\exists P \in O_p(\mathbb{R}) \,, \quad \mathcal{M} \subseteq \left\{ P^\top \operatorname{Diag}(d_1, \ldots, d_p) P \,,\, (d_i)_{i=1}^p \in (0, +\infty)^p \right\} \ . \tag{\textbf{HM}}$$

We define $(u_i)_{i=1}^p$ by $\forall j \in \{1, \ldots, p\},\ u_j = P^\top e_j$, and we then estimate Σ by

$$\widehat{\Sigma}_{\mathrm{HM}} = P \operatorname{Diag}(a(u_1), \ldots, a(u_p)) P^\top \ ,$$

The selected parameter then is

$$\widehat{M}_{\mathrm{HM}} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2 \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma}_{\mathrm{HM}} \otimes I_n) \right) \right\} \ . \tag{2.5}$$

We can then enunciate the oracle inequality as follows,

Theorem 2.3. *Let $\alpha = 2$, $\delta \geq 2$ and suppose Assumptions* (**Hdf**) *and* (**HM**) *hold. Then, there exists constants $L_2 > 0$, κ'', a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$, verifying $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa'' p n^{-\delta}$, such that, if $n \geq n_1(\delta)$, on $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_{HM}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 \sigma_{\max} (2 + \delta)^2 \frac{p \ln(n)^3}{n} \ .$$

This is a simplified version of Theorem 3.4, page 58. The original result also has a version that is stated in expectation.

We can remark several things.

– Gaining additional information on the repartition of the tasks allows to obtain a simplified algorithm which has a stronger convergence guarantee.
– In order for the oracle inequalities to constrain the risk of the estimator, that is so that the right-hand term is neglictible in front of the infimum, n and p have to be constrained, which excludes situations like $n \gg p$. Those constraints are discussed in Remark 3.13, page 59.
– Simulated experiments confirm the behaviour of our estimator in situations that do not constrain its risk via the oracle inequalities (Section 3.6, page 60).

2.3 Does multi-task work ?

We now have a multi-task estimator which is able to adapt to a parameter family $\mathcal{M}$, that is to choose a parameter $\widehat{M} \in \mathcal{M}$ whose risk is close to the best risk on $\mathcal{M}$. We can then ask the following question: does the estimator thus obtained performs better than the associated single-task estimator ? Considering our preceding results, it suffices to compare the risks of the oracle estimators. We suppose for simplicity that $\Sigma = \sigma^2 I_p$.

2.3.1 Decomposition of the risk

We will study the case where the p regression functions belonging to each task are supposed to be close. We know some matrices which can be used to regularize those functions and their differences. However, a quick study of this risk shows that it is smarter to regularize both the mean of those functions and their variance. Henceforth, we use the following set of matrices:

$$\mathcal{M}_{\mathrm{AV}} = \left\{ M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}^2 \right\} \quad ,$$

with

$$M_{\mathrm{AV}}(\lambda, \mu) := \frac{\lambda}{p} \frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p} \left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right) \quad .$$

This leads to the following criterion:

$$\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (y_i^j - G^j(X_i))^2 + \lambda \left\| \frac{\sum_{j=1}^p G^j}{p} \right\|_{\mathcal{F}}^2 + \mu \left[\frac{\sum_{j=1}^p \left\| G^j \right\|_{\mathcal{F}}^2}{p} - \left\| \frac{\sum_{j=1}^p G^j}{p} \right\|_{\mathcal{F}}^2 \right] \quad .$$

The single-task oracle then is

$$\mathfrak{R}_{\mathrm{ST}}^\star = \inf_{(\lambda^1, \ldots, \lambda^p) \in \mathbb{R}_+^p} \left\{ \frac{1}{np} \mathbb{E} \left[\sum_{j=1}^{p} \left\| \widehat{f}_{\lambda^j}^j - f^j \right\|_2^2 \right] \right\}$$

and the multi-task oracle risk is

$$\mathfrak{R}_{\mathrm{MT}}^\star = \inf_{(\lambda, \mu) \in \mathbb{R}_+^2} \left\{ \frac{1}{np} \mathbb{E} \left[\left\| \widehat{f}_{M_{\mathrm{AV}}(\lambda, \mu)} - f \right\|_2^2 \right] \right\} \quad .$$

We can first remark that the single-task oracle risk $\mathfrak{R}_{\mathrm{ST}}^\star$ is an infimum over p parameters, while the multi-task oracle risk $\mathfrak{R}_{\mathrm{MT}}^\star$ is an infimum over only two parameters. The single-task

oracle thus has more degrees of freedom than the multi-task one, but cannot simultaneously use data from different tasks. Hence, it is not obvious that the sole guarantee "$\mathfrak{R}^\star_{\mathrm{MT}} \leq \mathfrak{R}^\star_{\mathrm{ST}}$" can be obtained.

Let us denote by $(\gamma_i)_{i=1}^n$ the eigenvalues of K and, for every $j \in \{1,\dots,p\}$, by $(h_i^j)_{i=1}^n$ the coordinates of f^j on the orthonormal basis that diagonalised K. Let us also denote by

$$\mu_i = \nu_i^1 = \frac{h_i^1 + \cdots + h_i^p}{\sqrt{p}}$$

and the inter-task "variance"[1]

$$\varsigma_i^2 = \frac{\sum_{j=1}^p (h_i^j)^2}{p} - \left(\frac{\sum_{j=1}^p h_i^j}{p}\right)^2 = \frac{1}{p}\sum_{j=1}^p \left(h_i^j - \frac{\sum_{j=1}^p h_i^j}{p}\right)^2 \quad .$$

We can then write the risk of the estimator $\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)}$, thanks to the bias-variance decomposition:

$$n\lambda^2 \sum_{i=1}^n \frac{\frac{\mu_i^2}{p}}{(\gamma_i + n\lambda)^2} + \frac{\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\lambda}\right)^2 + n\mu^2 \sum_{i=1}^n \frac{\varsigma_i^2}{(\gamma_i + n\mu)^2} + \frac{(p-1)\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\mu}\right)^2 \quad .$$

We can then study each part separately to obtain a control on the multi-task oracle risk

2.3.2 Control of the multi-task oracle risk

We need some assumptions to control the decay of the sequences (γ_i), (μ_i) and (ς_i). Those very classical assumptions are for instance verified in the case where the RKHS is a Sobolev space W_m and where the regression functions are regular enough. Those assumptions are:

$$1 < 2\delta < 4\beta + 1 \ . \qquad (\mathbf{H_M}(\beta,\delta))$$

$$\forall i \in \{1,\dots,n\},\ \gamma_i = n i^{-2\beta} \ . \qquad (\mathbf{H_K}(\beta))$$

$$\forall i \in \{1,\dots,n\},\ \begin{cases} \frac{\mu_i^2}{p} &= C_1 n i^{-2\delta} \\ \varsigma_i^2 &= C_2 n i^{-2\delta} \end{cases} \ . \qquad (\mathbf{H_{AV}}(\delta, C_1, C_2))$$

Under those assumptions, the minimax rate is known and is of the order of $(n/\sigma^2)^{1/2\delta-1}$, and can be matched by kernel ridge estimators. See the article of Johnstone [Joh94], or the books of Wasserman [Was06] and Massart [Mas07] for more details.

We can then study the oracle risk, which leads to the following result, denoting by $\kappa(\beta,\delta)$ a constant which only depends on β and δ.

1. This inter-task variance is not a probabilistic variance. To remind its interpretation as a variance, and to differenciate it from σ, we will always denote it by the variant of the letter sigma, as it is written when located at the end of a word: ς.

Theorem 2.4. *For every n, p, C_1, C_2, σ^2, β and δ such that Assumption $(\mathbf{H_M}(\beta,\delta))$ holds, we have*

$$\mathfrak{R}^\star_{\mathrm{MT}} \leq 2^{1/(2\delta)} \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta,\delta) \left[C_1^{1/(2\delta)} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta}\right] \ .$$

Moreover, there exists constants N and $\alpha \in (0,1)$ such that, if $n \geq N$, $p/\sigma^2 \leq n$ and $2 < 2\delta < 4\beta$, we have

$$\mathfrak{R}^\star_{\mathrm{MT}} \geq \alpha \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta,\delta) \left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta}\right] \ .$$

2.3.3 Control of the single-task oracle risk

We now try to obtain a similar result for the single-task oracle risk. Unfortunately, the former assumptions do not correspond to only one repartition of the tasks. We now specify two kinds of those repartitions, which represent our hypothesis : the tasks are grouped together.

– Assumption "2 points": suppose, for simplicity, that p is even and that

$$f^1 = \cdots = f^{p/2} \ \text{ et } \ f^{p/2+1} = \cdots = f^p \ . \tag{2Points}$$

– Assumption "1 outlier":

$$f^1 = \cdots = f^{p-1} \ . \tag{1Out}$$

Those assumptions assume, respectively, that the p regression functions are equally split over two points, or are all gathered on one point excepted for one outlier. They are extremely restrictive, but we shall relax them later by running simulations.

Under those assumptions, we can now study the single-task oracle risk.

Corollary 2.1. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > 1$, if Assumptions* (2Points)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ and $(\mathbf{H_K}(\beta))$ hold, then*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \frac{\kappa(\beta,\delta)}{2} \times p^{1-1/2\delta} \left[\left(\sqrt{C_1} + \sqrt{C_2}\right)^{1/\delta} + \left|\sqrt{C_1} - \sqrt{C_2}\right|^{1/\delta}\right] \ .$$

Corollary 2.2. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > 1$, if Assumptions* (1Out)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ and $(\mathbf{H_K}(\beta))$ hold, then*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta,\delta) \times p^{1-1/2\delta} \left[\frac{p-1}{p}\left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|\sqrt{C_1} - \sqrt{(p-1)C_2}\right|^{1/\delta}\right] \ .$$

Notice that, as expected, the single-task oracle have a risk which is close, up to a constant, to the minimax risk. Those estimations are precise enough to ensure that the multi-task oracle introduced before has a quadratic risk which is neglictible compared to the single-task oracle risk, in favourable situations.

2.3.4 Comparison between single-task and multi-task oracle risks

We can now compare the results obtained previously. We will look at the quantity

$$\rho = \frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} \ ,$$

which we will express in terms of

$$r = \frac{C_2}{C_1} \ .$$

The parameter r controls the amount of the signal held in the mean of the tasks. If r is small, all the regression functions are close to their mean and the multi-task oracle should perform better than the single-task one. However, if r is large, the contrary should happen. We can then obtain the following results.

Corollary 2.3. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > 1$, if Assumptions* (2Points)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ and $(\mathbf{H_K}(\beta))$ hold, then*

$$\rho \asymp \frac{p^{1/(2\delta)-1} + (\frac{p-1}{p})^{1-(1/2\delta)} r^{1/2\delta}}{(1+\sqrt{r})^{1/\delta} + |1-\sqrt{r}|^{1/\delta}} \ .$$

Corollary 2.4. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > 1$, if Assumptions* (1Out)*, $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ and $(\mathbf{H_K}(\beta))$ hold, then*

$$\rho \asymp \frac{p^{1/(2\delta)-1} + \left(\frac{p-1}{p}\right)^{1-(1/2\delta)} r^{1/2\delta}}{\frac{p-1}{p}\left(1+\sqrt{\frac{r}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|1-\sqrt{r(p-1)}\right|^{1/\delta}} \ .$$

Those two examples show different behaviours.

- When r is small, in both situations, ρ goes to $Cst \times p^{1/2\delta-1}$, that is, the multi-task oracle performs similarly than the single-task oracle with a p times larger sample.
- When r is large, the two situations differ. On the one side, under Assumption (2Points), ρ stays bounded: the multi-task oracle cannot perform arbitrarily worse than the single-task one. On the other side, under Assumption (1Out), ρ goes to $+\infty$: the multi-task oracle performs arbitrarily worse than the single-task one.

Those behaviours are confirmed on simulated examples (Section 4.8, page 106) in a broader setting.

The positive behaviour of the multi-task estimator is hardly surprising. However, the situation of the case where the multi-task fails is unclear: it is probable that, more than an impossibility to obtain any multi-task procedure here, it is the model $\mathcal{M}_{\mathrm{AV}}$ itself which is not fitted and that induces this behaviour. Committing a modelisation error by wrongly including a task in a cluster it does not belong to can therefore be extremely damaging to this kind of multi-task procedure, which is now showed to be non robust to such errors. It is therefore crucial to develop procedures that can adapt better to the data !

Chapitre 3

Multi-task Regression using Minimal Penalties

Résumé. Dans ce chapitre, nous introduisons et étudions une méthode de régression ridge, à noyau, dans un cadre multi-tâches, en utilisant une technique de pénalisation. L'analyse théorique qui y est menée montre que la calibration optimale de cette méthode repose sur l'estimation de la matrice de covariance du bruit, entre les différentes tâches. Nous avons recours à un nouvel algorithme permettant de mener à bien cette estimation, fondé sur le concept de pénalité minimale—qui est utilisée dans un contexte mono-tâches pour estimer la variance du bruit. Ensuite, nous nous assurons de la consistance de cet estimateur, dans un cadre non asymptotique et sous de faibles hypothèses. Enfin, l'injection de cet estimateur dans la pénalité correspondante permet d'obtenir une inégalité oracle, qui certifie une certaine forme d'optimalité. Des simulations sur un jeu de donnée artificiel viennent compléter notre étude de cet estimateur, qui confirment les analyses décritent précédemment.

3.1 Introduction

A classical paradigm in statistics is that increasing the sample size (that is, the number of observations) improves the performance of the estimators. However, in some cases it may be impossible to increase the sample size, for instance because of experimental limitations. Hopefully, in many situations practicioners can find many related and similar problems, and might use these problems as if more observations were available for the initial problem. The techniques using this heuristic are called "multi-task" techniques. In this paper we study the kernel ridge regression procedure in a multi-task framework.

One-dimensional kernel ridge regression, which we refer to as "single-task" regression, has been widely studied. As we briefly review in Section 3.3 one has, given n data points $(X_i, Y_i)_{i=1}^n$, to estimate a function f, often the conditional expectation $f(X_i) = \mathbb{E}[Y_i|X_i]$, by minimizing the quadratic risk of the estimator regularized by a certain norm. A practically important task is to calibrate a regularization parameter, that is, to estimate the regularization parameter directly from data. For kernel ridge regression (a.k.a. smoothing splines),

many methods have been proposed based on different principles, for example, Bayesian criteria through a Gaussian process interpretation [RW06] or generalized cross-validation [Wah90]. In this paper, we focus on the concept of minimal penalty, which was first introduced by Birgé and Massart[BM07] and Arlot and Massart[AM09] for model selection, then extended to linear estimators such as kernel ridge regression by Arlot and Bach [AB11].

In this article we consider $p \geq 2$ different (but related) regression tasks, a framework we refer to as "multi-task" regression. This setting has already been studied in different papers. Some empirically show that it can lead to performance improvement [TO96, Car97, BH03]. Liang et al. [LBBJ10] also obtained a theoretical criterion (unfortunately non observable) which tells when this phenomenon asymptotically occurs. Several different paths have been followed to deal with this setting. Some consider a setting where $p \gg n$, and formulate a sparsity assumption which enables to use the group Lasso, assuming all the different functions have a small set of common active covariates [OWJ11, LPTvdG11]. We exclude this setting from our analysis, because of the Hilbertian nature of our problem, and thus will not consider the similarity between the tasks in terms of sparsity, but rather in terms of an Euclidean similarity. Another theoretical approach has also been taken (see for example, Brown and Zidek [BZ80], Evgeniou et al. [EMP05] or Ando and Zhang [AZ05] on semi-supervised learning), the authors often defining a theoretical framework where the multi-task problem can easily be expressed, and where sometimes solutions can be computed. The main remaining theoretical problem is the calibration of a matricial parameter M (typically of size p), which characterizes the relationship between the tasks and extends the regularization parameter from single-task regression. Because of the high dimensional nature of the problem (i.e., the small number of training observations) usual techniques, like cross-validation, are not likely to succeed. Argyriou et al. [AEP08] have a similar approach to ours, but solve this problem by adding a convex constraint to the matrix, which will be discussed at the end of Section 3.5.

Through a penalization technique we show in Section 3.2 that the only element we have to estimate is the correlation matrix Σ of the noise between the tasks. We give here a new algorithm to estimate Σ, and show that the estimation is sharp enough to derive an oracle inequality for the estimation of the task similarity matrix M, both with high probability and in expectation. Finally we give some simulation experiment results and show that our technique correctly deals with the multi-task settings with a low sample-size.

The notations used here are recapitulated at the end of the introduction (page 33)

3.2 Multi-task Regression: Problem Set-up

We consider p kernel ridge regression tasks. Treating them simultaneously and sharing their common structure (e.g., being close in some metric space) will help in reducing the overall prediction error.

3.2.1 Multi-task with a Fixed Kernel

Let $\mathcal{X}$ be some set and $\mathcal{F}$ a set of real-valued functions over $\mathcal{X}$. We suppose $\mathcal{F}$ has a reproducing kernel Hilbert space (RKHS) structure [Aro50], with kernel k and feature map $\Phi : \mathcal{X} \to \mathcal{F}$. We observe $\mathcal{D}_n = (X_i, Y_i^1, \dots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$, which gives us the positive

semidefinite kernel matrix $K = (k(X_i, X_\ell))_{1\leq i,\ell\leq n} \in \mathcal{S}_n^+(\mathbb{R})$. For each task $j \in \{1, \dots, p\}$, $\mathcal{D}_n^j = (X_i, y_i^j)_{i=1}^n$ is a sample with distribution $\mathcal{P}_j$, for which a simple regression problem has to be solved. In this paper we consider for simplicity that the different tasks have the same design $(X_i)_{i=1}^n$. When the designs of the different tasks are different the analysis is carried out similarly by defining $X_i = (X_i^1, \dots, X_i^p)$, but the notations would be more complicated.

We now define the model. We assume $(F^1, \dots, F^p) \in \mathcal{F}^p$, Σ is a symmetric positive-definite matrix of size p such that the vectors $(\varepsilon_i^j)_{j=1}^p$ are i.i.d. with normal distribution $\mathcal{N}(0, \Sigma)$, with mean zero and covariance matrix Σ, and

$$\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, p\},\ y_i^j = F^j(X_i) + \varepsilon_i^j \ . \tag{3.1}$$

This means that, while the observations are independent, the outputs of the different tasks can be correlated, with correlation matrix Σ between the tasks. We now place ourselves in the fixed-design setting, that is, $(X_i)_{i=1}^n$ is deterministic and the goal is to estimate $\big(F^1(X_i), \dots, F^p(X_i)\big)_{i=1}^n$. Let us introduce some notation:

- $\mu_{\min} = \mu_{\min}(\Sigma)$ (resp. $\mu_{\max}$) denotes the smallest (resp. largest) eigenvalue of Σ.
- $c(\Sigma) := \mu_{\max}/\mu_{\min}$ is the condition number of Σ.

To obtain compact equations, we will use the following definition:

Definition 3.1. We denote by F the $n \times p$ matrix $(f^j(X_i))_{1\leq i\leq n,\, 1\leq j\leq p}$ and introduce the vector $f := \mathrm{vec}(F) = (f^1(X_1), \dots, f^1(X_n), \dots, f^p(X_1), \dots, f^p(X_n)) \in \mathbb{R}^{np}$, obtained by stacking the columns of F. Similarly we define $Y := (y_i^j) \in \mathcal{M}_{n\times p}(\mathbb{R})$, $y := \mathrm{vec}(Y)$, $E := (\varepsilon_i^j) \in \mathcal{M}_{n\times p}(\mathbb{R})$ and $\varepsilon := \mathrm{vec}(E)$.

In order to estimate f, we use a regularization procedure, which extends the classical ridge regression of the single-task setting. Let M be a $p \times p$ matrix, symmetric and positive-definite. Generalizing the work of Evgeniou et al. [EMP05], we estimate $(f^1, \dots, f^p) \in \mathcal{F}^p$ by

$$\widehat{F}_M \in \operatorname*{argmin}_{g\in\mathcal{F}^p} \left\{ \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (y_i^j - g^j(X_i))^2 + \sum_{j=1}^p \sum_{\ell=1}^p M_{j,l} \langle g^j, g^\ell \rangle_{\mathcal{F}} \right\} \tag{3.2}$$

and we denote by $\widehat{f}_M$ its fixed-design analogous. Although M could have a general unconstrained form we may restrict M to certain forms, for either computational or statistical reasons.

Remark 3.1. *Requiring that $M \succeq 0$ implies that Eq. (3.2) is a convex optimization problem, which can be solved through the resolution of a linear system, as explained later. Moreover it allows an RKHS interpretation, which will also be explained later.*

Example 3.1. *The case where the p tasks are treated independently can be considered in this setting: taking $M = M_{\mathrm{ind}}(\lambda) := \frac{1}{p}\,\mathrm{Diag}(\lambda_1, \dots, \lambda_p)$ for any $\lambda \in \mathbb{R}^p$ leads to the criterion*

$$\frac{1}{p} \sum_{j=1}^p \left[\frac{1}{n} \sum_{i=1}^n (y_i^j - g^j(X_i))^2 + \lambda_j \|g^j\|_{\mathcal{F}}^2 \right] \ , \tag{3.3}$$

that is, the sum of the single-task criteria described in Section 3.3. Hence, minimizing Eq. (3.3) over $\lambda \in \mathbb{R}^p$ amounts to solve independently *p single task problems.*

Example 3.2. *As done by Evgeniou et al. [EMP05], for every* $\lambda, \mu \in (0, +\infty)^2$*, define*

$$M_{\mathrm{SD}}(\lambda, \mu) := (\lambda + p\mu) I_p - \mu \mathbf{1}\mathbf{1}^\top = \begin{pmatrix} \lambda + (p-1)\mu & & -\mu \\ & \ddots & \\ -\mu & & \lambda + (p-1)\mu \end{pmatrix} .$$

Taking $M = M_{\mathrm{SD}}(\lambda, \mu)$ *in Eq.* (3.2) *leads to the criterion*

$$\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (y_i^j - g^j(X_i))^2 + \lambda \sum_{j=1}^{p} \left\| g^j \right\|_{\mathcal{F}}^2 + \frac{\mu}{2} \sum_{j=1}^{p} \sum_{k=1}^{p} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 . \tag{3.4}$$

Minimizing Eq. (3.4) *enforces a regularization on both the norms of the functions* g^j *and the norms of the differences* $g^j - g^k$*. Thus, matrices of the form* $M_{\mathrm{SD}}(\lambda, \mu)$ *are useful when the functions* g^j *are assumed to be similar in* $\mathcal{F}$*. One of the main contributions of the paper is to go beyond this case and learn from data a more general similarity matrix* M *between tasks.*

Example 3.3. *We extend Example 3.2 to the case where the* p *tasks consist of two groups of close tasks. Let* I *be a subset of* $\{1, \ldots, p\}$*, of cardinality* $1 \leq k \leq p-1$*. Let us denote by* I^c *the complementary of* I *in* $\{1, \ldots, p\}$*,* $\mathbf{1}_I$ *the vector* v *with components* $v_i = \mathbf{1}_{i \in I}$*, and* $\mathrm{Diag}(I)$ *the diagonal matrix* d *with components* $d_{i,i} = \mathbf{1}_{i \in I}$*. We then define*

$$M_I(\lambda, \mu, \nu) := \lambda I_p + \mu \, \mathrm{Diag}(I) + \nu \, \mathrm{Diag}(I^c) - \frac{\mu}{k} \mathbf{1}_I \mathbf{1}_I^\top - \frac{\nu}{p-k} \mathbf{1}_{I^c} \mathbf{1}_{I^c}^\top .$$

This matrix leads to the following criterion, which enforces a regularization on both the norms of the functions g^j *and the norms of the differences* $g^j - g^k$ *inside the groups* I *and* I^c*:*

$$\begin{aligned} \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (y_i^j - g^j(X_i))^2 + \lambda \sum_{j=1}^{p} \left\| g^j \right\|_{\mathcal{F}}^2 \\ + \frac{\mu}{2k} \sum_{j \in I} \sum_{k \in I} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 + \frac{\nu}{2(p-k)} \sum_{j \in I^c} \sum_{k \in I^c} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 . \end{aligned} \tag{3.5}$$

As shown in Section 3.6, we can estimate the set I *from data (see the work of Jacob et al. [JBV08] for a more general formulation).*

Remark 3.2. *Since* I_p *and* $\mathbf{1}\mathbf{1}^\top$ *can be diagonalized simultaneously, minimizing Eq.* (3.4) *and Eq.* (3.5) *is quite easy: it only demands optimization over two independent parameters, which can be done with the procedure of Arlot and Bach [AB11].*

Remark 3.3. *As stated below (Property 3.1),* M *acts as a scalar product between the tasks. Selecting a general matrix* M *is thus a way to express a similarity between tasks.*

Following Evgeniou et al. [EMP05], we define the vector-space $\mathcal{G}$ of real-valued functions over $\mathcal{X} \times \{1, \ldots, p\}$ by

$$\mathcal{G} := \left\{ g : \mathcal{X} \times \{1, \ldots, p\} \to \mathbb{R} \,/\, \forall j \in \{1, \ldots, p\} \ , \ g(\cdot, j) \in \mathcal{F} \right\} .$$

We now define a bilinear symmetric form over $\mathcal{G}$,

$$\forall g, h \in \mathcal{G} \ , \quad \langle g, h \rangle_{\mathcal{G}} := \sum_{j=1}^{p} \sum_{l=1}^{p} M_{j,l} \langle g(\cdot, j), h(\cdot, l) \rangle_{\mathcal{F}},$$

which is a scalar product as soon as M is positive semi-definite (see proof in Appendix 3.A) and leads to a RKHS (see proof in Appendix 3.B):

Property 3.1. *With the preceding notations $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ is a scalar product on $\mathcal{G}$.*

Corollary 3.1. *$(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ is a RKHS.*

In order to write down the kernel matrix in compact form, we introduce the following notations.

Definition 3.2 (Kronecker Product)**.** Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $B \in \mathcal{M}_{p,q}(\mathbb{R})$. We define the Kronecker product $A \otimes B$ as being the $(mp) \times (nq)$ matrix built with $p \times q$ blocks, the block of index (i, j) being $A_{i,j} \cdot B$:

$$A \otimes B = \begin{pmatrix} A_{1,1}B & \dots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \dots & A_{m,n}B \end{pmatrix} \ .$$

The Kronecker product is a widely used tool to deal with matrices and tensor products. Some of its classical properties are given in Section 3.E; see also Horn and Johnson [HJ91].

Property 3.2. *The kernel matrix associated with the design $\widetilde{X} := (X_i, j)_{i,j} \in \mathcal{X} \times \{1, \dots, p\}$ and the RKHS $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ is $\widetilde{K}_M := M^{-1} \otimes K$.*

Property 3.2 is proved in Appendix 3.C. We can then apply the representer's theorem [SS02] to the minimization problem (3.2) and deduce that $\widehat{f}_M = A_M y$ with

$$A_M = A_{M,K} := \widetilde{K}_M (\widetilde{K}_M + npI_{np})^{-1} = (M^{-1} \otimes K) \left((M^{-1} \otimes K) + npI_{np} \right)^{-1} \ .$$

3.2.2 Optimal Choice of the Kernel

Now when working in multi-task regression, a set $\mathcal{M} \subset \mathcal{S}_p^{++}(\mathbb{R})$ of matrices M is given, and the goal is to select the "best" one, that is, minimizing over M the quadratic risk $n^{-1} \| \widehat{f}_M - f \|_2^2$. For instance, the single-task framework corresponds to $p = 1$ and $\mathcal{M} = (0, +\infty)$. The multi-task case is far richer. The oracle risk is defined as

$$\inf_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - f \right\|_2^2 \right\} \ . \tag{3.6}$$

The ideal choice, called the oracle, is any matrix

$$M^\star \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - f \right\|_2^2 \right\} \ .$$

Nothing here ensures the oracle exists. However in some special cases (see for instance Example 3.4) the infimum of $\|\widehat{f}_M - f\|^2$ over the set $\{\widehat{f}_M,\ M \in \mathcal{M}\}$ may be attained by a function $f^* \in \mathcal{F}^p$—which we will call "oracle" by a slight abuse of notation—while the former problem does not have a solution.

From now on we always suppose that the infimum of $\{\|\widehat{f}_M - f\|^2\}$ over $\mathcal{M}$ is attained by some function $f^\star \in \mathcal{F}^p$. However the oracle $M^\star$ is not an estimator, since it depends on f.

Example 3.4 (Partial computation of the oracle in a simple setting)**.** *It is possible in certain simple settings to exactly compute the oracle (or, at least, some part of it). Consider for instance the set-up where the p functions are taken to be equal (that is, $f^1 = \cdots = f^p$). In this setting it is natural to use the set*

$$\mathcal{M}_{\mathrm{SD}} := \left\{ M_{\mathrm{SD}}\left(\lambda, \mu\right) = (\lambda + p\mu) I_p - \frac{\mu}{p} \mathbf{1}\mathbf{1}^\top \,/\, (\lambda, \mu) \in (0, +\infty)^2 \right\} \ .$$

Using the estimator $\widehat{f}_M = A_M y$ we can then compute the quadratic risk using the bias-variance decomposition given in Equation (3.33)*:*

$$\mathbb{E}\left[\left\| \widehat{f}_M - f \right\|_2^2 \right] = \|(A_M - I_{np}) f\|_2^2 + \mathrm{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ .$$

Computations (*reported in Appendix 3.D*) *show that, with the change of variables $\widetilde{\mu} = \lambda + p\mu$, the bias does not depend on $\widetilde{\mu}$ and the variance is a decreasing function of $\widetilde{\mu}$. Thus the oracle is obtained when $\widetilde{\mu} = +\infty$, leading to a situation where the oracle functions $f^{1,\star}, \ldots, f^{p,\star}$ verify $f^{1,\star} = \cdots = f^{p,\star}$. It is also noticeable that, if one assumes the maximal eigenvalue of Σ stays bounded with respect to p, the variance is of order $\mathcal{O}(p^{-1})$ while the bias is bounded with respect to p.*

As explained by Arlot and Bach [AB11], we choose

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \{\mathrm{crit}(M)\} \quad \text{with} \quad \mathrm{crit}(M) = \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \mathrm{pen}(M) \ ,$$

where the penalty term $\mathrm{pen}(M)$ has to be chosen appropriately.

Remark 3.4. *Our model* (3.1) *does not constrain the functions $f^1, \ldots, f^p$. Our way to express the similarities between the tasks (that is, between the f^j) is via the set $\mathcal{M}$, which represents the a priori knowledge the statistician has about the problem. Our goal is to build an estimator whose risk is the closest possible to the oracle risk. Of course using an inappropriate set $\mathcal{M}$ (with respect to the target functions $f^1, \ldots, f^p$) may lead to bad overall performances. Explicit multi-task settings are given in Examples 3.1, 3.2 and 3.3 and through simulations in Section 3.6.*

fixed

The unbiased risk estimation principle [Aka70, introduced by] requires

$$\mathbb{E}\left[\mathrm{crit}(M)\right] \approx \mathbb{E}\left[\frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right] \ ,$$

which leads to the (deterministic) *ideal penalty*

$$\mathrm{pen}_{\mathrm{id}}(M) := \mathbb{E}\left[\frac{1}{np}\|\widehat{f}_M - f\|_2^2\right] - \mathbb{E}\left[\frac{1}{np}\left\|y - \widehat{f}_M\right\|_2^2\right] \ .$$

Since $\widehat{f}_M = A_M y$ and $y = f + \varepsilon$, we can write

$$\left\|\widehat{f}_M - y\right\|_2^2 = \left\|\widehat{f}_M - f\right\|_2^2 + \|\varepsilon\|_2^2 - 2\langle\varepsilon, A_M\varepsilon\rangle + 2\langle\varepsilon, (I_{np} - A_M)f\rangle \ .$$

Since ε is centered and M is deterministic, we get, up to an additive factor independent of M,

$$\mathrm{pen}_{\mathrm{id}}(M) = \frac{2\mathbb{E}\left[\langle\varepsilon, A_M\varepsilon\rangle\right]}{np} \ ,$$

that is, as the covariance matrix of ε is $\Sigma \otimes I_n$,

$$\mathrm{pen}_{\mathrm{id}}(M) = \frac{2\,\mathrm{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} \ . \tag{3.7}$$

In order to approach this penalty as precisely as possible, we have to sharply estimate Σ. In the single-task case, such a problem reduces to estimating the variance σ^2 of the noise and was tackled by Arlot and Bach [AB11]. Since our approach for estimating Σ heavily relies on these results, they are summarized in the next section.

Note that estimating Σ is a mean towards estimating M. The technique we develop later for this purpose is not purely a multi-task technique, and may also be used in a different context.

3.3 Single Task Framework: Estimating a Single Variance

This section recalls some of the main results from Arlot and Bach [AB11] which can be considered as solving a special case of Section 3.2, with $p = 1$, $\Sigma = \sigma^2 > 0$ and $\mathcal{M} = [0, +\infty]$. Writing $M = \lambda$ with $\lambda \in [0, +\infty]$, the regularization matrix is

$$\forall \lambda \in (0, +\infty) \,, \quad A_\lambda = A_{\lambda,K} = K(K + n\lambda I_n)^{-1} \ ,$$

$A_0 = I_n$ and $A_{+\infty} = 0$; the ideal penalty becomes

$$\mathrm{pen}_{\mathrm{id}}(\lambda) = \frac{2\sigma^2\,\mathrm{tr}(A_\lambda)}{n} \ .$$

By analogy with the case where A_λ is an orthogonal projection matrix, $\mathrm{df}(\lambda) := \mathrm{tr}(A_\lambda)$ is called the effective degree of freedom, first introduced by Mallows [Mal73]; see also the work by Zhang [Zha05]. The ideal penalty however depends on σ^2; in order to have a fully data-driven penalty we have to replace σ^2 by an estimator $\widehat{\sigma}^2$ inside $\mathrm{pen}_{\mathrm{id}}(\lambda)$. For every $\lambda \in [0, +\infty]$, define

$$\mathrm{pen}_{\mathrm{min}}(\lambda) = \mathrm{pen}_{\mathrm{min}}(\lambda, K) := \frac{(2\,\mathrm{tr}(A_{\lambda,K}) - \mathrm{tr}(A_{\lambda,K}^\top A_{\lambda,K}))}{n} \ .$$

We shall see now that it is a *minimal penalty* in the following sense. If for every $C > 0$

$$\widehat{\lambda}_0(C) \in \operatorname*{argmin}_{\lambda \in [0,+\infty]} \left\{ \frac{1}{n} \|A_{\lambda,K} Y - Y\|_2^2 + C \operatorname{pen}_{\min}(\lambda, K) \right\} \ ,$$

then—up to concentration inequalities—$\widehat{\lambda}_0(C)$ acts as a mimimizer of

$$g_C(\lambda) = \mathbb{E}\left[\frac{1}{n} \|A_\lambda Y - Y\|_2^2 + C \operatorname{pen}_{\min}(\lambda) \right] - \sigma^2 = \frac{1}{n} \|(A_\lambda - I_n) f\|_2^2 + (C - \sigma^2) \operatorname{pen}_{\min}(\lambda) \ .$$

The former theoretical arguments show that

- if $C < \sigma^2$, $g_C(\lambda)$ decreases with $\mathrm{df}(\lambda)$ so that $\mathrm{df}(\widehat{\lambda}_0(C))$ is huge: the procedure overfits;
- if $C > \sigma^2$, $g_C(\lambda)$ increases with $\mathrm{df}(\lambda)$ when $\mathrm{df}(\lambda)$ is large enough so that $\mathrm{df}(\widehat{\lambda}_0(C))$ is much smaller than when $C < \sigma^2$.

The following algorithm was introduced by Arlot and Bach [AB11] and uses this fact to estimate σ^2.

Algorithm 3.1. **Input:** $Y \in \mathbb{R}^n$, $K \in \mathcal{S}_n^{++}(\mathbb{R})$

1. *For every $C > 0$, compute*

$$\widehat{\lambda}_0(C) \in \operatorname*{argmin}_{\lambda \in [0,+\infty]} \left\{ \frac{1}{n} \|A_{\lambda,K} Y - Y\|_2^2 + C \operatorname{pen}_{\min}(\lambda, K) \right\} \ .$$

2. **Output:** *$\widehat{C}$ such that* $\mathrm{df}(\widehat{\lambda}_0(\widehat{C})) \in [n/10, n/3]$.

An efficient algorithm for the first step of Algorithm 3.1 is detailed by Arlot and Massart [AM09], and we discuss the way we implemented Algorithm 3.1 in Section 3.6. The output $\widehat{C}$ of Algorithm 3.1 is a provably consistent estimator of σ^2, as stated in the following theorem.

Theorem 3.1 (Corollary of Theorem 1 of Arlot and Bach [AB11])**.** *Let $\beta = 150$. Suppose $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ with $\sigma^2 > 0$, and that $\lambda_0 \in (0, +\infty)$ and $d_n \geq 1$ exist such that*

$$\mathrm{df}(\lambda_0) \leq \sqrt{n} \text{ and } \frac{1}{n} \|(A_{\lambda_0} - I_n) F\|_2^2 \leq d_n \sigma^2 \sqrt{\frac{\ln n}{n}} \ . \tag{3.8}$$

Then for every $\delta \geq 2$, some constant $n_0(\delta)$ and an event Ω exist such that $\mathbb{P}(\Omega) \geq 1 - n^{-\delta}$ and if $n \geq n_0(\delta)$, on Ω,

$$\left(1 - \beta(2+\delta)\sqrt{\frac{\ln n}{n}}\right) \sigma^2 \leq \widehat{C} \leq \left(1 + \beta(2+\delta) d_n \sqrt{\frac{\ln(n)}{n}}\right) \sigma^2 \ . \tag{3.9}$$

Remark 3.5. *The values $n/10$ and $n/3$ in Algorithm 3.1 have no particular meaning and can be replaced by n/k, n/k', with $k > k' > 2$. Only β depends on k and k'. Also the bounds required in Assumption* (3.8) *only impact the right hand side of Equation* (3.9) *and are chosen to match the left hand side. See Property 10 of Arlot and Bach [AB11] for more details.*

3.4 Estimation of the Noise Covariance Matrix Σ

Thanks to the results developped by Arlot and Bach [AB11] (recapitulated in Section 3.3), we know how to estimate a variance for any one-dimensional problem. In order to estimate Σ, which has $p(p+1)/2$ parameters, we can use several one-dimensional problems. Projecting Y onto some direction $z \in \mathbb{R}^p$ yields

$$Y_z := Y \cdot z = F \cdot z + E \cdot z = F_z + \varepsilon_z \ , \tag{P_z}$$

with $\varepsilon_z \sim \mathcal{N}(0, \sigma_z^2 I_n)$ and $\sigma_z^2 := \mathrm{Var}[\varepsilon \cdot z] = z^\top \Sigma z$. Therefore, we will estimate σ_z^2 for $z \in \mathcal{Z}$ a well chosen set, and use these estimators to build back an estimation of Σ.

We now explain how to estimate Σ using those one-dimensional projections.

Definition 3.3. Let $a(z)$ be the output $\widehat{C}$ of Algorithm 3.1 applied to problem (P_z), that is, with inputs $Y_z \in \mathbb{R}^n$ and $K \in \mathcal{S}_n^{++}(\mathbb{R})$.

The idea is to apply Algorithm 3.1 to the elements z of a carefully chosen set $\mathcal{Z}$. Noting e_i the i-th vector of the canonical basis of $\mathbb{R}^p$, we introduce $\mathcal{Z} = \{e_i,\ i \in \{1, \dots, p\}\} \cup \{e_i + e_j,\ 1 \leq i < j \leq p\}$. We can see that $a(e_i)$ estimates $\Sigma_{i,i}$, while $a(e_i + e_j)$ estimates $\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}$. Henceforth, $\Sigma_{i,j}$ can be estimated by $(a(e_i + e_j) - a(e_i) - a(e_j))/2$. This leads to the definition of the following map J, which builds a symmetric matrix using the latter construction.

Definition 3.4. Let $J : \mathbb{R}^{\frac{p(p+1)}{2}} \to \mathcal{S}_p(\mathbb{R})$ be defined by

$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,i} = a_i \text{ if } 1 \leq i \leq p \ ,$$
$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,j} = \frac{a_{i,j} - a_i - a_j}{2} \text{ if } 1 \leq i < j \leq p \ .$$

This map is bijective, and for all $B \in \mathcal{S}_p(\mathbb{R})$

$$J^{-1}(B) = (B_{1,1}, \dots, B_{p,p}, B_{1,1} + B_{2,2} + 2B_{1,2}, \dots, B_{p-1,p-1} + B_{p,p} + 2B_{p-1,p}) \ .$$

This leads us to defining the following estimator of Σ:

$$\widehat{\Sigma} := J\left(a(e_1), \dots, a(e_p), a(e_1 + e_2), \dots, a(e_1 + e_p), \dots, a(e_{p-1} + e_p)\right) \ . \tag{3.10}$$

Remark 3.6. *If a diagonalization basis $(e'_1, \dots, e'_p)$ (whose basis matrix is P) of Σ is known, or if Σ is diagonal, then a simplified version of the algorithm defined by Eq.* (3.10) *is*

$$\widehat{\Sigma}_{\text{simplified}} = P^\top \operatorname{Diag}(a(e'_1), \dots, a(e'_p)) P \ . \tag{3.11}$$

This algorithm has a smaller computational cost and leads to better theoretical bounds (see Remark 3.10 and Section 3.5.2).

Let us recall that $\forall \lambda \in (0, +\infty)$, $A_\lambda = A_{\lambda,K} = K(K + n\lambda I_n)^{-1}$. Following Arlot and Bach [AB11] we make the following assumption from now on:

$$\left.\begin{array}{l} \forall j \in \{1, \dots, p\}\,, \exists \lambda_{0,j} \in (0, +\infty)\,, \\ \qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{and} \quad \dfrac{1}{n} \left\| (A_{\lambda_{0,j}} - I_n) F_{e_j} \right\|_2^2 \leq \Sigma_{j,j} \sqrt{\dfrac{\ln n}{n}} \end{array}\right\} \tag{\textbf{Hdf}}$$

We can now state the first main result of the paper.

3.4. ESTIMATION OF THE NOISE COVARIANCE MATRIX Σ

Theorem 3.2. *Let $\widehat{\Sigma}$ be defined by Eq.* (3.10)*, $\alpha = 2$ and assume* (**Hdf**) *holds. For every $\delta \geq 2$, a constant $n_0(\delta)$, an absolute constant $L_1 > 0$ and an event Ω exist such that $\mathbb{P}(\Omega) \geq 1 - p(p+1)/2 \times n^{-\delta}$ and if $n \geq n_0(\delta)$, on Ω,*

$$(1-\eta)\Sigma \preceq \widehat{\Sigma} \preceq (1+\eta)\Sigma \tag{3.12}$$

$$\textit{where} \qquad \eta := L_1(2+\delta)p\sqrt{\frac{\ln(n)}{n}}c(\Sigma)^2 \ .$$

Theorem 3.2 is proved in Section 3.E. It shows $\widehat{\Sigma}$ estimates Σ with a "multiplicative" error controlled with large probability, in a non-asymptotic setting. The multiplicative nature of the error is crucial for deriving the oracle inequality stated in Section 3.5, since it allows to show the ideal penalty defined in Equation (3.7) is precisely estimated when Σ is replaced by $\widehat{\Sigma}$.

An important feature of Theorem 3.2 is that it holds under very mild assumptions on the mean f of the data (see Remark 3.8). Therefore, it shows $\widehat{\Sigma}$ is able to estimate a covariance matrix *without prior knowledge on the regression function*, which, to the best of our knowledge, has never been obtained in multi-task regression.

Remark 3.7 (Scaling of (n,p) for consistency)**.** *A sufficient condition for ensuring $\widehat{\Sigma}$ is a consistent estimator of Σ is*

$$pc(\Sigma)^2\sqrt{\frac{\ln(n)}{n}} \longrightarrow 0 \ ,$$

which enforces a scaling between n, p and $c(\Sigma)$. Nevertheless, this condition is probably not necessary since the simulation experiments of Section 3.6 show that Σ can be well estimated (at least for estimator selection purposes) in a setting where $\eta \gg 1$.

Remark 3.8 (On assumption (**Hdf**))**.** *Assumption* (**Hdf**) *is a single-task assumption (made independently for each task). The upper bound $\sqrt{\ln(n)/n}$ can be multiplied by any factor $1 \leq d_n \ll \sqrt{n/\ln(n)}$ (as in Theorem 3.1), at the price of multiplying η by d_n in the upper bound of Eq.* (3.12)*. More generally the bounds on the degree of freedom and the bias in* (**Hdf**) *only influence the upper bound of Eq.* (3.12)*. The rates are chosen here to match the lower bound, see Property 10 of Arlot and Bach [AB11] for more details.*

Assumption (**Hdf**) *is rather classical in model selection, see Arlot and Bach [AB11] for instance. In particular, (a weakened version of)* (**Hdf**) *holds if the bias $n^{-1}\|(A_\lambda - I_n)F_{e_i}\|_2^2$ is bounded by $C_1 \operatorname{tr}(A_\lambda)^{-C_2}$, for some $C_1, C_2 > 0$.*

Remark 3.9 (Choice of the set $\mathcal{Z}$)**.** *Other choices could have been made for $\mathcal{Z}$, however ours seems easier in terms of computation, since $|\mathcal{Z}| = p(p+1)/2$. Choosing a larger set $\mathcal{Z}$ leads to theoretical difficulties in the reconstruction of $\widehat{\Sigma}$, while taking other basis vectors leads to more complex computations. We can also note that increasing $|\mathcal{Z}|$ decreases the probability in Theorem 3.2, since it comes from an union bound over the one-dimensional estimations.*

Remark 3.10. *When $\widehat{\Sigma} = \widehat{\Sigma}_{\text{simplified}}$ as defined by Eq.* (3.11)*, that is, when a diagonalization basis of Σ is known, Theorem 3.2 still holds on a set of larger probability $1 - \kappa p n^{-\delta}$ with a reduced error $\eta = L_1(\alpha+\delta)\sqrt{\ln(n)/n}$. Then, a consistent estimation of Σ is possible whenever $p = O(n^\delta)$ for some $\delta \geq 0$.*

3.5 Oracle Inequality

This section aims at proving "oracle inequalities", as usually done in a model selection setting: given a set of models or of estimators, the goal is to upper bound the risk of the selected estimator by the oracle risk (defined by Eq. (3.6)), up to an additive term and a multiplicative factor. We show two oracle inequalities (Theorems 3.3 and 3.4) that correspond to two possible definitions of $\widehat{\Sigma}$.

Note that "oracle inequality" sometimes has a different meaning in the literature [LPTvdG11, see for instance] when the risk of the proposed estimator is controlled by the risk of an estimator using information coming from the true parameter (that is, available only if provided by an oracle).

3.5.1 A General Result for Discrete Matrix Sets $\mathcal{M}$

We first show that the estimator introduced in Eq. (3.10) is precise enough to derive an oracle inequality when plugged in the penalty defined in Eq. (3.7) in the case where $\mathcal{M}$ is finite.

Definition 3.5. Let $\widehat{\Sigma}$ be the estimator of Σ defined by Eq. (3.10). We define

$$\widehat{M} \in \operatorname*{argmin}_{M\in\mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2\operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\} \ .$$

We assume now the following holds true:

$$\exists (C, \alpha_{\mathcal{M}}) \in (0, +\infty)^2, \quad \operatorname{card}(\mathcal{M}) < C n^{\alpha_{\mathcal{M}}} \ . \tag{3.13}$$

Theorem 3.3. *Let $\alpha = \max(\alpha_{\mathcal{M}}, 2)$, $\delta \geq 2$ and assume* (**Hdf**) *and* (3.13) *hold true. Absolute constants $L_2, \kappa' > 0$, a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$ exist such that $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa' p(p+C) n^{-\delta}$ and the following holds as soon as $n \geq n_1(\delta)$. First, on $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M\in\mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 c(\Sigma)^4 \operatorname{tr}(\Sigma)(\alpha+\delta)^2 \frac{p^3 \ln(n)^3}{np} \ . \tag{3.14}$$

Second, an absolute constant L_3 exists such that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \right] \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \mathbb{E}\left[\inf_{M\in\mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} \right] \\ + L_2 c(\Sigma)^4 \operatorname{tr}(\Sigma)(\alpha+\delta)^2 \frac{p^3 \ln(n)^3}{np} + L_3 \frac{\sqrt{p(p+C)}}{n^{\delta/2}} \left(|\!|\!| \Sigma |\!|\!| + \frac{\|f\|_2^2}{np} \right) \ . \end{aligned} \tag{3.15}$$

Theorem 3.3 is proved in Section 3.F.

Remark 3.11. *If $\widehat{\Sigma} = \widehat{\Sigma}_{\text{simplified}}$ is defined by Eq.* (3.11) *the result still holds on a set of larger probability $1 - \kappa' p(1+C) n^{-\delta}$ with a reduced error, similar to the one in Theorem 3.4.*

3.5.2 A Result for a Continuous Set of Jointly Diagonalizable Matrices

We now show a similar result when matrices in $\mathcal{M}$ can be jointly diagonalized. It turns out a faster algorithm can be used instead of Eq. (3.10) with a reduced error and a larger probability event in the oracle inequality. Note that we no longer assume $\mathcal{M}$ is finite, so it can be parametrized by continuous parameters.

Suppose now the following holds, which means the matrices of $\mathcal{M}$ are jointly diagonalizable:

$$\exists P \in O_p(\mathbb{R})\,, \quad \mathcal{M} \subseteq \left\{ P^\top \operatorname{Diag}(d_1, \dots, d_p) P\,,\, (d_i)_{i=1}^p \in (0, +\infty)^p \right\} \quad . \tag{HM}$$

Let P be the matrix defined in Assumption (**HM**), $\widetilde{\Sigma} = P \Sigma P^\top$ and recall that $A_\lambda = K(K + n\lambda I_n)^{-1}$. Computations detailed in Appendix 3.D show that the ideal penalty introduced in Eq. (3.7) can be written as

$$\forall M = P^\top \operatorname{Diag}(d_1, \dots, d_p) P \in \mathcal{M},$$
$$\operatorname{pen}_{\mathrm{id}}(M) = \frac{2 \operatorname{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} = \frac{2}{np} \left(\sum_{j=1}^p \operatorname{tr}(A_{pd_j}) \widetilde{\Sigma}_{j,j} \right) \quad . \tag{3.16}$$

Eq. (3.16) shows that under Assumption (**HM**), we do not need to estimate the entire matrix Σ in order to have a good penalization procedure, but only to estimate the variance of the noise in p directions.

Definition 3.6. Let $(e_1, \dots, e_p)$ be the canonical basis of $\mathbb{R}^p$, $(u_1, \dots, u_p)$ be the orthogonal basis defined by $\forall j \in \{1, \dots, p\},\ u_j = P^\top e_j$. We then define

$$\widehat{\Sigma}_{\mathrm{HM}} = P \operatorname{Diag}(a(u_1), \dots, a(u_p)) P^\top \quad ,$$

where for every $j \in \{1, \dots, p\}$, $a(u_j)$ denotes the output of Algorithm 3.1 applied to Problem (**Pu$_j$**), and

$$\widehat{M}_{\mathrm{HM}} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2 \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma}_{\mathrm{HM}} \otimes I_n) \right) \right\} \quad . \tag{3.17}$$

Theorem 3.4. *Let $\alpha = 2$, $\delta \geq 2$ and assume (**Hdf**) and (**HM**) hold true. Absolute constants $L_2 > 0$, and κ'', a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$ exist such that $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa'' p n^{-\delta}$ and the following holds as soon as $n \geq n_1(\delta)$. First, on $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_{HM}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 \operatorname{tr}(\Sigma)(2+\delta)^2 \frac{\ln(n)^3}{n} \quad . \tag{3.18}$$

Second, an absolute constant L_4 exists such that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_{HM}} - f \right\|_2^2 \right] \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \mathbb{E}\left[\inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} \right] \\ + L_4 \operatorname{tr}(\Sigma)(2+\delta)^2 \frac{\ln(n)^3}{n} + \frac{p}{n^{\delta/2}} \frac{\|f\|_2^2}{np} \quad . \end{aligned} \tag{3.19}$$

Theorem 3.4 is proved in Section 3.F.

3.5.3 Comments on Theorems 3.3 and 3.4

Remark 3.12. *Taking $p = 1$ (hence $c(\Sigma) = 1$ and $\operatorname{tr}(\Sigma) = \sigma^2$), we recover Theorem 3 of Arlot and Bach [AB11] as a corollary of Theorem 3.3.*

Remark 3.13 (Scaling of (n, p))**.** *When assumption* (3.13) *holds, Eq.* (3.14) *implies the asymptotic optimality of the estimator $\widehat{f}_{\widehat{M}}$ when*

$$c(\Sigma)^4 \frac{\operatorname{tr}\Sigma}{p} \times \frac{p^3 \left(\ln(n)\right)^3}{n} \ll \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} \quad .$$

In particular, only (n, p) such that $p^3 \ll n/(\ln(n))^3$ are admissible. When assumption (**HM**) *holds, the scalings required to ensure optimality in Eq.* (3.18) *are more favorable:*

$$\operatorname{tr}\Sigma \times \frac{\left(\ln(n)\right)^3}{n} \ll \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} \quad .$$

It is to be noted that p still influences the left hand side via $\operatorname{tr}\Sigma$.

Remark 3.14. *Theorems 3.3 and 3.4 are non asymptotic oracle inequalities, with a multiplicative term of the form $1 + o(1)$. This allows us to claim that our selection procedure is nearly optimal, since our estimator is close (with regard to the empirical quadratic norm) to the oracle one. Furthermore the term $1+(\ln(n))^{-1}$ in front of the infima in Equations* (3.14), (3.18), (3.15) *and* (3.19) *can be further diminished, but this yields a greater remainder term as a consequence.*

Remark 3.15 (On assumption (**HM**))**.** *Assumption* (**HM**) *actually means all matrices in $\mathcal{M}$ can be diagonalized in a unique orthogonal basis, and thus can be parametrized by their eigenvalues as in Examples 3.1, 3.2 and 3.3.*

In that case the optimization problem is quite easy to solve, as detailed in Remark 3.18. If not, solving (3.17) *may turn out to be a hard problem, and our theoretical results do not cover this setting. However, it is always possible to discretize the set $\mathcal{M}$ or, in practice, to use gradient descent.*

Compared to the setting of Theorem 3.3, assumption (**HM**) *allows a simpler estimator for the penalty* (3.16)*, with an increased probability and a reduced error in the oracle inequality.*

The main theoretical limitation comes from the fact that the probabilistic concentration tools used apply to discrete sets $\mathcal{M}$ (through union bounds). The structure of kernel ridge regression allows us to have a uniform control over a continuous set for the single-task estimators at the "cost" of n pointwise controls, which can then be extended to the multi-task setting via (**HM**)*. We conjecture Theorem 3.4 still holds without* (**HM**) *as long as $\mathcal{M}$ is not "too large", which could be proved similarly up to some uniform concentration inequalities.*

Note also that if $\mathcal{M}_1, \ldots, \mathcal{M}_K$ all satisfy (**HM**) *(with different matrices P_k), then Theorem 3.4 still holds for $\mathcal{M} = \bigcup_{k=1}^K \mathcal{M}_k$ with the penalty defined by Eq.* (3.17) *with $P = P_k$ when $M \in \mathcal{M}_k$, and $\mathbb{P}(\widetilde{\Omega}) \geq 1 - 9Kp^2n^{-\delta}$, by applying the union bound in the proof.*

Remark 3.16 (Relationship with the trace norm)**.** *Our approach relies on the minimization of Equation* (3.2) *with respect to f. Argyriou et al. [AEP08] have shown that if we also*

minimize Equation (3.2) *with respect to the matrix* M *subject to the constraint* $\operatorname{tr} M^{-1} = 1$*, then we obtain an equivalent regularization by the nuclear norm (a.k.a. trace norm), which implies the prior knowledge that our* p *prediction functions may be obtained as the linear combination of* $r \ll p$ *basis functions. This situation corresponds to cases where the matrix* M^{-1} *is singular.*

Note that the link between our framework and trace norm (i.e., nuclear norm) regularization is the same than between multiple kernel learning and the single task framework of Arlot and Bach [AB11]. In the multi-task case, the trace-norm regularization, though efficient computationally, does not lead to an oracle inequality, while our criterion is an unbiased estimate of the generalization error, which turns out to be non-convex in the matrix M*. While DC programming techniques [GRC09, and references therin] could be brought to bear to find local optima, the goal of the present work is to study the theoretical properties of our estimators, assuming we can minimize the cost function (e.g., in special cases, where we consider spectral variants, or by brute force enumeration).*

3.6 Simulation Experiments

In all the experiments presented in this section, we consider the framework of Section 3.2 with $\mathcal{X} = \mathbb{R}^d$, $d = 4$, and the kernel defined by $\forall x, y \in \mathcal{X}$, $k(x, y) = \prod_{j=1}^{d} e^{-|x_j - y_j|}$. The design points $X_1, \dots, X_n \in \mathbb{R}^d$ are drawn (repeatedly and independently for each sample) independently from the multivariate standard Gaussian distribution. For every $j \in \{1, \dots, p\}$, $f^j(\cdot) = \sum_{i=1}^{m} \alpha_i^j k(\cdot, z_i)$ where $m = 4$ and $z_1, \dots, z_m \in \mathbb{R}^d$ are drawn (once for all experiments except in Experiment D) independently from the multivariate standard Gaussian distribution, independently from the design $(X_i)_{1 \leq i \leq n}$. Thus, the expectations that will be considered are taken conditionally to the z_i. The coefficients $(\alpha_i^j)_{1 \leq i \leq m\,, 1 \leq j \leq p}$ differ according to the setting. Matlab code is available online. [1]

3.6.1 Experiments

Five experimental settings are considered:

A⌋ Various numbers of tasks: $n = 10$ and $\forall i, j$, $\alpha_i^j = 1$, that is, $\forall j$, $f^j = f_A := \sum_{i=1}^{m} k(\cdot, z_i)$. The number of tasks is varying: $p \in \{2k \,/\, k = 1, \dots, 25\}$. The covariance matrix is $\Sigma = 10 \cdot I_p$.

B⌋ Various sample sizes: $p = 5$, $\forall j$, $f^j = f_A$ and $\Sigma = \Sigma_B$ has been drawn (once for all) from the Whishart $W(I_5, 10, 5)$ distribution; the condition number of Σ_B is $c(\Sigma_B) \approx 22.05$. The only varying parameter is $n \in \{50k \,/\, k = 1, \dots, 20\}$.

C⌋ Various noise levels: $n = 100$, $p = 5$ and $\forall j$, $f^j = f_A$. The varying parameter is $\Sigma = \Sigma_{C,t} := 5t \cdot I_5$ with $t \in \{0.2k \,/\, k = 1, \dots, 50\}$. We also ran the experiments for $t = 0.01$ and $t = 100$.

D⌋ Clustering of two groups of functions: $p = 10$, $n = 100$, $\Sigma = \Sigma_E$ has been drawn (once for all) from the Whishart $W(I_{10}, 20, 10)$ distribution; the condition number of Σ_E is $c(\Sigma_E) \approx 24.95$. We pick the function $f_D := \sum_{i=1}^{m} \alpha_i k(\cdot, z_i)$ by drawing

1. Matlab code can be found at `http://www.di.ens.fr/~solnon/multitask_minpen_en.html`.

$(\alpha_1, \ldots, \alpha_m)$ and $(z_1, \ldots, z_m)$ from standard multivariate normal distribution (independently in each replication) and finally $f^1 = \cdots = f^5 = f_D$, $f^6 = \cdots = f^{10} = -f_D$.

E⌋ Comparison to cross-validation parameter selection: $p = 5$, $\Sigma = 10 \cdot I_5$, $\forall j$, $f^j = f_A$. The sample size is taken in $\{10, 50, 100, 250\}$.

3.6.2 Collections of Matrices

Two different sets of matrices $\mathcal{M}$ are considered in the Experiments A–C, following Examples 3.1 and 3.2:

$$\mathcal{M}_{\text{SD}} := \left\{ M_{\text{SD}}(\lambda, \mu) = (\lambda + p\mu) I_p - \frac{\mu}{p} \mathbf{1}\mathbf{1}^\top \,/\, (\lambda, \mu) \in (0, +\infty)^2 \right\}$$
$$\text{and} \quad \mathcal{M}_{\text{ind}} := \{ M_{\text{ind}}(\lambda) = \text{Diag}(\lambda_1, \ldots, \lambda_p) \,/\, \lambda \in (0, +\infty)^p \} \quad .$$

In Experiment D, we also use two different sets of matrices, following Example 3.3:

$$\mathcal{M}_{\text{clus}} := \bigcup_{I \subset \{1,\ldots,p\}, I \notin \{\{1,\ldots,p\}, \emptyset\}} \left\{ M_I(\lambda, \mu, \mu) \,/\, (\lambda, \mu) \in (0, +\infty)^2 \right\} \cup \mathcal{M}_{\text{SD}}$$
$$\text{and} \quad \mathcal{M}_{\text{interval}} := \bigcup_{1 \leq k \leq p-1} \left\{ M_I(\lambda, \mu, \mu) \,/\, (\lambda, \mu) \in (0, +\infty)^2, I = \{1, \ldots, k\} \right\} \cup \mathcal{M}_{\text{SD}} \quad .$$

Remark 3.17. *The set $\mathcal{M}_{\text{clus}}$ contains $2^p - 1$ models, a case we will denote by "clustering". The other set, $\mathcal{M}_{\text{interval}}$, only has p models, and is adapted to the structure of the Experiment D. We call this setting "segmentation into intervals".*

3.6.3 Estimators

In Experiments A–C, we consider four estimators obtained by combining two collections $\mathcal{M}$ of matrices with two formulas for Σ which are plugged into the penalty (3.7) (that is, either Σ known or estimated by $\widehat{\Sigma}$):

$$\forall \alpha \in \{\text{SD}, \text{ind}\} \;, \forall S \in \left\{ \Sigma, \widehat{\Sigma}_{\text{HM}} \right\}, \quad \widehat{f}_{\alpha,S} := \widehat{f}_{\widehat{M}_{\alpha,S}} = A_{\widehat{M}_{\alpha,S}} y$$
$$\text{where} \quad \widehat{M}_{\alpha,S} \in \operatorname*{argmin}_{M \in \mathcal{M}_\alpha} \left\{ \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (S \otimes I_n) \right) \right\}$$

and $\widehat{\Sigma}_{\text{HM}}$ is defined in Section 3.5.2. As detailed in Examples 3.1–3.2, $\widehat{f}_{\text{ind},\widehat{\Sigma}_{\text{HM}}}$ and $\widehat{f}_{\text{ind},\Sigma}$ are concatenations of single-task estimators, whereas $\widehat{f}_{\text{SD},\widehat{\Sigma}_{\text{HM}}}$ and $\widehat{f}_{\text{SD},\Sigma}$ should take advantage of a setting where the functions f^j are close in $\mathcal{F}$ thanks to the regularization term $\sum_{j,k} \|f^j - f^k\|_{\mathcal{F}}^2$. In Experiment D we consider the following three estimators, that depend on the choice of the collection $\mathcal{M}$:

$$\forall \beta \in \{\text{clus}, \text{interval}, \text{ind}\} \;, \quad \widehat{f}_\beta := \widehat{f}_{\widehat{M}_\beta} = A_{\widehat{M}_\beta} y$$
$$\text{where} \quad \widehat{M}_\beta \in \operatorname*{argmin}_{M \in \mathcal{M}_\beta} \left\{ \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\}$$

and $\widehat{\Sigma}$ is defined by Equation (3.10).

In Experiment E we consider the estimator $\widehat{f}_{\mathrm{SD},\widehat{\Sigma}_{\mathrm{HM}}}$. As explained in the following remark the parameters of the former estimator are chosen by optimizing (3.17), in practice by choosing a grid. We also consider the estimator $\widehat{f}_{\mathrm{SD,CV}}$ where the parameters are selected by performing 5-fold cross-validation on the mentionned grid.

Remark 3.18 (Optimization of (3.17))**.** *Thanks to Assumption* (**HM**) *the optimization problem* (3.17) *can be solved easily. It suffices to diagonalize in a common basis the elements of* $\mathcal{M}$ *and the problem splits into several multi-task problems, each with one real parameter. The optimization was then done by using a grid on the real parameters, chosen such that the degree of freedom takes all integer values from* 0 *to* n*.*

Remark 3.19 (Finding the jump in Algorithm 3.1)**.** *Algorithm 3.1 raises the question of how to detect the jump of* $\mathrm{df}(\lambda)$*, which happens around* $C = \sigma^2$*. We chose to select an estimator* $\widehat{C}$ *of* σ^2 *corresponding to the smallest index such that* $\mathrm{df}(\widehat{\lambda}_0(\widehat{C})) < n/2$*. Another approach is to choose the index corresponding to the largest instantaneous jump of* $\mathrm{df}(\widehat{\lambda}_0(C))$ (*which is piece-wise constant and non-increasing*)*. This approach has a major drawback, because it sometimes selects a jump far away from the "real" jump around* σ^2*, when the real jump consists of several small jumps. Both approaches gave similar results in terms of prediction error, and we chose the first one because of its direct link to the theoretical criterion given in Theorem 3.1.*

3.6.4 Results

In each experiment, $N = 1000$ independent samples $y \in \mathbb{R}^{np}$ have been generated. Expectations are estimated thanks to empirical means over the N samples. Error bars correspond to the classical Gaussian 95% confidence interval (that is, empirical standard-deviation over the N samples multiplied by $1.96/\sqrt{N}$). The results of Experiments A–C are reported in Figures 3.2–3.8. The results of Experiments C–E are reported in Tables 3.1–3.3. The p-values correspond to the classical Gaussian difference test, where the hypotheses tested are of the shape $\mathbb{H}_0 = \{q > 1\}$ against the hypotheses $\mathbb{H}_1 = \{q \leq 1\}$, where the different quantities q are detailed in Tables 3.2–3.3.

t	0.01	100
$\mathbb{E}[\|\widehat{f}_{\mathrm{SD},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$	1.80 ± 0.02	0.300 ± 0.003
$\mathbb{E}[\|\widehat{f}_{\mathrm{SD},\widehat{\Sigma}} - f\|^2]$	$(2.27 \pm 0.38) \times 10^{-2}$	0.357 ± 0.048
$\mathbb{E}[\|\widehat{f}_{\mathrm{SD},\Sigma} - f\|^2]$	$(1.20 \pm 0.28) \times 10^{-2}$	0.823 ± 0.080
$\mathbb{E}[\|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$	$(1.26 \pm 0.26) \times 10^{-2}$	1.51 ± 0.07
$\mathbb{E}[\|\widehat{f}_{\mathrm{ind},\Sigma} - f\|^2]$	$(1.20 \pm 0.24) \times 10^{-2}$	4.47 ± 0.13

Table 3.1: Results of Experiment C for the extreme values of t.

3.6.5 Comments

As expected, multi-task learning significantly helps when all f^j are equal, as soon as p is large enough (Figure 3.1), especially for small n (Figure 3.6) and large noise-levels

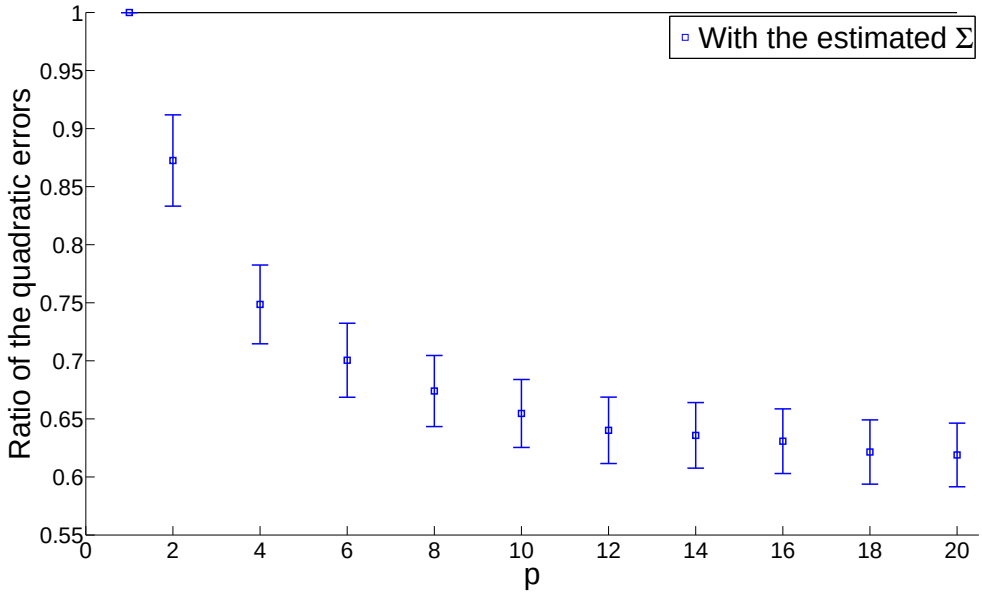

Figure 3.1: Increasing the number of tasks p (Experiment A), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\mathrm{SD},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$.

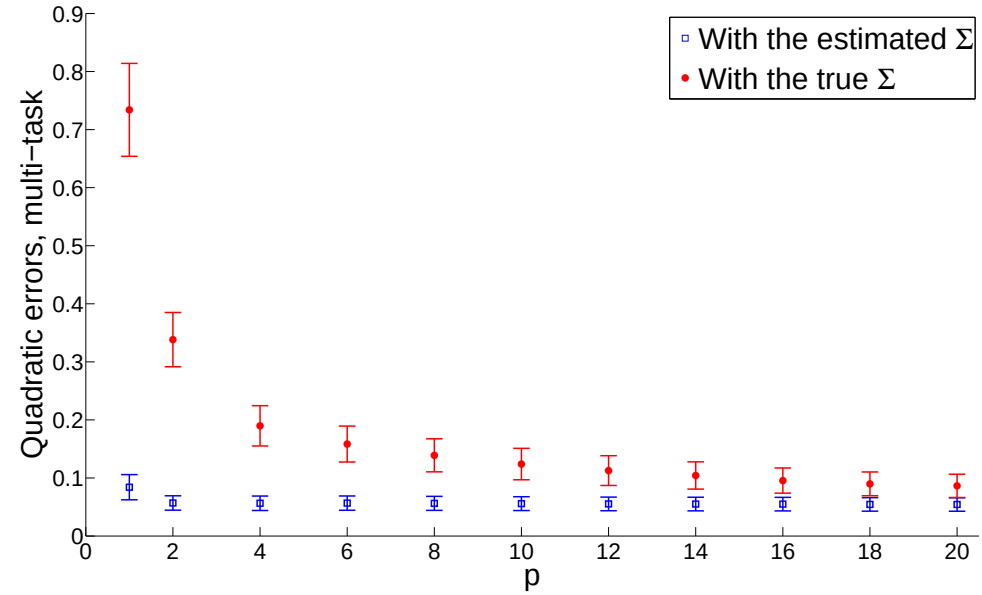

Figure 3.2: Increasing the number of tasks p (Experiment A), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{SD},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

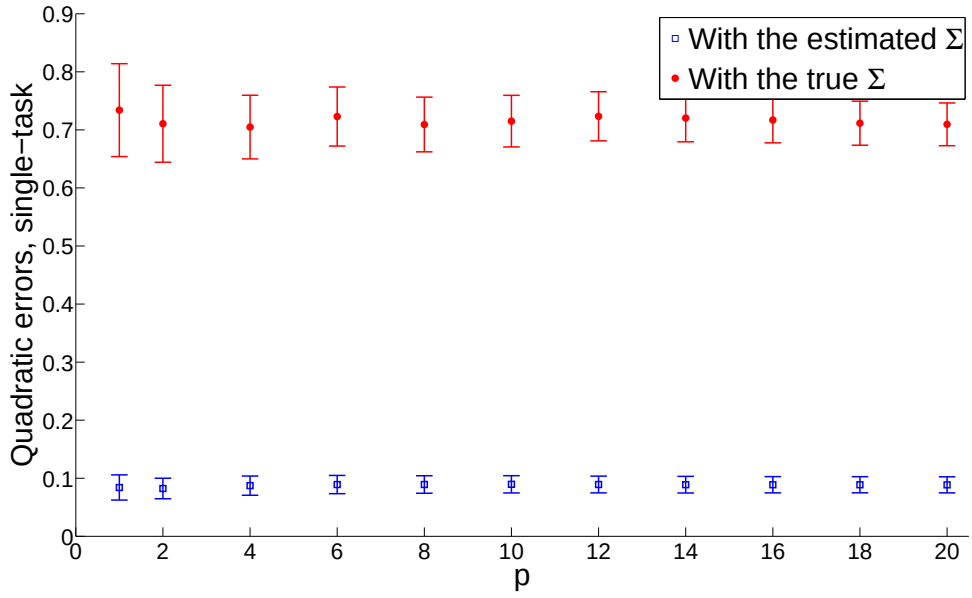

Figure 3.3: Increasing the number of tasks p (Experiment A), quadratic errors of single-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{ind},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

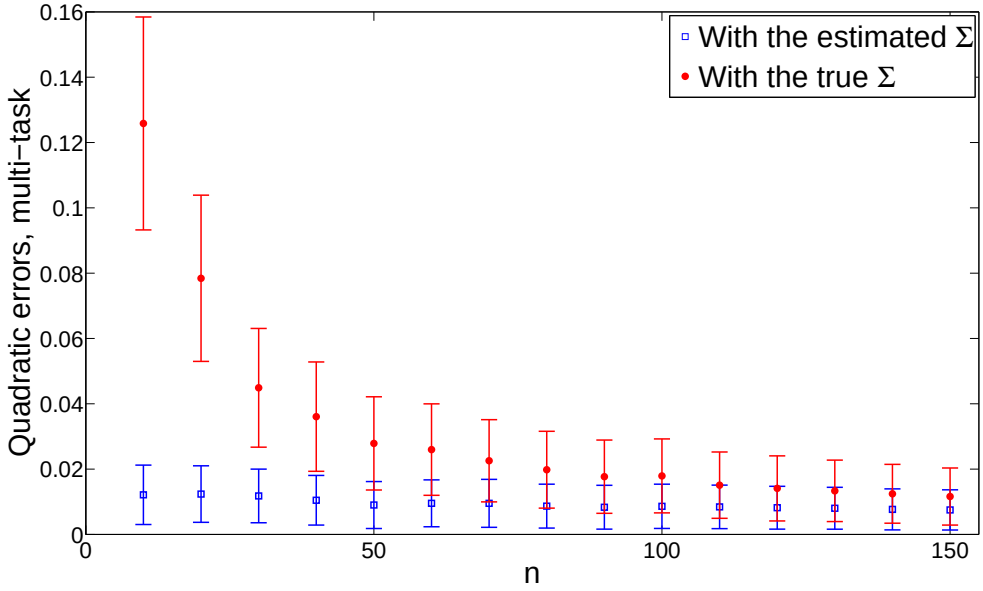

Figure 3.4: Increasing the sample size n (Experiment B), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{SD},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

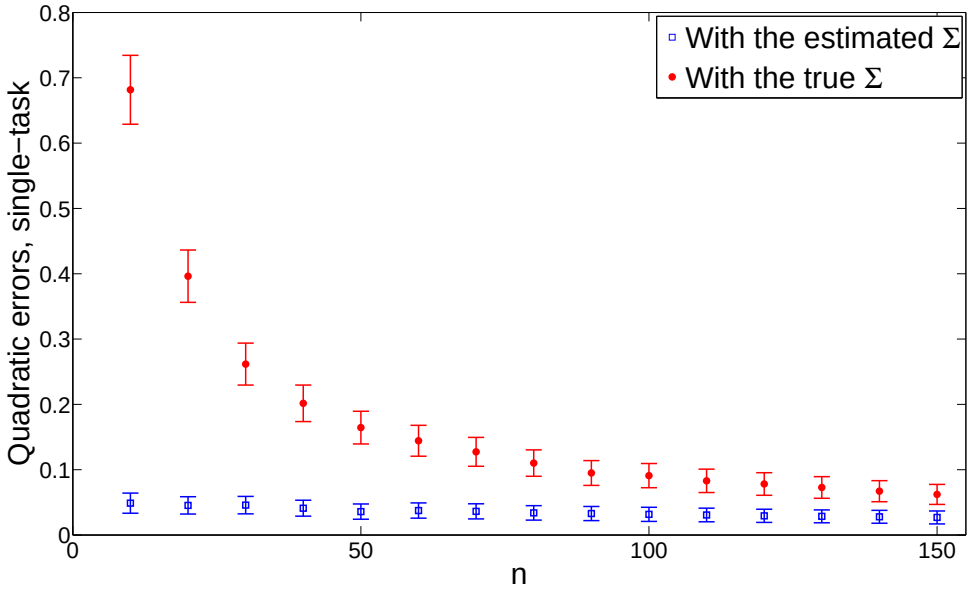

Figure 3.5: Increasing the sample size n (Experiment B), quadratic errors of single-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{ind},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

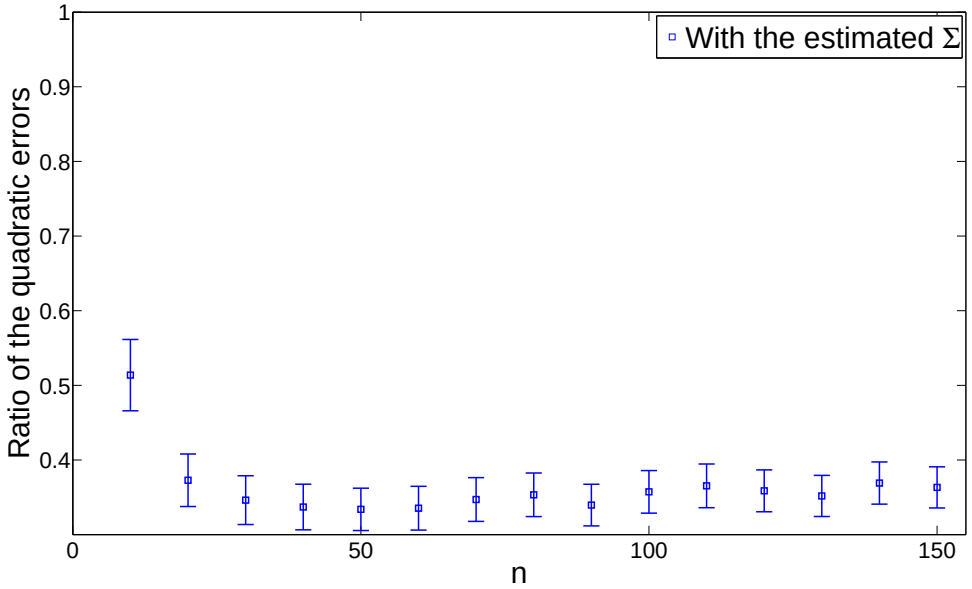

Figure 3.6: Increasing the sample size n (Experiment B), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\mathrm{SD},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$.

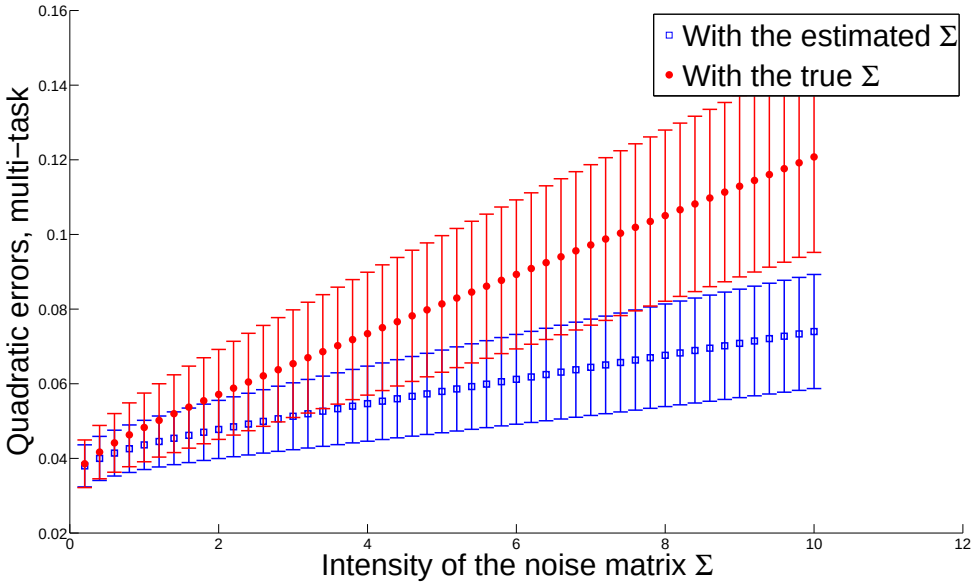

Figure 3.7: Increasing the signal-to-noise ratio (Experiment C), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{SD},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

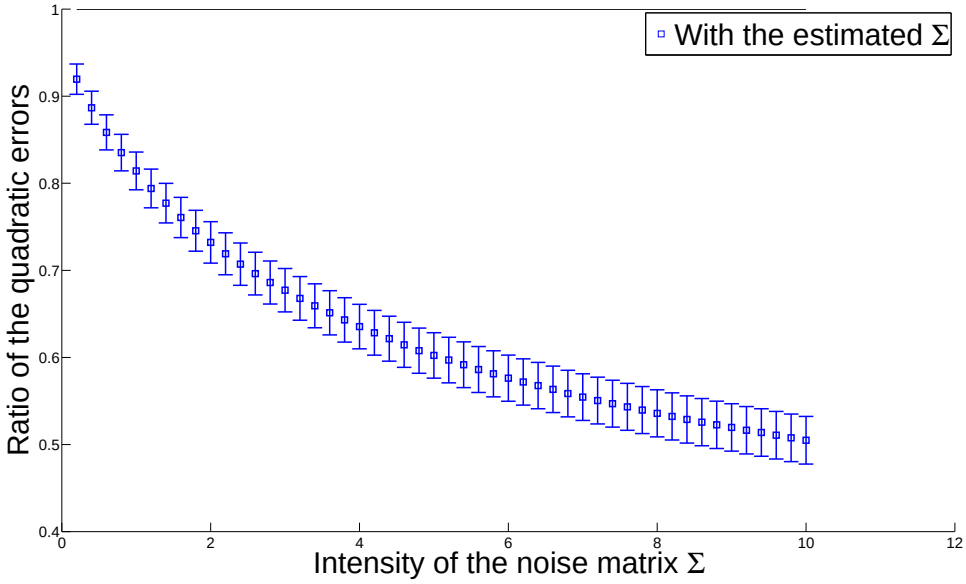

Figure 3.8: Increasing the signal-to-noise ratio (Experiment C), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\mathrm{SD},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$.

q	$\mathbb{E}[q]$	Std$[q]$	p-value for $\mathbb{H}_0 = \{q > 1\}$
$\|\widehat{f}_{\text{clus}} - f\|^2 / \|\widehat{f}_{\text{ind}} - f\|^2$	0.668	0.294	$< 10^{-15}$
$\|\widehat{f}_{\text{interval}} - f\|^2 / \|\widehat{f}_{\text{ind}} - f\|^2$	0.660	0.270	$< 10^{-15}$
$\|\widehat{f}_{\text{interval}} - f\|^2 / \|\widehat{f}_{\text{clus}} - f\|^2$	1.00	0.165	0.50

Table 3.2: Clustering and segmentation (Experiment D).

q	n	$\mathbb{E}[q]$	Std$[q]$	p-value for $\mathbb{H}_0 = \{q > 1\}$
$\|\widehat{f}_{\text{SD},\widehat{\Sigma}_{\text{HM}}} - f\|^2 / \|\widehat{f}_{\text{SD,CV}} - f\|^2$	10	0.35	0.46	$< 10^{-15}$
$\|\widehat{f}_{\text{SD},\widehat{\Sigma}_{\text{HM}}} - f\|^2 / \|\widehat{f}_{\text{SD,CV}} - f\|^2$	50	0.56	0.42	$< 10^{-15}$
$\|\widehat{f}_{\text{SD},\widehat{\Sigma}_{\text{HM}}} - f\|^2 / \|\widehat{f}_{\text{SD,CV}} - f\|^2$	100	0.71	0.34	$< 10^{-15}$
$\|\widehat{f}_{\text{SD},\widehat{\Sigma}_{\text{HM}}} - f\|^2 / \|\widehat{f}_{\text{SD,CV}} - f\|^2$	250	0.87	0.19	$< 10^{-15}$

Table 3.3: Comparison of our method to 5-fold cross-validation (Experiment E).

(Figure 3.8 and Table 3.1). Increasing the number of tasks rapidly reduces the quadratic error with multi-task estimators (Figure 3.2) contrary to what happens with single-task estimators (Figure 3.3).

A noticeable phenomenon also occurs in Figure 3.2 and even more in Figure 3.3: the estimator $\widehat{f}_{\text{ind},\Sigma}$ (that is, obtained knowing the true covariance matrix Σ) is less efficient than $\widehat{f}_{\text{ind},\widehat{\Sigma}}$ where the covariance matrix is estimated. It corresponds to the combination of two facts: (i) multiplying the ideal penalty by a small factor $1 < C_n < 1 + o(1)$ is known to often improve performances in practice when the sample size is small [Arl09, see Section 6.3.2 of], and (ii) minimal penalty algorithms like Algorithm 3.1 are conjectured to overpenalize slightly when n is small or the noise-level is large [Ler11] (as confirmed by Figure 3.7). Interestingly, this phenomenon is stronger for single-task estimators (differences are smaller in Figure 3.2) and disappears when n is large enough (Figure 3.5), which is consistent with the heuristic motivating multi-task learning: "increasing the number of tasks p amounts to increase the sample size".

Figures 3.4 and 3.5 show that our procedure works well with small n, and that increasing n does not seem to significantly improve the performance of our estimators, except in the single-task setting with Σ known, where the over-penalization phenomenon discussed above disappears.

Table 3.2 shows that using the multitask procedure improves the estimation accuracy, both in the clustering setting and in the segmentation setting. The last line of Table 3.2 does not show that the clustering setting improves over the "segmentation into intervals" one, which was awaited if a model close to the oracle is selected in both cases.

Table 3.3 finally shows that our parameter tuning procedure outperforms 5-fold cross-validation.

3.7 Conclusion and Future Work

This paper shows that taking into account the unknown similarity between p regression tasks can be done optimally (Theorem 3.3). The crucial point is to estimate the $p \times p$ covariance matrix Σ of the noise (covariance between tasks), in order to learn the task similarity matrix M. Our main contributions are twofold. First, an estimator of Σ is defined in Section 3.4, where non-asymptotic bounds on its error are provided under mild assumptions on the mean of the sample (Theorem 3.2). Second, we show an oracle inequality (Theorem 3.3), more particularly with a simplified estimation of Σ and increased performances when the matrices of $\mathcal{M}$ are jointly diagonalizable (which often corresponds to cases where we have a prior knowledge of what the relations between the tasks would be). We do plan to expand our results to larger sets $\mathcal{M}$, which may require new concentration inequalities and new optimization algorithms.

Simulation experiments show that our algorithm works with reasonable sample sizes, and that our multi-task estimator often performs much better than its single-task counterpart. Up to the best of our knowledge, a theoretical proof of this point remains an open problem that we intend to investigate in a future work.

Appendices

We give here the proofs of the different results stated in Sections 3.2, 3.4 and 3.5. The proofs of our main results are contained in Sections 3.E and 3.F.

3.A Proof of Property 3.1

Proof. It is sufficient to show that $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ is positive-definite on $\mathcal{G}$. Take $g \in \mathcal{G}$ and $S = (S_{i,j})_{1 \leq i \leq j \leq p}$ the symmetric postive-definite matrix of size p verifying $S^2 = M$, and denote $T = S^{-1} = (T_{i,j})_{1 \leq i,j \leq p}$. Let f be the element of $\mathcal{G}$ defined by $\forall i \in \{1 \dots p\},\ g(\cdot, i) = \sum_{k=1}^{n} T_{i,k} f(\cdot, k)$. We then have:

$$\begin{aligned}
\langle g, g \rangle_{\mathcal{G}} &= \sum_{i=1}^{p} \sum_{j=1}^{p} M_{i,j} \langle g(\cdot, i), g(\cdot, j) \rangle_{\mathcal{F}} \\
&= \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{p} M_{i,j} T_{i,k} T_{j,l} \langle f(\cdot, k), f(\cdot, l) \rangle_{\mathcal{F}} \\
&= \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{p} T_{l,j} \langle f(\cdot, k), f(\cdot, l) \rangle_{\mathcal{F}} \sum_{i=1}^{p} M_{j,i} T_{i,k} \\
&= \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{p} T_{l,j} \langle f(\cdot, k), f(\cdot, l) \rangle_{\mathcal{F}} (M \cdot T)_{j,k} \\
&= \sum_{k=1}^{p} \sum_{l=1}^{p} T_{l,j} \langle f(\cdot, k), f(\cdot, l) \rangle_{\mathcal{F}} \sum_{j=1}^{p} T_{l,j} (M \cdot T)_{j,k} \\
&= \sum_{k=1}^{p} \sum_{l=1}^{p} \langle f(\cdot, k), f(\cdot, l) \rangle_{\mathcal{F}} (T \cdot M \cdot T)_{k,l} \\
&= \sum_{k=1}^{p} \| f(\cdot, k) \|_{\mathcal{F}}^2.
\end{aligned}$$

This shows that $\langle g, g \rangle_{\mathcal{G}} \geq 0$ and that $\langle g, g \rangle_{\mathcal{G}} = 0 \Rightarrow f = 0 \Rightarrow g = 0$. □

3.B Proof of Corollary 3.1

Proof. If $(x, j) \in \mathcal{X} \times \{1, \dots, p\}$, the application $(f^1, \dots, f^p) \mapsto f^j(x)$ is clearly continuous. We now show that $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ is complete. If $(g_n)_{n\in\mathbb{N}}$ is a Cauchy sequence of $\mathcal{G}$ and if we define, as in Section 3.A, the functions f_n by $\forall n \in \mathbb{N},\ \forall i \in \{1 \dots p\},\ g_n(\cdot, i) = \sum_{k=1}^{p} T_{i,k} f_n(\cdot, k)$. The same computations show that $(f_n(\cdot, i))_{n\in\mathbb{N}}$ are Cauchy sequences of $\mathcal{F}$, and thus converge. So the sequence $(f_n)_{n\in\mathbb{N}}$ converges in $\mathcal{G}$, and $(g_n)_{n\in\mathbb{N}}$ does likewise. □

3.C Proof of Property 3.2

Proof. We define

$$\widetilde{\Phi}(x, j) = M^{-1} \cdot \begin{pmatrix} \delta_{1,j}\Phi(x) \\ \vdots \\ \delta_{p,j}\Phi(x) \end{pmatrix} \ ,$$

with $\delta_{i,j} = \mathbf{1}_{i=j}$ being the Kronecker symbol, that is, $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. We now show that $\widetilde{\Phi}$ is the feature function of the RKHS. For $g \in \mathcal{G}$ and $(x, l) \in \mathcal{X} \times \{1, \dots, p\}$, we have:

$$\begin{aligned}
\langle g, \widetilde{\Phi}(x, l) \rangle_{\mathcal{G}} &= \sum_{j=1}^{p} \sum_{i=1}^{p} M_{j,i} \langle g(\cdot, j), \widetilde{\Phi}(x, l)^i \rangle_{\mathcal{F}} \\
&= \sum_{j=1}^{p} \sum_{i=1}^{p} \sum_{m=1}^{p} M_{j,i} M^{-1}_{i,m} \delta_{m,l} \langle g(\cdot, j), \Phi(x) \rangle_{\mathcal{F}} \\
&= \sum_{j=1}^{p} \sum_{m=1}^{p} (M \cdot M^{-1})_{j,m} \delta_{m,l} g(x, j) \\
&= \sum_{j=1}^{p} \delta_{j,l} g(x, j) = g(x, l) \ .
\end{aligned}$$

Thus we can write:

$$\begin{aligned}
\widetilde{k}((x, i), (y, j)) &= \langle \widetilde{\Phi}(x, i), \widetilde{\Phi}(y, j) \rangle_{\mathcal{G}} \\
&= \sum_{h=1}^{p} \sum_{h'=1}^{p} M_{h,h'} \langle M^{-1}_{h,i} \Phi(x), M^{-1}_{h',j} \Phi(y) \rangle_{\mathcal{F}} \\
&= \sum_{h=1}^{p} \sum_{h'=1}^{p} M_{h,h'} M^{-1}_{h,i} M^{-1}_{h',j} K(x, y) \\
&= \sum_{h=1}^{p} M^{-1}_{h,i} (M \cdot M^{-1})_{h,j} K(x, y) \\
&= \sum_{h=1}^{p} M^{-1}_{h,i} \delta_{h,j} K(x, y) = M^{-1}_{i,j} K(x, y) \ .
\end{aligned}$$

□

3.D Computation of the Quadratic Risk in Example 3.4

We consider here that $f^1 = \cdots = f^p$. We use the set $\mathcal{M}_{\text{SD}}$:

$$\mathcal{M}_{\text{SD}} := \left\{ M_{\text{SD}}(\lambda,\mu) = (\lambda + p\mu) I_p - \frac{\mu}{p}\mathbf{1}\mathbf{1}^\top \,/\, (\lambda,\mu) \in (0,+\infty)^2 \right\}$$

Using the estimator $\widehat{f}_M = A_M y$ we can then compute the quadratic risk using the bias-variance decomposition given in Equation (3.33):

$$\mathbb{E}\left[\left\| \widehat{f}_M - f \right\|_2^2 \right] = \|(A_M - I_{np}) f\|_2^2 + \operatorname{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ .$$

Les us denote by $(e_1, \ldots, e_p)$ the canonical basis of $\mathbb{R}^p$. The eigenspaces of $p^{-1}\mathbf{1}\mathbf{1}^\top$ are:

- span $\{e_1 + \cdots + e_p\}$ corresponding to eigenvalue p,
- span $\{e_2 - e_1, \ldots, e_p - e_1\}$ corresponding to eigenvalue 0.

Thus, with $\widetilde{\mu} = \lambda + p\mu$ we can diagonalize in an orthonormal basis any matrix $M_{\lambda,\mu} \in \mathcal{M}$ as $M = P^\top D_{\lambda,\widetilde{\mu}} P$, with $D = D_{\lambda,\widetilde{\mu}} = \text{Diag}\{\lambda, \widetilde{\mu}, \ldots, \widetilde{\mu}\}$. Les us also diagonalise in an orthonormal basis K: $K = Q^\top \Delta Q$, $\Delta = \text{Diag}\{\mu_1, \ldots, \mu_n\}$. Thus we can write (see Properties 3.3 and 3.4 for basic properties of the Kronecker product):

$$A_M = A_{M_{\lambda,\mu}} = (P^\top \otimes Q^\top) \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right] (P \otimes Q) \ .$$

We can then note that $(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1}$ is a diagonal matrix, whose diagonal entry of index $(j-1)n + i$ ($i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, p\}$) is

$$\begin{cases} \frac{\mu_i}{\mu_i + np\lambda} \text{ if } j = 1 \ , \\ \frac{\mu_i}{\mu_i + np\widetilde{\mu}} \text{ if } j > 1 \ . \end{cases}$$

We can now compute both bias and variance.

Bias: We can first remark that $(P^\top \otimes Q^\top) = (P \otimes Q)^\top$ is an orthogonal matrix and that $P \times \mathbf{1} = (1, 0, \ldots, 0)^\top$. Thus, as in this setting $f^1 = \cdots = f^p$, we have $f = \mathbf{1} \otimes (f^1(X_1), \ldots, f^1(X_n))^\top$ and $(P^\top \otimes Q^\top) f = (1, 0, \ldots, 0)^\top \otimes Q(f^1(X_1), \ldots, f^1(X_n))^\top$. To keep notations simple we note $Q(f^1(X_1), \ldots, f^1(X_n))^\top := (g_1, \ldots, g_n)^\top$. Thus

$$\begin{aligned} \|(A_M - I_{np}) f\|_2^2 &= \|(P \otimes Q)^\top \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} - I_{np} \right] (P \otimes Q) f\|_2^2 \\ &= \| \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} - I_{np} \right] \\ &\qquad \times (1, 0, \ldots, 0)^\top \otimes (g_1, \ldots, g_n)^\top \|_2^2 \ . \end{aligned}$$

As only the first n terms of $(P \otimes Q) f$ are non-zero we can finally write

$$\|(A_M - I_{np}) f\|_2^2 = \sum_{i=1}^{n} \left(\frac{np\lambda}{\mu_i + np\lambda} \right)^2 g_i^2 \ .$$

Variance: First note that

$$(P \otimes Q)(\Sigma \otimes I_n)(P \otimes Q)^\top = (P\Sigma P^\top \otimes I_n) \ .$$

We can also note that $\widetilde{\Sigma} := P\Sigma P^\top$ is a symmetric positive definite matrix, with positive diagonal coefficients. Thus we can finally write

$$\begin{aligned}
\mathrm{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) &= \mathrm{tr}\left(P \otimes Q)^\top \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right]^2 \right. \\
&\qquad \left. \times (P \otimes Q)(\Sigma \otimes I_n) \right) \\
&= \mathrm{tr}\left(\left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right]^2 \right. \\
&\qquad \left. \times (P \otimes Q)(\Sigma \otimes I_n)(P \otimes Q)^\top \right) \\
&= \sum_{i=1}^{n} \left[\left(\frac{\mu_i}{\mu_i + np\lambda} \right)^2 \widetilde{\Sigma}_{1,1} + \left(\frac{\mu_i}{\mu_i + np\widetilde{\mu}} \right)^2 \sum_{j=2}^{p} \widetilde{\Sigma}_{j,j} \right] \ .
\end{aligned}$$

As noted at the end of Example 3.4 this leads to an oracle which has all its p functions equal.

3.D.1 Proof of Equation (3.16) in Section 3.5.2

Let $M \in \mathcal{S}_p^{++}(\mathbb{R})$, $P \in \mathcal{O}_p(\mathbb{R})$ such that $M = P^\top \mathrm{Diag}(d_1, \ldots, d_p)P$ and $\widetilde{\Sigma} = P\Sigma P^\top$. We recall that $A_\lambda = K(K + n\lambda I_n)^{-1}$. The computations detailed above also show that the ideal penalty introduced in Eq. (3.7) can be written as

$$\mathrm{pen}_{\mathrm{id}}(M) = \frac{2\,\mathrm{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} = \frac{2}{np}\left(\sum_{j=1}^{p} \mathrm{tr}(A_{pd_j})\widetilde{\Sigma}_{j,j} \right) \ .$$

3.E Proof of Theorem 3.2

Theorem 3.2 is proved in this section, after stating some classical linear algebra results (Section 3.E.1).

3.E.1 Some Useful Tools

We now give two properties of the Kronecker product, and then introduce a useful norm on $\mathcal{S}_p(\mathbb{R})$, upon which we give several properties. Those are the tools needed to prove Theorem 3.2.

Property 3.3. *The Kronecker product is bilinear, associative and for every matrices A, B, C, D such that the dimensions fit, $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.*

Property 3.4. *Let $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_B(\mathbb{R})$, $(A \otimes B)^\top = (A^\top \otimes B^\top)$.*

Definition 3.7. We now introduce the norm $|||\cdot|||$ on $\mathcal{S}_p(\mathbb{R})$, which is the modulus of the eigenvalue of largest magnitude, and can be defined by

$$|||S||| := \sup_{z\in\mathbb{R}^p, \|z\|_2=1} \left| z^\top S z \right| \ .$$

This norm has several interesting properties, some of which we will use are stated below.

Property 3.5. *The norm $|||\cdot|||$ is a matricial norm: $\forall (A,B) \in \mathcal{S}_p(\mathbb{R})^2,\ |||AB||| \leq |||A|||B|||$.*

We will use the following result, which is a consequence of the preceding Property.

$$\forall S \in \mathcal{S}_p(\mathbb{R}),\ \forall T \in \mathcal{S}_p^{++}(\mathbb{R}),\ |||T^{-\frac{1}{2}} S T^{-\frac{1}{2}}||| \leq |||S|||T^{-1}||| \ .$$

We also have:

Property 3.6.

$$\forall \Sigma \in \mathcal{S}_p(\mathbb{R}),\ |||\Sigma \otimes I_n||| = |||\Sigma||| \ .$$

Proof. We can diagonalize Σ in an orthonormal basis: $\exists U \in \mathcal{O}_n(\mathbb{R}),\ \exists D = \mathrm{Diag}(\mu_1, \ldots, \mu_p),\ \Sigma = U^\top D U$. We then have, using the properties of the Kronecker product:

$$\begin{aligned} \Sigma \otimes I_n &= (U^\top \otimes I_n)(D \otimes I_n)(U \otimes I_n) \\ &= (U \otimes I_n)^\top (D \otimes I_n)(U \otimes I_n) \ . \end{aligned}$$

We just have to notice that $U \otimes I_n \in \mathcal{O}_{np}(\mathbb{R})$ and that:

$$D \otimes I_n = \mathrm{Diag}(\underbrace{\mu_1, \ldots, \mu_1}_{n \text{ times}}, \ldots, \underbrace{\mu_p, \ldots, \mu_p}_{n \text{ times}}) \ .$$

□

This norm can also be written in other forms:

Property 3.7. *If $M \in \mathcal{M}_n(\mathbb{R})$, the operator norm $\|M\|_2 := \sup_{t\in\mathbb{R}^n\setminus\{0\}} \left\{ \frac{\|Mt\|_2}{\|t\|_2} \right\}$ is equal to the greatest singular value of M: $\sqrt{\rho(M^\top M)}$. Henceforth, if S is symmetric, we have $|||S||| = \|S\|_2$*

3.E.2 The Proof

We now give a proof of Theorem 3.2, using Lemmas 3.1, 3.2 and 3.3, which are stated and proved in Section 3.E.3. The outline of the proof is the following:

1. Apply Theorem 3.1 to problem (P_z) for every $z \in \mathcal{Z}$ in order to
2. control $\|s - \zeta\|_\infty$ with a large probability, where $s, \zeta \in \mathbb{R}^{p(p+1)/2}$ are defined by

$$\begin{aligned} s &:= (\Sigma_{1,1}, \ldots, \Sigma_{p,p}, \Sigma_{1,1} + \Sigma_{2,2} + 2\Sigma_{1,2}, \ldots, \Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}, \ldots) \\ \text{and} \quad \zeta &:= (a(e_1), \ldots, a(e_p), a(e_1+e_2), \ldots, a(e_1+e_p), a(e_2+e_3), \ldots, a(e_{p-1}+e_p)) \ . \end{aligned}$$

3. Deduce that $\widehat{\Sigma} = J(\zeta)$ is close to $\Sigma = J(s)$ by controlling the Lipschitz norm of J.

3.E. PROOF OF THEOREM 3.2

Proof. 1. Apply Theorem 3.1: We start by noticing that Assumption (**Hdf**) actually holds true with all $\lambda_{0,j}$ equal. Indeed, let $(\lambda_{0,j})_{1\leq j\leq p}$ be given by Assumption (**Hdf**) and define $\lambda_0 := \min_{j=1,\dots,p}\lambda_{0,j}$. Then, $\lambda_0 \in (0,+\infty)$ and $\mathrm{df}(\lambda_0) \leq \sqrt{n}$ since all $\lambda_{0,j}$ satisfy these two conditions. For the last condition, remark that for every $j \in \{1,\dots,p\}$, $\lambda_0 \leq \lambda_{0,j}$ and $\lambda \mapsto \|(A_\lambda - I)F_{e_j}\|_2^2$ is a nonincreasing function [AB11, as noticed, for instance, in], so that

$$\frac{1}{n}\left\|(A_{\lambda_0} - I_n)F_{e_j}\right\|_2^2 \leq \frac{1}{n}\left\|(A_{\lambda_{0,j}} - I_n)F_{e_j}\right\|_2^2 \leq \Sigma_{j,j}\sqrt{\frac{\ln(n)}{n}} \ . \tag{3.20}$$

In particular, Eq. (3.8) holds with $d_n = 1$ for problem (P_z) whatever $z \in \{e_1,\dots,e_p\}$.

Let us now consider the case $z = e_i + e_j$ with $i \neq j \in \{1,\dots,p\}$. Using Eq. (3.20) and that $F_{e_i+e_j} = F_{e_i} + F_{e_j}$, we have

$$\left\|(A_{\lambda_0} - I_n)F_{e_i+e_j}\right\|_2^2 \leq \|(A_{\lambda_0} - I_n)F_{e_i}\|_2^2 + \left\|(A_{\lambda_0} - I_n)F_{e_j}\right\|_2^2 + 2\langle (A_{\lambda_0} - I_n)F_{e_i}, (A_{\lambda_0} - I_n)F_{e_j}\rangle \ .$$

The last term is bounded as follows:

$$\begin{aligned}
2\langle (B_{\lambda_0} - I_n)F_{e_i}, (B_{\lambda_0} - I_n)F_{e_j}\rangle &\leq 2\|(B_{\lambda_0} - I_n)F_{e_i}\| \cdot \|(B_{\lambda_0} - I_n)F_{e_j}\| \\
&\leq 2\sqrt{n\ln(n)}\sqrt{\Sigma_{i,i}\Sigma_{j,j}} \\
&\leq \sqrt{n\ln(n)}(\Sigma_{i,i} + \Sigma_{j,j}) \\
&\leq \frac{1+c(\Sigma)}{2}\sqrt{n\ln(n)}(\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}) \\
&= \frac{1+c(\Sigma)}{2}\sqrt{n\ln(n)}\sigma^2_{e_i+e_j} \ ,
\end{aligned}$$

because Lemma 3.1 shows

$$2(\Sigma_{i,i} + \Sigma_{j,j}) \leq (1 + c(\Sigma))(\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}) \ .$$

Therefore, Eq. (3.8) holds with $d_n = (1 + c(\Sigma))/2$ for problem (P_z) whatever $z \in \mathcal{Z}$.

2. Control $\|s - \zeta\|_\infty$: Let us define

$$\eta_1 := \beta(2+\delta)(1 + c(\Sigma))\sqrt{\frac{\ln(n)}{n}} \ .$$

By Theorem 3.1, for every $z \in \mathcal{Z}$, an event Ω_z of probability greater than $1 - n^{-\delta}$ exists on which, if $n \geq n_0(\delta)$,

$$(1 - \eta_1)\sigma_z^2 \leq a(z) \leq (1 + \eta_1)\sigma_z^2 \ .$$

So, on $\Omega := \bigcap_{z\in Z}\Omega_z$,

$$\|\zeta - s\|_\infty \leq \eta_1 \|s\|_\infty \ , \tag{3.21}$$

and $\mathbb{P}(\Omega) \geq 1 - p(p+1)/2 \times n^{-\delta}$ by the union bound. Let

$$\|\Sigma\|_\infty := \sup_{i,j}|\Sigma_{i,j}| \quad \text{and} \quad C_1(p) := \sup_{\Sigma\in\mathcal{S}_p(\mathbb{R})}\left\{\frac{\|\Sigma\|_\infty}{|||\Sigma|||}\right\} \ .$$

Since $\|s\|_\infty \leq 4\|\Sigma\|_\infty$ and $C_1(p) = 1$ by Lemma 3.2, Eq. (3.21) implies that on Ω,

$$\|\zeta - s\|_\infty \leq 4\eta_1\|\Sigma\|_\infty \leq 4\eta_1|||\Sigma||| \ . \tag{3.22}$$

3. Conclusion of the proof: Let

$$C_2(p) := \sup_{\zeta \in \mathbb{R}^{p(p+1)/2}} \left\{ \frac{|||J(\zeta)|||}{\|\zeta\|_\infty} \right\} \ .$$

By Lemma 3.3, $C_2(p) \leq \frac{3}{2}p$. By Eq. (3.22), on Ω,

$$|||\widehat{\Sigma} - \Sigma||| = |||J(\zeta) - J(s)||| \leq C_2(p) \, \|\zeta - s\|_\infty \leq 4\eta_1 C_2(p) |||\Sigma||| \ . \tag{3.23}$$

Since

$$|||\Sigma^{-\frac{1}{2}} \widehat{\Sigma} \Sigma^{-\frac{1}{2}} - I_p||| = |||\Sigma^{-\frac{1}{2}} (\Sigma - \widehat{\Sigma}) \Sigma^{-\frac{1}{2}}||| \leq |||\Sigma^{-1}||| |||\Sigma - \widehat{\Sigma}||| \ ,$$

and $|||\Sigma||| |||\Sigma^{-1}||| = c(\Sigma)$, Eq. (3.23) implies that on Ω,

$$|||\Sigma^{-\frac{1}{2}} \widehat{\Sigma} \Sigma^{-\frac{1}{2}} - I_p||| \leq 4\eta_1 C_2(p) |||\Sigma||| |||\Sigma^{-1}||| = 4\eta_1 C_2(p) c(\Sigma) \leq 6\eta_1 p c(\Sigma) \ .$$

To conclude, Eq. (3.12) holds on Ω with

$$\eta = 6pc(\Sigma)\beta(2+\delta)(1+c(\Sigma))\sqrt{\frac{\ln(n)}{n}} \leq L_1(2+\delta)p\sqrt{\frac{\ln(n)}{n}} c(\Sigma)^2 \tag{3.24}$$

for some numerical constant L_1. □

Remark 3.20. *As stated in Arlot and Bach [AB11], we need* $\sqrt{n_0(\delta)/\ln(n_0(\delta))} \geq 504$ *and* $\sqrt{n_0(\delta)}/\ln(n_0(\delta)) \geq 24(290+\delta)$.

Remark 3.21. *To ensure that the estimated matrix* $\widehat{\Sigma}$ *is positive-definite we need that* $\eta < 1$, *that is,*

$$\sqrt{\frac{n}{\ln(n)}} > 6\beta(2+\delta)pc(\Sigma)\,(1+c(\Sigma)) \ .$$

3.E.3 Useful Lemmas

Lemma 3.1. *Let* $p \geq 1$, $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$ *and* $c(\Sigma)$ *its condition number. Then,*

$$\forall 1 \leq i < j \leq p \,, \quad \Sigma_{i,j} \geq -\frac{c(\Sigma)-1}{c(\Sigma)+1} \frac{\Sigma_{i,i} + \Sigma_{j,j}}{2} \ , \tag{3.25}$$

Remark 3.22. *The proof of Lemma 3.1 shows the constant* $\frac{c(\Sigma)-1}{c(\Sigma)+1}$ *cannot be improved without additional assumptions on* Σ.

Proof. It suffices to show the result when $p = 2$. Indeed, (3.25) only involves 2×2 submatrices $\widetilde{\Sigma}(i,j) \in \mathcal{S}_2^{++}(\mathbb{R})$ for which

$$1 \leq c(\widetilde{\Sigma}) \leq c\,(\Sigma) \quad \text{hence} \quad 0 \leq \frac{c(\widetilde{\Sigma})-1}{c(\widetilde{\Sigma})+1} \leq \frac{c(\Sigma)-1}{c(\Sigma)+1} \ .$$

So, some $\theta \in \mathbb{R}$ exists such that $\Sigma = |||\Sigma||| R_\theta^\top D R_\theta$ where

$$R_\theta := \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \qquad D = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \quad \text{and} \quad \lambda := \frac{1}{c(\Sigma)} \ .$$

Therefore,

$$\Sigma = |||\Sigma||| \begin{pmatrix} \cos^2(\theta) + \lambda \sin^2(\theta) & \frac{1-\lambda}{2}\sin(2\theta) \\ \frac{1-\lambda}{2}\sin(2\theta) & \lambda\cos^2(\theta) + \sin^2(\theta) \end{pmatrix} \ .$$

So, Eq. (3.25) is equivalent to

$$\frac{(1-\lambda)\sin(2\theta)}{2} \geq -\frac{1-\lambda}{1+\lambda}\frac{1+\lambda}{2} \ ,$$

which holds true for every $\theta \in \mathbb{R}$, with equality for $\theta \equiv \pi/2 \ (\text{mod. } \pi)$. □

Lemma 3.2. *For every $p \geq 1$, $C_1(p) := \sup_{\Sigma \in \mathcal{S}_p(\mathbb{R})} \frac{\|\Sigma\|_\infty}{|||\Sigma|||} = 1$.*

Proof. With $\Sigma = I_p$ we have $\|\Sigma\|_\infty = |||\Sigma||| = 1$, so $C_1(p) \geq 1$.
Let us introduce (i,j) such that $|\Sigma_{i,j}| = \|\Sigma\|_\infty$. We then have, with e_k being the k^{th} vector of the canonical basis of $\mathbb{R}^p$,

$$|\Sigma_{i,j}| = |e_i^\top \Sigma e_j| \leq |e_i^\top \Sigma e_i|^{1/2} |e_j^\top \Sigma e_j|^{1/2} \leq (\|\Sigma\|_2^{1/2})^2 \ .$$

□

Lemma 3.3. *For every $p \geq 1$, let $C_2(p) := \sup_{\zeta \in \mathbb{R}^{p(p+1)/2}} \frac{|||J(\zeta)|||}{\|\zeta\|_\infty}$. Then,*

$$\frac{p}{4} \leq C_2(p) \leq \frac{3}{2}p \ .$$

Proof. For the lower bound, we consider

$$\zeta_1 = (\underbrace{1, \ldots, 1}_{p \text{ times}}, \underbrace{4, \ldots, 4}_{\frac{p(p-1)}{2} \text{ times}}), \quad \text{then} \quad J(\zeta_1) = \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$$

so that $|||J(\zeta)||| = p$ and $\|\zeta\|_\infty = 4$.

For the upper bound, we have for every $\zeta \in \mathbb{R}^{p(p+1)/2}$ and $z \in \mathbb{R}^p$ such that $\|z\|_2 = 1$

$$\left| z^\top J(\zeta) z \right| = \left| \sum_{1 \leq i,j \leq p} z_i z_j J(\zeta)_{i,j} \right| \leq \sum_{1 \leq i,j \leq p} |z_i| \, |z_j| \, |J(\zeta)_{i,j}| \leq \|J(\zeta)\|_\infty \|z\|_1^2 \ .$$

By definition of J, $\|J(\zeta)\|_\infty \leq 3/2 \, \|\zeta\|_\infty$. Remarking that $\|z\|_1^2 \leq p \, \|z\|_2^2$ yields the result. □

3.F Proof of Theorem 3.3

The proof of Theorem 3.3 is similar to the proof of Theorem 3 in Arlot and Bach [AB11]. We give it here for the sake of completeness. We also show how to adapt its proof to demonstrate Theorem 3.4. The two main mathematical results used here are Theorem 3.2 and a gaussian concentration inequality from Arlot and Bach [AB11].

3.F.1 Key Quantities and their Concentration Around their Means

Definition 3.8. We introduce, for $S \in \mathcal{S}_p^{++}(\mathbb{R})$,

$$\widehat{M}_o(S) \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{F}_M - Y \right\|_2 + 2\,\mathrm{tr}\left(A_M \cdot (S \otimes I_n)\right) \right\} \tag{3.26}$$

Definition 3.9. Let $S \in \mathcal{S}_p(\mathbb{R})$, we note S_+ the symmetric matrix where the eigenvalues of S have been thresholded at 0. That is, if $S = U^\top D U$, with $U \in \mathcal{O}_p(\mathbb{R})$ and $D = \mathrm{Diag}(d_1, \ldots, d_p)$, then

$$S_+ := U^\top \mathrm{Diag}\left(\max\{d_1, 0\}, \ldots, \max\{d_n, 0\}\right) U \ .$$

Definition 3.10. For every $M \in \mathcal{M}$, we define

$$\begin{aligned}
b(M) &= \|(A_M - I_{np}) f\|_2^2 \ , \\
v_1(M) &= \mathbb{E}\left[\langle \varepsilon, A_M \varepsilon \rangle\right] = \mathrm{tr}(A_M \cdot (\Sigma \otimes I_n)) \ , \\
\delta_1(M) &= \langle \varepsilon, A_M \varepsilon \rangle - \mathbb{E}\left[\langle \varepsilon, A_M \varepsilon \rangle\right] = \langle \varepsilon, A_M \varepsilon \rangle - \mathrm{tr}(A_M \cdot (\Sigma \otimes I_n)) \ , \\
v_2(M) &= \mathbb{E}\left[\|A_M \varepsilon\|_2^2\right] = \mathrm{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ , \\
\delta_2(M) &= \|A_M \varepsilon\|_2^2 - \mathbb{E}\left[\|A_M \varepsilon\|_2^2\right] = \|A_M \varepsilon\|_2^2 - \mathrm{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ , \\
\delta_3(M) &= 2\langle A_M \varepsilon, (A_M - I_{np}) f \rangle \ , \\
\delta_4(M) &= 2\langle \varepsilon, (I_{np} - A_M) f \rangle \ , \\
\widehat{\Delta}(M) &= -2\delta_1(M) + \delta_4(M) \ .
\end{aligned}$$

Definition 3.11. Let $C_A, C_B, C_C, C_D, C_E, C_F$ be fixed nonnegative constants. For every $x \geq 0$ we define the event

$$\Omega_x = \Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)$$

on which, for every $M \in \mathcal{M}$ and $\theta_1, \theta_2, \theta_3, \theta_4 \in (0, 1]$:

$$|\delta_1(M)| \leq \theta_1\,\mathrm{tr}\left(A_M^\top A_M \cdot (\Sigma \otimes I_n)\right) + (C_A + C_B \theta_1^{-1}) x |\!|\!| \Sigma |\!|\!| \tag{3.27}$$

$$|\delta_2(M)| \leq \theta_2\,\mathrm{tr}\left(A_M^\top A_M \cdot (\Sigma \otimes I_n)\right) + (C_C + C_D \theta_2^{-1}) x |\!|\!| \Sigma |\!|\!| \tag{3.28}$$

$$|\delta_3(M)| \leq \theta_3 \|(I_{np} - A_M) f\|_2^2 + C_E \theta_3^{-1} x |\!|\!| \Sigma |\!|\!| \tag{3.29}$$

$$|\delta_4(M)| \leq \theta_4 \|(I_{np} - A_M) f\|_2^2 + C_F \theta_4^{-1} x |\!|\!| \Sigma |\!|\!| \tag{3.30}$$

Of key interest is the concentration of the empirical processes δ_i, uniformly over $M \in \mathcal{M}$. The following Lemma introduces such a result, when $\mathcal{M}$ contains symmetric matrices parametrized with their eigenvalues (with fixed eigenvectors).

Lemma 3.4. *Let*

$$C_A = 2,\ C_B = 1,\ C_C = 2,\ C_D = 1,\ C_E = 306.25,\ C_F = 306.25 \ .$$

Suppose that (**HM**) *holds. Then* $\mathbb{P}(\Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)) \geq 1 - p e^{1027 + \ln(n)} e^{-x}$. *Suppose that* (3.13) *holds. Then* $\mathbb{P}(\Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)) \geq 1 - 6p\,\mathrm{card}(\mathcal{M}) e^{-x}$.

.

3.F. PROOF OF THEOREM 3.3

Proof. **First common step.** Let $M \in \mathcal{M}$, $P_M \in \mathcal{O}_p(\mathbb{R})$ such that $M = P_M^\top D P_M$, with $D = \mathrm{Diag}(d_1, \dots, d_p)$. We can write:

$$\begin{aligned} A_M = A_{d_1,\dots,d_p} &= (P_M \otimes I_n)^\top \left[(D^{-1} \otimes K) \left(D^{-1} \otimes K + npI_{np} \right)^{-1} \right] (P_M \otimes I_n) \\ &= Q^\top \widetilde{A}_{d_1,\dots,d_p} Q \ , \end{aligned}$$

with $Q = P_M \otimes I_n$ and $\widetilde{A}_{d_1,\dots,d_p} = (D^{-1} \otimes K)(D^{-1} \otimes K + npI_{np})^{-1}$. Remark that $\widetilde{A}_{d_1,\dots,d_p}$ is block-diagonal, with diagonal blocks being $B_{d_1}, \dots, B_{d_p}$ using the notations of Section 3.3. With $\widetilde{\varepsilon} = Q\varepsilon = (\widetilde{\varepsilon_1}^\top, \dots, \widetilde{\varepsilon_p}^\top)^\top$ and $\widetilde{f} = Qf = (\widetilde{f_1}^\top, \dots, \widetilde{f_p}^\top)^\top$ we can write

$$\begin{aligned} |\delta_1(M)| &= \langle \widetilde{\varepsilon}, \widetilde{A}_{d_1,\dots,d_p} \widetilde{\varepsilon} \rangle - \mathbb{E}\left[\langle \widetilde{\varepsilon}, \widetilde{A}_{d_1,\dots,d_p} \widetilde{\varepsilon} \rangle \right] \ , \\ |\delta_2(M)| &= \left\| \widetilde{A}_{d_1,\dots,d_p} \widetilde{\varepsilon} \right\|_2^2 - \mathbb{E}\left[\left\| \widetilde{A}_{d_1,\dots,d_p} \widetilde{\varepsilon} \right\|_2^2 \right] \ , \\ |\delta_3(M)| &= 2\langle \widetilde{A}_{d_1,\dots,d_p} \widetilde{\varepsilon}, (\widetilde{A}_{d_1,\dots,d_p} - I_{np}) \widetilde{f} \rangle \ , \\ |\delta_4(M)| &= 2\langle \widetilde{\varepsilon}, (I_{np} - \widetilde{A}_{d_1,\dots,d_p}) \widetilde{f} \rangle \ . \end{aligned}$$

We can see that the quantities δ_i decouple, therefore

$$\begin{aligned} |\delta_1(M)| &= \sum_{i=1}^p \langle \widetilde{\varepsilon}_i, A_{pd_i} \widetilde{\varepsilon}_i \rangle - \mathbb{E}\left[\langle \widetilde{\varepsilon}_i, A_{pd_i} \widetilde{\varepsilon}_i \rangle \right] \ , \\ |\delta_2(M)| &= \sum_{i=1}^p \| A_{pd_i} \widetilde{\varepsilon}_i \|_2^2 - \mathbb{E}\left[\| A_{pd_i} \widetilde{\varepsilon}_i \|_2^2 \right] \ , \\ |\delta_3(M)| &= \sum_{i=1}^p 2\langle A_{pd_i} \widetilde{\varepsilon}_i, (A_{pd_i} - I_n) \widetilde{f}_i \rangle \ , \\ |\delta_4(M)| &= \sum_{i=1}^p 2\langle \widetilde{\varepsilon}_i, (I_n - A_{pd_i}) \widetilde{f}_i \rangle \ . \end{aligned}$$

Supposing (HM). Assumption (**HM**) implies that the matrix P used above is the same for all the matrices M of $\mathcal{M}$. Using Lemma 9 of Arlot and Bach [AB11], where we have p concentration results on the sets $\widetilde{\Omega}_i$, each of probability at least $1 - e^{1027+\ln(n)} e^{-x}$

we can state that, on the set $\bigcap_{i=1}^{p} \widetilde{\Omega}_i$, we have uniformly on $\mathcal{M}$

$$
\begin{aligned}
|\delta_1(M)| &\leq \sum_{i=1}^{p} \theta_1 \operatorname{Var}[\widetilde{\varepsilon}_i] \operatorname{tr}(A_{pd_i}^\top A_{pd_i}) + (C_A + C_B \theta_1^{-1}) x \operatorname{Var}[\widetilde{\varepsilon}_i] \ , \\
|\delta_2(M)| &\leq \sum_{i=1}^{p} \theta_2 \operatorname{Var}[\widetilde{\varepsilon}_i] \operatorname{tr}(A_{pd_i}^\top A_{pd_i}) + (C_C + C_D \theta_2^{-1}) x \operatorname{Var}[\widetilde{\varepsilon}_i] \ , \\
|\delta_3(M)| &\leq \sum_{i=1}^{p} \theta_3 \left\| (I_n - A_{pd_i}) \widetilde{f}_i \right\|_2^2 + C_E \theta_3^{-1} x \operatorname{Var}[\widetilde{\varepsilon}_i] \ , \\
|\delta_4(M)| &\leq \sum_{i=1}^{p} \theta_4 \left\| (I_n - A_{pd_i}) \widetilde{f}_i \right\|_2^2 + C_F \theta_4^{-1} x \operatorname{Var}[\widetilde{\varepsilon}_i] \ .
\end{aligned}
$$

Supposing (3.13)**.** We can use Lemma 8 of Arlot and Bach [AB11] where we have p concentration results on the sets $\widetilde{\Omega}_{j,M}$, each of probability at least $1 - 6e^{-x}$ we can state that, on the set $\bigcap_{j=1}^{p} \bigcap_{M \in \mathcal{M}} \widetilde{\Omega}_i$, we have uniformly on $\mathcal{M}$ the same inequalities written above.

Final common step. To conclude, it suffices to see that $\forall i \in \{1, \ldots, p\}$, $\operatorname{Var}[\widetilde{\varepsilon}_i] \leq \|\!|\Sigma|\!\|$. □

3.F.2 Intermediate Result

We first prove a general oracle inequality, under the assumption that the penalty we use (with an estimator of Σ) does not underestimate the ideal penalty (involving Σ) too much.

Property 3.8. *Let $C_A, C_B, C_C, C_D, C_E \geq 0$ be fixed constants, $\gamma > 0$, $\theta_S \in [0, 1/4)$ and $K_S \geq 0$. On $\Omega_{\gamma \ln(n)}(\mathcal{M}, C_A, C_B, C_C, C_D, C_E)$, for every $S \in \mathcal{S}_p^{++}(\mathbb{R})$ such that*

$$
\begin{aligned}
&\operatorname{tr}\left(A_{\widehat{M}_o(S)} \cdot ((S - \Sigma) \otimes I_n) \right) \\
&\qquad \geq -\theta_S \operatorname{tr}\left(A_{\widehat{M}_o(S)} \cdot (\Sigma \otimes I_n) \right) \inf_{M \in \mathcal{M}} \left\{ \frac{b(M) + v_2(M) + K_S \ln(n) \|\!|\Sigma|\!\|}{v_1(M)} \right\}
\end{aligned} \tag{3.31}
$$

and for every $\theta \in (0, (1 - 4\theta_S)/2)$, we have:

$$
\begin{aligned}
&\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_o(S)} - f \right\|_2^2 \leq \frac{1 + 2\theta}{1 - 2\theta - 4\theta_S} \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{F}_M - F \right\|_2^2 + \frac{2 \operatorname{tr}\left(A_M \cdot ((S - \Sigma)_+ \otimes I_n) \right)}{np} \right\} \\
&+ \frac{1}{1 - 2\theta - 4\theta_S} \left[(2C_A + 3C_C + 6C_D + 6C_E + \frac{2}{\theta}(C_B + C_F))\gamma + \frac{\theta_S K_S}{4} \right] \frac{\ln(n) \|\!|\Sigma|\!\|}{np}
\end{aligned} \tag{3.32}
$$

Proof. The proof of Property 3.8 is very similar to the one of Property 5 in Arlot and Bach [AB11]. First, we have

$$
\left\| \widehat{f}_M - f \right\|_2^2 = b(M) + v_2(M) + \delta_2(M) + \delta_3(M) \ , \tag{3.33}
$$

$$
\left\| \widehat{f}_M - y \right\|_2^2 = \| \widehat{f}_M - f \|_2^2 - 2v_1(M) - 2\delta_1(M) + \delta_4(M) + \|\varepsilon\|_2^2 \ . \tag{3.34}
$$

3.F. PROOF OF THEOREM 3.3

Combining Eq. (3.26) and (3.34), we get:

$$\begin{aligned}&\left\|\widehat{f}_{\widehat{M}_o(S)} - f\right\|_2^2 + 2\,\mathrm{tr}\left(A_{\widehat{M}_o(S)} \cdot ((S-\Sigma)_+ \otimes I_n)\right) + \widehat{\Delta}(\widehat{M}_o(S)) \\ &\leq \inf_{M\in\mathcal{M}} \left\{\left\|\widehat{f}_M - f\right\|_2^2 + 2\,\mathrm{tr}\left(A_M \cdot ((S-\Sigma)\otimes I_n)\right) + \widehat{\Delta}(M)\right\} \ .\end{aligned} \tag{3.35}$$

On the event $\Omega_{\gamma\ln(n)}$, for every $\theta \in (0,1]$ and $M \in \mathcal{M}$, using Eq. (3.27) and (3.30) with $\theta = \theta_1 = \theta_4$,

$$|\widehat{\Delta}(M)| \leq \theta(b(M) + v_2(M)) + (C_A + \frac{1}{\theta}(C_B + C_F))\gamma\ln(n)|||\Sigma||| \ . \tag{3.36}$$

Using Eq. (3.28) and (3.29) with $\theta_2 = \theta_3 = 1/2$ we get that, for every $M \in \mathcal{M}$,

$$\left\|\widehat{F}_M - F\right\|_2^2 \geq \frac{1}{2}(b(M) + v_2(M)) - (C_C + 2C_D + 2C_E)\gamma\ln(n)|||\Sigma||| \ ,$$

which is equivalent to

$$b(M) + v_2(M) \leq 2\left\|\widehat{F}_M - F\right\|_2^2 + 2(C_C + 2C_D + 2C_E)\gamma\ln(n)|||\Sigma||| \ . \tag{3.37}$$

Combining Eq. (3.36) and (3.37), we get

$$|\widehat{\Delta}(M)| \leq 2\theta\left\|\widehat{F}_M - F\right\|_2^2 + \left(C_A + (2C_C + 4C_D + 4C_E)\theta + (C_B + C_F)\frac{1}{\theta}\right)\gamma\ln(n)|||\Sigma||| \ .$$

With Eq. (3.35), and with $C_1 = C_A$, $C_2 = 2C_C + 4C_D + 4C_E$ and $C_3 = C_B + C_F$ we get

$$\begin{aligned}&(1-2\theta)\left\|\widehat{f}_{\widehat{M}_o(S)} - f\right\|_2^2 + 2\,\mathrm{tr}\left(A_{\widehat{M}_o(S)} \cdot ((S-\Sigma)_+ \otimes I_n)\right) \leq (1+2\theta) \\ &\times \inf_{M\in\mathcal{M}} \left\{\left\|\widehat{f}_M - f\right\|_2^2 + 2\,\mathrm{tr}\left(A_M \cdot ((S-\Sigma)\otimes I_n)\right)\right\} + \left(C_1 + C_2\theta + \frac{C_3}{\theta}\right)\gamma\ln(n)|||\Sigma||| \ .\end{aligned} \tag{3.38}$$

Using Eq. (3.31) we can state that

$$\begin{aligned}&\mathrm{tr}\left(A_{\widehat{M}_o(S)} \cdot ((S-\Sigma)\otimes I_n)\right) \geq \\ &\qquad -\theta_S \frac{b(\widehat{M}_o(S)) + v_2(\widehat{M}_o(S)) + K_S\ln(n)|||\Sigma|||}{v_1(\widehat{M}_o(S))}\,\mathrm{tr}\left(A_{\widehat{M}_o(S)} \cdot (\Sigma\otimes I_n)\right)\end{aligned}$$

so that

$$\mathrm{tr}\left(A_{\widehat{M}_o(S)} \cdot ((S-\Sigma)\otimes I_n)\right) \geq -\theta_S\left((b(\widehat{M}_o(S)) + v_2(\widehat{M}_o(S)) + K_S\ln(n)|||\Sigma|||\right) \ ,$$

which then leads to Eq. (3.32) using Eq. (3.37) and (3.38). □

3.F.3 The Proof Itself

We now show Theorem 3.3 as a consequence of Property 3.8. It actually suffices to show that $\widehat{\Sigma}$ does not underestimate Σ too much, and that the second term in the infimum of Eq. (3.32) is negligible in front of the quadratic error $(np)^{-1}\|\widehat{f}_M - f\|^2$.

Proof. On the event Ω introduced in Theorem 3.2, Eq. (3.12) holds. Let

$$\gamma = pc(\Sigma)\left(1 + c(\Sigma)\right) \ .$$

By Lemma 3.5 below, we have:

$$\inf_{M\in\mathcal{M}}\left\{\frac{b(M) + v_2(M) + K_S\ln(n)|||\Sigma|||}{v_1(M)}\right\} \geq 2\sqrt{\frac{K_S\ln(n)|||\Sigma|||}{n\,\mathrm{tr}(\Sigma)}} \ .$$

We supposed Assumption (3.13) holds. Using elementary algebra it is easy to show that, for every symmetric positive definite matrices A, M and N of size p, $M \succeq N$ implies that $\mathrm{tr}(AM) \geq \mathrm{tr}(AN)$. In order to have $\widehat{M}_o(\widehat{\Sigma})$ satisfying Eq. (3.31), Theorem 3.2 shows that it suffices to have, for every $\theta_S > 0$,

$$2\theta_S\sqrt{\frac{K_S\ln(n)|||\Sigma|||}{n\,\mathrm{tr}(\Sigma)}} = 6\beta(2+\delta)\gamma\sqrt{\frac{\ln(n)}{n}} \ ,$$

which leads to the choice

$$K_S = \left(\frac{3\beta(\alpha+\delta)\gamma\,\mathrm{tr}(\Sigma)}{\theta_S|||\Sigma|||}\right)^2 \ .$$

We now take $\theta_S = \theta = (9\ln(n))^{-1}$. Let Ω be the set given by Theorem 3.2. Using Eq. (3.32) and requiring that $\ln(n) \geq 6$ we get, on the set $\widetilde{\Omega} = \Omega \cap \Omega_{(\alpha+\delta)\ln(n)}(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)$ of probability $1 - (p(p+1)/2 + 6pC)n^{-\delta}$, using that $\alpha \geq 2$:

$$\frac{1}{np}\left\|\widehat{f}_{\widehat{M}} - f\right\|_2 \leq \left(1 + \frac{1}{\ln(n)}\right)\inf_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2 + \frac{2\,\mathrm{tr}\left(A_M\cdot((\widehat{\Sigma}-\Sigma)_+\otimes I_n)\right)}{np}\right\}$$
$$+\left(1 - \frac{2}{3\ln(n)}\right)^{-1}\left[2C_A + 3C_C + 6C_D + 6C_E + \ln(n)\left(18C_B + 18C_F + \frac{729\beta^2\gamma^2\,\mathrm{tr}(\Sigma)^2}{4|||\Sigma|||^2}\right)\right]$$
$$\times(\alpha+\delta)^2\frac{\ln(n)^2|||\Sigma|||}{np} \ .$$

Using Eq. (3.24) and defining

$$\eta_2 := 12\beta(\alpha+\delta)\gamma\sqrt{\frac{\ln(n)}{n}} \ ,$$

we get

$$\frac{1}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2 \leq \left(1+\frac{1}{\ln(n)}\right)\inf_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M-f\right\|_2^2+\eta_2\frac{\mathrm{tr}(A_M\cdot(\Sigma\otimes I_n))}{np}\right\}$$
$$+\left(1-\frac{2}{3\ln(n)}\right)^{-1}\left[2C_A+3C_C+6C_D+6C_E+\ln(n)\left(18C_B+18C_F+\frac{729\beta^2\gamma^2\,\mathrm{tr}(\Sigma)^2}{4|||\Sigma|||^2}\right)\right]$$
$$\times(\alpha+\delta)^2\frac{\ln(n)^2|||\Sigma|||}{np}\ . \tag{3.39}$$

Now, to get a classical oracle inequality, we have to show that $\eta_2 v_1(M)=\eta_2\,\mathrm{tr}(A_M\cdot(\Sigma\otimes I_n))$ is negligible in front of $\|\widehat{f}_M-f\|^2$. Lemma 3.5 ensures that:

$$\forall M\in\mathcal{M}\,,\,\forall x\geq 0\,,\quad 2\sqrt{\frac{x|||\Sigma|||}{n\,\mathrm{tr}(\Sigma)}}v_1(M)\leq v_2(M)+x|||\Sigma|||\ .$$

With $0<C_n<1$, taking x to be equal to $72\beta^2\ln(n)\gamma^2\,\mathrm{tr}(\Sigma)/(C_n|||\Sigma|||)$ leads to

$$\eta_2 v_1(M)\leq 2C_n v_2(M)+\frac{72\beta^2\ln(n)\gamma^2\,\mathrm{tr}(\Sigma)}{C_n}\ . \tag{3.40}$$

Then, since $v_2(M)\leq v_2(M)+b(M)$ and using also Eq. (3.33), we get

$$v_2(M)\leq\left\|\widehat{f}_M-f\right\|_2^2+|\delta_2(m)|+|\delta_3(M)|\ .$$

On $\widetilde{\Omega}$ we have that for every $\theta\in(0,1)$, using Eq. (3.28) and (3.29),

$$|\delta_2(M)|+|\delta_3(M)|\leq 2\theta\left(\left\|\widehat{f}_M-f\right\|_2^2-|\delta_2(M)|-|\delta_3(M)|\right)+(C_C+(C_D+C_E)\theta^{-1})(\alpha+\delta)\ln(n)|||\Sigma|||\ ,$$

which leads to

$$|\delta_2(M)|+|\delta_3(M)|\leq\frac{2\theta}{1+2\theta}\left\|\widehat{f}_M-f\right\|_2^2+\frac{C_C+(C_D+C_E)\theta^{-1}}{1+2\theta}(\alpha+\delta)\ln(n)|||\Sigma|||\ .$$

Now, combining this equation with Eq. (3.40), we get

$$\eta_2 v_1(M)\leq\left(1+\frac{4C_n\theta}{1+2\theta}\right)\left\|\widehat{f}_M-f\right\|_2^2+2C_n\frac{C_C+(C_D+C_E)\theta^{-1}}{1+2\theta}(\alpha+\delta)\ln(n)|||\Sigma|||$$
$$+\frac{72\beta^2\ln(n)\gamma^2\,\mathrm{tr}(\Sigma)}{C_n}\ .$$

Taking $\theta=1/2$ then leads to

$$\eta_2 v_1(M)\leq(1+C_n)\left\|\widehat{f}_M-f\right\|_2^2+C_n(C_C+2(C_D+C_E))(\alpha+\delta)\ln(n)|||\Sigma|||$$
$$+\frac{72\beta^2\ln(n)\gamma\,\mathrm{tr}(\Sigma)}{C_n}\ .$$

We now take $C_n = 1/\ln(n)$. We now replace the constants C_A, C_B, C_C, C_D, C_E, C_F by their values in Lemma 3.4 and we get, for some constant L_2,

$$\left(1-\frac{2}{3\ln(n)}\right)^{-1}\left[1851.5+\ln(n)\left(5530.5+\frac{729\beta^2\gamma^2}{4|||\Sigma|||^2}\right)+616.5\left(1+\frac{1}{\ln(n)}\right)\frac{1}{\ln(n)}\right] + \frac{72\beta^2\ln(n)\gamma^2\operatorname{tr}(\Sigma)}{C_n} \leq L_2\ln(n)\gamma^2\frac{\operatorname{tr}(\Sigma)^2}{|||\Sigma|||^2}$$

From this we can deduce Eq. (3.14) by noting that $\gamma \leq 2pc(\Sigma)^2$.

Finally we deduce an oracle inequality in expectation by noting that if $n^{-1}\|f_{\widehat{M}} - f\|^2 \leq R_{n,\delta}$ on $\widetilde{\Omega}$, using Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^2\right] &= \mathbb{E}\left[\frac{\mathbf{1}_{\widetilde{\Omega}}}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^2\right]+\mathbb{E}\left[\frac{\mathbf{1}_{\widetilde{\Omega}^c}}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^2\right] \\ &\leq \mathbb{E}\left[R_{n,\delta}\right]+\frac{1}{np}\sqrt{\frac{4p(p+1)+6pC}{n^\delta}}\sqrt{\mathbb{E}\left[\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^4\right]}\ . \qquad (3.41)\end{aligned}$$

We can remark that, since $|||A_M||| \leq 1$,

$$\left\|\widehat{f}_M - f\right\|_2^2 \leq 2\left\|A_M\varepsilon\right\|_2^2 + 2\left\|(I_{np}-A_M)f\right\|_2^2 \leq 2\left\|\varepsilon\right\|_2^2 + 8\left\|f\right\|_2^2\ .$$

So

$$\mathbb{E}\left[\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^4\right] \leq 12\left(np|||\Sigma||| + 4\left\|f\right\|_2^2\right)^2\ ,$$

together with Eq. (3.39) and Eq. (3.41), induces Eq. (3.15), using that for some constant $L_3 > 0$,

$$12\sqrt{\frac{p(p+1)/2+6pC}{n^\delta}}\left(|||\Sigma||| + \frac{4}{np}\left\|f\right\|_2^2\right) \leq L_3\frac{\sqrt{p(p+C)}}{n^{\delta/2}}\left(|||\Sigma||| + \frac{1}{np}\left\|f\right\|_2^2\right)\ .$$

□

Lemma 3.5. *Let $n, p \geq 1$ be two integers, $x \geq 0$ and $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$. Then,*

$$\inf_{A\in\mathcal{M}_{np}(\mathbb{R}),|||A|||\leq 1}\left\{\frac{\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n))+x|||\Sigma|||}{\operatorname{tr}(A\cdot(\Sigma\otimes I_n))}\right\} \geq 2\sqrt{\frac{x|||\Sigma|||}{n\operatorname{tr}(\Sigma)}}$$

Proof. First note that the bilinear form on $\mathcal{M}_{np}(\mathbb{R})$, $(A,B) \mapsto \operatorname{tr}(A^\top B\cdot(\Sigma\otimes I_n))$ is a scalar product. By Cauchy-Schwarz inequality, for every $A \in \mathcal{M}_{np}(\mathbb{R})$,

$$\operatorname{tr}(A\cdot(\Sigma\otimes I_n))^2 \leq \operatorname{tr}(\Sigma\otimes I_n)\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n))\ .$$

Thus, since $\operatorname{tr}(\Sigma\otimes I_n) = n\operatorname{tr}(\Sigma)$, if $c = \operatorname{tr}(A\cdot(\Sigma\otimes I_n)) > 0$, $\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n)) \geq \dfrac{c^2}{n\operatorname{tr}(\Sigma)}$. Therefore,

$$\begin{aligned}\inf_{A\in\mathcal{M}_{np}(\mathbb{R}),|||A|||\leq 1}\left\{\frac{\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n))+x|||\Sigma|||}{\operatorname{tr}(A\cdot(\Sigma\otimes I_n))}\right\} &\geq \inf_{c>0}\left\{\frac{c}{n\operatorname{tr}(\Sigma)}+\frac{x|||\Sigma|||}{c}\right\} \\ &\geq 2\sqrt{\frac{x|||\Sigma|||}{n\operatorname{tr}(\Sigma)}}\ .\end{aligned}$$

□

3.F.4 Proof of Theorem 3.4

We now prove Theorem 3.4, first by proving that $\widehat{\Sigma}_{\mathrm{HM}}$ leads to a sharp enough approximation of the penalty.

Lemma 3.6. *Let $\widehat{\Sigma}_{HM}$ be defined as in Definition 3.6, $\alpha = 2$, $\kappa > 0$ be the numerical constant defined in Theorem 3.1 and assume* (**Hdf**) *and* (**HM**) *hold. For every $\delta \geq 2$, a constant $n_0(\delta)$, an absolute constant $L_1 > 0$ and an event Ω exist such that $\mathbb{P}(\Omega_{HM}) \geq 1 - pn^{-\delta}$ and for every $n \geq n_0(\delta)$, on Ω_{HM}, for every M in $\mathcal{M}$*

$$(1-\eta)\operatorname{tr}(A_M \cdot (\Sigma \otimes I_n)) \leq \operatorname{tr}(A_M \cdot (\widehat{\Sigma}_{HM} \otimes I_n)) \leq (1+\eta)\operatorname{tr}(A_M \cdot (\Sigma \otimes I_n)) \ , \tag{3.42}$$

where $\eta := L_1(\alpha+\delta)\sqrt{\dfrac{\ln(n)}{n}}$.

Proof. Let P be defined by (**HM**). Let $M \in \mathcal{M}$, and $(d_1, \ldots, d_p) \in (0, +\infty)^p$ such that $M = P^\top \operatorname{Diag}(d_1, \ldots, d_p)P$. Thus, as shown in Section 3.D, we have with $\widetilde{\Sigma} = P\Sigma P^\top$:

$$\operatorname{tr}(A_M \cdot (\Sigma \otimes I_n)) = \sum_{j=1}^{p} \operatorname{tr}(A_{pd_j})\widetilde{\Sigma}_{j,j} \ .$$

let $\widetilde{\sigma}_j$ be defined as in Definition 3.6 (and thus $\widehat{\Sigma}_{\mathrm{HM}} = P\operatorname{Diag}(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_p)P^\top$), we then have by Theorem 3.1 that for every $j \in \{1, \ldots, p\}$ an event Ω^j of probability $1 - \kappa n^{-\delta}$ exists such that on Ω^j $|\widetilde{\Sigma}_{j,j} - \widetilde{\sigma}_j| \leq \eta\widetilde{\Sigma}_{j,j}$. Since

$$\operatorname{tr}(A_M \cdot (\widehat{\Sigma}_{\mathrm{HM}} \otimes I_n)) = \sum_{j=1}^{p} \operatorname{tr}(A_{pd_j})\widetilde{\sigma}_j \ ,$$

taking $\Omega_{\mathrm{HM}} = \cap_{j=1}^{p}\Omega^j$ suffices to conclude. □

Proof of Theorem 3.3. Adapting the proof of Theorem 3.3 to Assumption (**HM**) first requires to take $\gamma = 1$ as Lemma 3.6 allows us. It then suffices to take the set $\widetilde{\Omega} = \Omega_{\mathrm{HM}} \cap \Omega_{(2+\delta)\ln(n)}(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)$ (thus replacing α by 2) of probability $1-(p(p+1)/2+p)n^{-\delta} \geq 1-p^2n^{-\delta}$—supposing $p \geq 2$—if we require that $2\ln(n) \geq 1027$.

To get to the oracle inequality in expectation we use the same technique than above, but we note that $\sqrt{\mathbb{P}(\widetilde{\Omega}^c)} \leq \widetilde{L_4} \times p/n^{\delta/2}$. We can finally define the constant L_4 by:

$$L_3 \operatorname{tr}(\Sigma)(2+\delta)^2\frac{p\ln(n)^3}{np} + \frac{p}{n^{\delta/2}}|\!|\!|\Sigma|\!|\!| \leq L_4\gamma^2 \operatorname{tr}(\Sigma)(\alpha+\delta)^2\frac{p\ln(n)^3}{np} \ .$$

□

Chapitre 4

Comparison between multi-task and single-task oracle risks in kernel ridge regression

Résumé. Dans ce chapitre, nous essayons de comprendre quand la procédure multi-tâches, introduite dans le chapitre précédent, donne de meilleures performances que la procédure mono-tâche, en termes de risque quadratique moyenné sur les tâches. Nous menons cette comparaison en considérant que les estimateurs sont parfaitement calibrés, ce qui revient à étudier leur risque oracle. Cela nous permet de dégager des situations favorables à la procédure multi-tâches, dans lesquelles cette dernière atteint des vitesses de convergence supérieures à celles de la procédure mono-tâche. Dans les cas contraires, où nous conjecturions que la procédure multi-tâches fonctionne moins bien que celle mono-tâche, nous montrons que les oracles respectifs se comportent de même. Des simulations viennent confirmer ces observations théoriques dans des situations moins contraintes. Il résulte donc de ces travaux que l'utilisation de ces méthodes multi-tâches peut être d'un grand secours, quand elles sont utilisées à bon escient. Cependant, comme nous le montrons, la moindre erreur de modélisation peut mener à de lourdes pertes.

4.1 Introduction

Increasing the sample size is the most common way to improve the performance of statistical estimators. In some cases (see, for instance, the experiments of Evgeniou et al. [EMP05] on customer data analysis or those of Jacob et al. [JBV08] on molecule binding problems), having access to some new data may be impossible, often due to experimental limitations. One way to circumvent those constraints is to use datasets from several related (and, hopefully, "similar") problems, as if it gave additional (in some sense) observations on the initial problem. The statistical methods using this heuristic are called "multi-task" techniques, as opposed to "single-task" techniques, where every problem is treated one at a

time. In this paper, we study kernel ridge regression in a multi-task framework and try to understand when multi-task can improve over single-task.

The first trace of a multi-task estimator can be found in the work of Stein [Ste56]. In this article, Charles Stein showed that the usual maximum-likelihood estimator of the mean of a Gaussian vector (of dimension larger than 3, every dimension representing here a task) is not admissible—that is, there exists another estimator that has a lower risk for every parameter. He showed the existence of an estimator that uniformly attains a lower quadratic risk by shrinking the estimators along the different dimensions towards an arbitrary point. An explicit form of such an estimator was given by James and Stein [JS61], yielding the famous James-Stein estimator. This phenomenon, now known as the "Stein's paradox", was widely studied in the following years and the behaviour of this estimator was confirmed by empirical studies, in particular the one from Efron and Morris [EM77]. This first example clearly shows the goals of the multi-task procedure: an advantage is gained by borrowing information from different tasks (here, by shrinking the estimators along the different dimensions towards a common point), the improvement being scored by the global (averaged) squared risk. Therefore, this procedure does not guarantee individual gains on every task, but a global improvement on the sum of those task-wise risks.

We consider here $p \geq 2$ different regression tasks, a framework we refer to as "multi-task" regression, and where the performance of the estimators is measured by the fixed-design quadratic risk. Kernel ridge regression is a classical framework to work with and comes with a natural norm, which often has desirable properties (such as, for instance, links with regularity). This norm is also a natural "similarity measure" between the regression functions. Evgeniou et al. [EMP05] showed how to extend kernel ridge regression to a multi-task setting, by adding a regularization term that binds the regression functions along the different tasks together. One of the main questions that is asked is to assert whether the multi-task estimator has a lower risk than any single-task estimator. It was recently proved by Solnon et al. [SAB12] that a fully data-driven calibration of this procedure is possible, given some assumptions on the set of matrices used to regularize—which correspond to prior knowledge on the tasks. Under those assumptions, the estimator is showed to verify an *oracle inequality*, that is, its risk matches (up to constants) the best possible one, the *oracle risk*. Thus, it suffices to compare the oracle risks for the multi-task procedure and the single-task one to provide an answer to this question.

The multi-task regression setting, which could also be called "multivariate regression", has already been studied in different papers. It was first introduced by Brown and Zidek [BZ80] in the case of ridge regression, and then adapted by Evgeniou et al. [EMP05] in its kernel form. Another view of the meaning of "task similarity" is that the functions all share a few common features, and can be expressed by a similar regularization term. This idea was expressed in a linear set up (also known as group lasso) by Obozinski et al. [OWJ11] and Lounici et al. [LPTvdG11], in multiple kernel learning by Koltchinskii and Yuan [KY10] or in semi-supervised learning by Ando and Zhang [AZ05]. The kernel version of this was also studied [AEP08, JBV08], a convex relaxation leading to a trace norm regularization and allowing the calibration of parameters. Another point of view was brought by Ben-David and Schuller [BDS03], defining a multi-task framework in classification, two classification problems being similar if, given a group of permutations of the input set, a dataset of the

one can be permuted in a dataset of the other. They followed the analysis of Baxter [Bax00], which shows very general bounds on the risk of a multi-task estimator in a model-selection framework, the sets of all models reflecting the insight the statistician has on the multi-task setting.

Advantages of the multi-task procedure over the single task one were first shown experimentally in various situations by, for instance, Thrun and O'Sullivan [TO96], Caruana [Car97] or Bakker and Heskes [BH03]. For classification, Ben-David and Schuller [BDS03] compare upper bounds on multi-task and single-task classification errors, and showed that the multi-task estimator could, in some settings, need less training data to reach the same upper bounds. The low dimensional linear regression setting was analysed by Rohde and Tsybakov [RT11], who showed that, under sparsity assumptions, restricted isometry conditions and using the trace-norm regularization, the multi-task estimator achieves the rates of a single-task estimator with a np-sample. Liang et al. [LBBJ10] also obtained a theoretical criterion, applicable to the linear regression setting and unfortunately non observable, which tells when the multi-task estimator asymptotically has a lower risk than the lower one. A step was recently carried by Feldman et al. [FGF12] in a kernel setting where every function is estimated by a constant. They give a closed-form expression of the oracle for two tasks and run simulations to compare the risk of the multi-task estimator to the risk of the single-task estimator.

In this chapter we study the oracle multi-task risk and compare it to the oracle single-task risk. We then find situations where the multi-task oracle is proved to have a lower risk than the single-task oracle. This allows us to better understand which situation favors the multi-task procedure and which does not. After having defined our model (Section 4.2.1), we write down the risk of a general multi-task ridge estimator and see that it admits a convenient decomposition using two key elements: the mean of the tasks and the resulting variance (Section 4.3). This decomposition allows us to optimize this risk and get a precise estimation of the oracle risk, in settings where the ridge estimator is known to be minimax optimal (Section 4.4). We then explore several repartitions of the tasks that give the latter multi-task rates, study their single-task oracle risk (Section 4.5) and compare it to their respective multi-task rates. This allows us to discriminate several situations, depending whether the multi-task oracle either outperforms its single-task counterpart, underperforms it or whether both behave similarly (Section 4.6). We also show that, in the cases favorable to the multi-task oracle detailed in the previous sections, the estimator proposed by Solnon et al. [SAB12] behaves accordingly and achieves a lower risk than the single-task oracle (Section 4.7). We finally study settings where we can no longer explicitly study the oracle risk, by running simulations, and we show that the multi-task oracle continues to retain the same virtues and disadvantages as before (Section 4.8).

The notations used here are recapitulated at the end of the introduction (page 33)

4.2 Kernel ridge regression in a multi-task setting

We consider here that each task is treated as a kernel ridge-regression problem and we will then extend the single-task ridge-regression estimator in a multi-task setting.

4.2.1 Model and estimator

Let Ω be a set, $\mathcal{A}$ be a σ-algebra on Ω and $\mathbb{P}$ be a probability measure on $\mathcal{A}$. We observe $\mathcal{D}_n = (X_i, Y_i^1, \dots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$. For each task $j \in \{1, \dots, p\}$, $\mathcal{D}_n^j = (X_i, y_i^j)_{i=1}^n$ is a sample with distribution $\mathcal{P}^j$, whose first marginal distribution is $\mathcal{P}$, for which a simple regression problem has to be solved.

We assume that for every $j \in \{1, \dots, p\}$, $F^j \in L^2(\mathbb{P})$, Σ is a symmetric positive-definite matrix of size p such that the vectors $(\varepsilon_i^j)_{j=1}^p$ are independent and identically distributed (i.i.d.) with normal distribution $\mathcal{N}(0, \Sigma)$, with mean zero and covariance matrix Σ, and

$$\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, p\}, \ y_i^j = F^j(X_i) + \varepsilon_i^j \ . \tag{4.1}$$

We suppose here, for simplicity, that $\Sigma = \sigma^2 I_p$, with $\sigma^2 \in \mathbb{R}_+^\star$.

Remark 4.1. *This implies that the outputs of every task are independent, which slightly simplifies the setting but allow lighter calculations. It is to be noted, though, that the analysis carried afterwards can still take place without this assumption. This can be dealt by diagonalizing Σ, majoring the quantities of interest using the largest eigenvalue of Σ and minoring those quantities by its smallest eigenvalue. The comparisons shown in Section 4.6 are still valid, only being enlarged by the condition number of Σ. A fully data-driven estimation of Σ was proposed by Solnon et al. [SAB12].*

We consider here a fixed-design setting, that is, we consider the input points as fixed and want to predict the output of the functions F^j on those input points only. The analysis could be transfered to the random-design setting by using tools developped by Hsu et al. [HKZ11].

For an estimator $(\widehat{F}^1, \dots, \widehat{F}^p)$, the natural quadratic risk to consider is

$$\mathbb{E}\left[\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}(\widehat{F}^j(X_i) - F^j(X_i))^2 | (X_1, \dots, X_n)\right] \ .$$

For the sake of simplicity, all the expectations that follow will implicitly be written conditional on $(X_1, \dots, X_n)$. This corresponds to the fixed-design setting, which treats the input points as fixed.

Remark 4.2. *We will use the following notations from now on :*

$$f = \mathrm{vec}\left((f^j(X_i))_{i,j}\right), \ F^j = \mathrm{vec}\left((f^j(X_i))_{i=1}^n\right) \quad \textit{and} \quad y = \mathrm{vec}\left((Y_i^j)_{i,j}\right) \ ,$$

so that, when using such vectorized notations, the elements are stacked task by task, the elements refering to the first task always being stored in the first entries of the vector, and so on.

We want to estimate f using elements of a particular function set. Let $\mathcal{F} \subset L^2(\mathbb{P})$ be a reproducing kernel Hilbert space (RKHS) [Aro50], with kernel k and feature map $\Phi : \mathcal{X} \to \mathcal{F}$, which give us the positive semidefinite kernel matrix $K = (k(X_i, X_\ell))_{1 \le i, \ell \le n} \in \mathcal{S}_n^+(\mathbb{R})$.

As done by Solnon et al. [SAB12] we extend the multi-task estimators generalizing the ridge-regression used in Evgeniou et al. [EMP05]. Given a positive-definite matrix $M \in \mathcal{S}_p^{++}(\mathbb{R})$, we consider the estimator

$$\widehat{F}_M \in \operatorname*{argmin}_{g\in\mathcal{F}^p} \left\{ \underbrace{\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2}_{\text{Empirical risk}} + \underbrace{\sum_{j=1}^{p}\sum_{\ell=1}^{p} M_{j,l}\langle g^j, g^\ell\rangle_{\mathcal{F}}}_{\text{Regularization term}} \right\} . \tag{4.2}$$

This leads to the fixed-design estimator

$$\widehat{f}_M = A_M y \in \mathbb{R}^{np} ,$$

with

$$A_M = A_{M,K} := \widetilde{K}_M(\widetilde{K}_M + npI_{np})^{-1} = (M^{-1}\otimes K)\left((M^{-1}\otimes K) + npI_{np}\right)^{-1} ,$$

where $\otimes$ denotes the Kronecker product (see the textbook of Horn and Johnson [HJ91] for simple properties of the Kronecker product).

Remark 4.3. *This setting also captures the single-task setting. Taking $j \in \{1,\dots,p\}$, $f^j = (f^j(X_1),\dots,f^j(X_n))^\top$ being the target-signal for the jth task and $y^j = (y_1^j,\dots,y_n^j)^\top$ being the observed output of the jth task, the single-task estimator for the jth task becomes (for $\lambda \in \mathbb{R}_+$)*

$$\widehat{f}_\lambda^j = A_\lambda y^j = K(K + n\lambda I_n)^{-1} y^j .$$

4.2.2 Two regularization terms for one problem

A common hypothesis that motivates the use of multi-task estimators is that all the target functions of the different tasks lie in a single cluster (that is, the p functions that are estimated are all close with respect to the norm defined on $\mathcal{F}$). Two different regularization terms are usually considered in this setting:

– one that penalizes the norms of the p function and their differences, introduced by Evgeniou et al. [EMP05], leading to the criterion (with $(g^j)_{j=1}^p \in \mathcal{F}^p$, $(\alpha,\beta) \in (\mathbb{R}_+)^2$)

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2 + \frac{\alpha}{p}\sum_{j=1}^{p}\left\|g^j\right\|_{\mathcal{F}}^2 + \frac{\beta}{2p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left\|g^j - g^k\right\|_{\mathcal{F}}^2 ; \tag{4.3}$$

– one that penalizes the norms of the average of the p functions and the resulting variance, leading to the criterion (with $(g^j)_{j=1}^p \in \mathcal{F}^p$, $(\lambda,\mu) \in (\mathbb{R}_+)^2$)

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2 + \lambda\left\|\frac{\sum_{j=1}^p g^j}{p}\right\|_{\mathcal{F}}^2 + \mu\left[\frac{\sum_{j=1}^p \left\|g^j\right\|_{\mathcal{F}}^2}{p} - \left\|\frac{\sum_{j=1}^p g^j}{p}\right\|_{\mathcal{F}}^2\right] . \tag{4.4}$$

As we will see, those two penalties are closely related. Lemma 4.1 indeed shows that the two former penalties can be obtained as a special case of Equation (4.2), the matrix M being respectively

$$M_{\mathrm{SD}}(\alpha,\beta) := \frac{\alpha}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\alpha + p\beta}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right)$$

and

$$M_{\mathrm{AV}}(\lambda,\mu) := \frac{\lambda}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) \ .$$

Thus, we see that those two criteria are related, since $M_{\mathrm{SD}}(\alpha,\beta) = M_{\mathrm{AV}}(\alpha,\alpha+p\beta)$ for every (α,β). Minimizing Equations (4.3) and (4.4) over $\mathcal{F}^p$ respectly give the ridge estimators $\widehat{f}_{SD}(\alpha,\beta) = A_{M_{\mathrm{SD}}(\alpha,\beta)}Y$ and $\widehat{f}_{\mathrm{AV}}(\lambda,\mu) = A_{M_{\mathrm{AV}}(\lambda,\mu)}Y$.

Remark 4.4. *We can now see that the regularization terms used in Equations* (4.3) *and* (4.4) *are equivalent when the parameters are not constrained to be positive. However, if one desires to use the regularization* (4.3) *(that is, with $\lambda = \alpha$ and $\mu = \alpha + p\beta$) and seeks to calibrate those parameters by taking them to be nonnegative (which is to be expected if they are seen as regularization parameters), the following problems could occur:*

- *if the optimization is carried over (λ,μ), then the selected parameter $\beta = \frac{\mu-\lambda}{p}$ may be negative;*
- *conversely, if the risk of the estimator defined by Equation* (4.3) *is optimized over the parameters $(\alpha,\alpha+p\beta)$ with the constraints $\alpha \geq 0$ and $\beta \geq 0$, then the infimum over $\mathbb{R}^2_+$ could never be approached.*

We will also show in the next section that the risk of $\widehat{f}_{\mathrm{AV}}(\lambda,\mu)$ nicely decomposes in two parts, the first part depending only on λ and the second only on μ, which is not the case for $\widehat{f}_{SD}(\alpha,\beta)$ because of the aforementioned phenomenon. This makes us prefer the second formulation and use the matrices M_{AV} instead of the matrices M_{SD}.

4.3 Decomposition of the risk

A fully data-driven selection of the hyper-parameters was proposed by Arlot and Bach [AB11], for the single-task ridge estimator, and by Solnon et al. [SAB12] for the multi-task estimator. The single-task estimator is shown to have a risk which is close to the single-task oracle-risk (with a fixed-design)

$$\mathfrak{R}^\star_{\mathrm{ST}} = \inf_{(\lambda^1,\dots,\lambda^p)\in\mathbb{R}^p_+} \left\{ \frac{1}{np}\mathbb{E}\left[\sum_{j=1}^p \left\| \widehat{f}^j_{\lambda^j} - f^j \right\|_2^2\right]\right\} \ ,$$

while the multi-task estimator is shown to have a risk which is close to the multi-task oracle risk

$$\mathfrak{R}^\star_{\mathrm{MT}} = \inf_{(\lambda,\mu)\in\mathbb{R}^2_+} \left\{ \frac{1}{np}\mathbb{E}\left[\left\| \widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)} - f \right\|_2^2\right]\right\} \ .$$

The purpose of this paper is to closely study both oracle risks and, ultimately, to compare them. We show in this section how to decompose the risk of an estimator obtained by minimizing Equation (4.4) over $(g^j)_{j=1}^p \in \mathcal{F}^p$. A key point of this analysis is that the matrix $M_{\mathrm{AV}}(\lambda,\mu)$ naturally decomposes over two orthogonal vector-subspaces of $\mathbb{R}^p$. By exploiting this decomposition we can simply use the classical bias-variance decomposition to analyse the Euclidean risk of those linear estimators.

4.3.1 Eigendecomposition of the matrix $M_{\mathrm{AV}}(\lambda, \mu)$

In this section we show that all the matrices $M_{\mathrm{AV}}(\lambda, \mu)$ have the same eigenvectors, which gives us a simple decomposition of the matrices $A_{M_{\mathrm{AV}}(\lambda,\mu)}$. Let us denote by $(e_1, \dots, e_p)$ the canonical basis of $\mathbb{R}^p$. The eigenspaces of $p^{-1}\mathbf{1}\mathbf{1}^\top$ are orthogonal and correspond to:

- span $\{e_1 + \cdots + e_p\}$ associated to eigenvalue 1,
- span $\{e_2 - e_1, \dots, e_p - e_1\}$ associated to eigenvalue 0.

Thus, with $(\lambda, \mu) \in (R^+)^2$, we can diagonalize in an orthonormal basis any matrix $M_{\mathrm{AV}}(\lambda, \mu)$ as $M = M_{\mathrm{AV}}(\lambda, \mu) = P^\top D_{\frac{\lambda}{p}, \frac{\mu}{p}} P$, with $D = \mathrm{Diag}\{\frac{\lambda}{p}, \frac{\mu}{p}, \dots, \frac{\mu}{p}\} = D_{\frac{\lambda}{p}, \frac{\mu}{p}}$. Let us also diagonalize K in an orthonormal basis : $K = Q^\top \Delta Q$, $\Delta = \mathrm{Diag}\{\gamma_1, \dots, \gamma_n\}$. Then

$$A_M = A_{M_{\mathrm{AV}}(\lambda,\mu)} = (P^\top \otimes Q^\top) \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right] (P \otimes Q) \ .$$

We can then note that $(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1}$ is a diagonal matrix, whose diagonal entry of index $(j-1)n + i$ ($i \in \{1, \dots, n\}$, $j \in \{1, \dots, p\}$) is

$$\begin{cases} \frac{\gamma_i}{\gamma_i + n\lambda} \text{ if } j = 1 \ , \\ \frac{\gamma_i}{\gamma_i + n\mu} \text{ if } j > 1 \ . \end{cases}$$

In the following section we will use the following notations :

- for every $j \in \{1, \dots, p\}$, $(h_i^j)_{i=1}^n$ denotes the coordinates of $(f^j(X_i))_{i=1}^n$ in the basis that diagonalizes K,
- for every $i \in \{1, \dots, n\}$, $(\nu_i^j)_{j=1}^p$ denotes the coordinates of $(h_i^j)_{j=1}^p$ in the basis that diagonalizes M.

Or, to sum up, we have :

$$\forall j \in \{1, \dots, p\}, \ \begin{pmatrix} h_1^j \\ \vdots \\ h_n^j \end{pmatrix} = Q \begin{pmatrix} f^j(X_1) \\ \vdots \\ f^j(X_n) \end{pmatrix}$$

and

$$\forall i \in \{1, \dots, n\}, \ \begin{pmatrix} \nu_i^1 \\ \vdots \\ \nu_i^p \end{pmatrix} = P \begin{pmatrix} h_i^1 \\ \vdots \\ h_i^p \end{pmatrix} \ .$$

With the usual notation $\nu^j = (\nu_1^j, \dots, \nu_n^j)^\top$ and f , we get, by using elementary properties of the Kronecker product,

$$\nu = \begin{pmatrix} \nu^1 \\ \vdots \\ \nu^p \end{pmatrix} = (P \otimes Q) f \ .$$

4.3.2 Bias-variance decomposition

We now use a classical bias-variance decomposition of the risk of $\widehat{f}_{\mathrm{AV}}(\lambda, \mu)$ and show that the quantities introduced above allow a simple expression of this risk. For any matrix

4.3. DECOMPOSITION OF THE RISK

$M \in \mathcal{S}_p^{++}(\mathbb{R})$, the classical bias-variance decomposition for the linear estimator $\widehat{f}_M = A_M y$ is

$$\begin{aligned}\frac{1}{np}\mathbb{E}\left[\left\|\widehat{f}_M - f\right\|_2^2\right] &= \frac{1}{np}\|(A_M - I_{np})f\|_2^2 + \frac{1}{np}\operatorname{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \\ &= \underbrace{\frac{1}{np}\|(A_M - I_{np})f\|_2^2}_{\text{Bias}} + \underbrace{\frac{\sigma^2}{np}\operatorname{tr}(A_M^\top A_M)}_{\text{Variance}} \quad .\end{aligned}$$

We can now compute both bias and variance of the estimator $\widehat{f}_{\text{AV}}(\lambda, \mu)$ by decomposing $A_{M_{\text{AV}}(\lambda,\mu)}$ on the eigenbasis introduced in the previous section.

$np\times$**Variance :**

$$\begin{aligned}& \sigma^2 \operatorname{tr}(A_M^\top A_M) \\ =\ & \sigma^2 \operatorname{tr}\left((P\otimes Q)^\top \left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta) + npI_{np}\right)^{-1}\right]^2 (P\otimes Q)\right) \\ =\ & \sigma^2 \operatorname{tr}\left(\left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta) + npI_{np}\right)^{-1}\right]^2\right) \\ =\ & \sigma^2 \sum_{i=1}^n \left[\left(\frac{\gamma_i}{\gamma_i + n\lambda}\right)^2 + (p-1)\left(\frac{\gamma_i}{\gamma_i + n\mu}\right)^2\right] \quad .\end{aligned}$$

$np\times$**Bias :**

$$\begin{aligned}& \|(A_M - I_{np})f\|_2^2 \\ =\ & \|(P\otimes Q)^\top \left[(D^{-1}\otimes K)\left((D^{-1}\otimes K) + npI_{np}\right)^{-1} - I_{np}\right](P\otimes Q)f\|_2^2 \\ =\ & \|\left[(D^{-1}\otimes\Delta)\left((D^{-1}\otimes\Delta) + npI_{np}\right)^{-1} - I_{np}\right]\nu\|_2^2 \\ =\ & (n\lambda)^2 \sum_{i=1}^n \frac{(\nu_i^1)^2}{(\gamma_i + n\lambda)^2} + (n\mu)^2 \sum_{i=1}^n \sum_{j=2}^p \frac{(\nu_i^j)^2}{(\gamma_i + n\mu)^2} \\ =\ & (n\lambda)^2 \sum_{i=1}^n \frac{(\nu_i^1)^2}{(\gamma_i + n\lambda)^2} + (n\mu)^2 \sum_{i=1}^n \frac{\sum_{j=2}^p (\nu_i^j)^2}{(\gamma_i + n\mu)^2} \quad .\end{aligned}$$

Thus, the risk of $\widehat{f}_{\text{AV}}(\lambda, \mu)$ becomes

$$\begin{aligned} n\lambda^2 \sum_{i=1}^n \frac{\frac{(\nu_i^1)^2}{p}}{(\gamma_i + n\lambda)^2} + \frac{\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\lambda}\right)^2 \\ + n\mu^2 \sum_{i=1}^n \frac{\frac{\sum_{j=2}^p (\nu_i^j)^2}{p}}{(\gamma_i + n\mu)^2} + \frac{\sigma^2(p-1)}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\mu}\right)^2 \quad . \end{aligned} \tag{4.5}$$

This decomposition has two direct consequences:

– the oracle risk of the multi-task procedure can be obtained by optimizing Equation (4.5) independently over λ and μ;

– the estimator $\widehat{f}_{\mathrm{AV}}$ can be calibrated by independently calibrating two parameters.

It is now easy to optimize over the quantities in Equation (4.5). An interesting fact is that both sides have a natural and interesting interpretation, which we give now.

4.3.3 Remark

To avoid further ambiguities and to simplify the formulas we introduce the following notations for every $i \in \{1, \dots, n\}$:

$$\mu_i = \nu_i^1 = \frac{h_i^1 + \cdots + h_i^p}{\sqrt{p}}$$

and

$$\varsigma_i^2 = \frac{\sum_{j=1}^p (h_i^j)^2}{p} - \left(\frac{\sum_{j=1}^p h_i^j}{p}\right)^2 = \frac{1}{p}\sum_{j=1}^p \left(h_i^j - \frac{\sum_{j=1}^p h_i^j}{p}\right)^2 \ ,$$

so that

$$p\varsigma_i^2 = \sum_{j=2}^p (\nu_i^j)^2 \ .$$

Remark 4.5. *We can see that for every $i \in \{1, \dots, n\}$, $\mu_i/\sqrt{p}$ is the average of the p target functions f^j, expressed on the basis diagonalizing K. Likewise, ς_i^2 can be seen as the variance between the p target functions f^j (which does not come from the noise).*

Henceforth, the risk of $\widehat{f}_{\mathrm{AV}}(\lambda, \mu)$ over (λ, μ) is decoupled into two parts.

– With the parameter λ, a part which corresponds to the risk of a single-task ridge estimator, which regularizes the mean of the tasks functions, with a noise variance σ^2/p:

$$n\lambda^2 \sum_{i=1}^n \frac{\frac{\mu_i^2}{p}}{(\gamma_i + n\lambda)^2} + \frac{\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\lambda}\right)^2 \ . \tag{4.6}$$

– With the parameter μ, a part which corresponds to the risk of a single-task ridge estimator, which regularizes the variance of the tasks functions, with a noise variance $(p-1)\sigma^2/p$:

$$n\mu^2 \sum_{i=1}^n \frac{\varsigma_i^2}{(\gamma_i + n\mu)^2} + \frac{(p-1)\sigma^2}{np}\sum_{i=1}^n \left(\frac{\gamma_i}{\gamma_i + n\mu}\right)^2 \ . \tag{4.7}$$

Remark 4.6. *Our analysis can also be used on any set of positive semi-definite matrices $\mathcal{M}$ that are jointly diagonalizable on an orthonormal basis, as was $\left\{M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}_+^2\right\}$. The element of interest then becomes the norms of the projections of the input tasks on the different eigenspaces (here, the mean and the resulting variance of the p tasks). An example of such a set is when the tasks are known to be split into several clusters, the assignement of each task to its cluster being known to the statistician. The matrices that can be used then regularize the mean of the tasks and, for each cluster, the variance of the tasks belonging to this cluster.*

4.4 Precise analysis of the multi-task oracle risk

In the latter section we showed that, in order to obtain the multi-task risk, we just had to optimize several functions, which have the form of the risk of a kernel ridge estimator. The risk of those estimators has already been widely studied. Johnstone [Joh94] (see also the article of Caponnetto and De Vito [CDV07] for random design) showed that, for a single-task ridge estimator, if the coefficients of the decomposition of the input function on the eigenbasis of the kernel decrease as $i^{-2\delta}$, with $2\delta > 1$, then the minimax rates for the estimation of this imput function is of order $n^{1/2\delta-1}$. The kernel ridge estimator is then known to be minimax optimal, under certain regularity assumptions (see the work of Bach [Bac13] for more details). If the eigenvalues of the kernel are known to decrease as $i^{-2\beta}$, then a single-task ridge estimator is minimax optimal under the following assumption:

$$1 < 2\delta < 4\beta + 1 \ . \qquad (\mathbf{H_M}(\beta, \delta))$$

The analysis carried in the former section shows that the key elements to express this risk are the components of the average of the signals (μ_i) and the components of the variance of the signals (ς_i) on the basis that diagonalises the kernel matrix K, together with the eigenvalues of this matrix (γ_i). It is then natural to impose the same natural assumptions that make the single-task ridge estimator optimal on those elements. We first suppose that the eigenvalues of the kernel matrix have a polynomial decrease rate:

$$\forall i \in \{1, \dots, n\},\ \gamma_i = n i^{-2\beta} \ . \qquad (\mathbf{H_K}(\beta))$$

Then, we assume that the components of the average of the signals and the variance of the signals also have a polynomial decrease rate:

$$\forall i \in \{1, \dots, n\},\ \left\{ \begin{array}{rcl} \frac{\mu_i^2}{p} & = & C_1 n i^{-2\delta} \\ \varsigma_i^2 & = & C_2 n i^{-2\delta} \end{array} \right. . \qquad (\mathbf{H_{AV}}(\delta, C_1, C_2))$$

Remark 4.7. *We assume for simplicity that both Assumptions* $(\mathbf{H_K}(\beta))$ *and* $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ *hold in equality, although the equivalence* $\asymp$ *is only needed.*

Example 4.1. *This example, related to Assumptions* $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ *and* $(\mathbf{H_K}(\beta))$ *by taking* $\beta = m$ *and* $2\delta = k + 2$*, is detailed by Wahba [Wah90] and by Gu [Gu02]. Let* $\mathcal{P}(2\pi)$ *the set of all square-integrable* 2π*-periodic functions on* $\mathbb{R}$*,* $m \in \mathbb{N}^\star$ *and define* $\mathcal{H} = \left\{ f \in \mathcal{P}(2\pi),\ f^{(m)}_{|[0,2\pi]} \in L^2[0, 2\pi] \right\}$*. This set* $\mathcal{H}$ *has a RKHS structure, with a reproducing kernel having the Fourier base functions as eigenvectors. The* i*-th eigenvalue of this kernel is* i^{-2m}*. For any function* $f \in \mathcal{P}[0, 2\pi] \cap \mathcal{C}^k[0, 2\pi]$*, then its Fourier coefficient are* $O(i^{-k})$*. For instance, if* $f \in \mathcal{P}[0, 2\pi]$ *such that* $\forall x \in [-\pi, \pi],\ f^{(k)}(x) = |x|$*, then its Fourier coefficients are* $\asymp i^{-(k+2)}$*.*

Under Assumptions ($\mathbf{H_K}(\beta)$) and ($\mathbf{H_{AV}}(\delta, C_1, C_2)$), we can now more precisely express the risk of a multi-task estimator. Equation (4.6) thus becomes

$$\begin{aligned} & n\lambda^2 \sum_{i=1}^{n} \frac{\frac{\mu_i^2}{p}}{(\gamma_i + n\lambda)^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{\gamma_i}{\gamma_i + n\lambda} \right)^2 \\ = \; & n\lambda^2 \sum_{i=1}^{n} \frac{C_1 n i^{-2\delta}}{(n i^{-2\beta} + n\lambda)^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{n i^{-2\beta}}{n i^{-2\beta} + n\lambda} \right)^2 \\ = \; & C_1 \lambda^2 \sum_{i=1}^{n} \frac{i^{4\beta - 2\delta}}{(1 + \lambda i^{2\beta})^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1 + \lambda i^{2\beta})^2} \\ = \; & R(n, p, \sigma^2, \lambda, \beta, \delta, C_1) \ , \end{aligned}$$

while Equation (4.7) becomes

$$\begin{aligned} & n\mu^2 \sum_{i=1}^{n} \frac{\varsigma_i^2}{(\gamma_i + n\mu)^2} + \frac{(p-1)\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{\gamma_i}{\gamma_i + n\mu} \right)^2 \\ = \; & n\mu^2 \sum_{i=1}^{n} \frac{C_2 n i^{-2\delta}}{(n i^{-2\beta} + n\mu)^2} + \frac{(p-1)\sigma^2}{np} \sum_{i=1}^{n} \left(\frac{n i^{-2\beta}}{n i^{-2\beta} + n\mu} \right)^2 \\ = \; & C_2 \mu^2 \sum_{i=1}^{n} \frac{i^{4\beta - 2\delta}}{(1 + \mu i^{2\beta})^2} + \frac{(p-1)\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1 + \mu i^{2\beta})^2} \\ = \; & R(n, p, (p-1)\sigma^2, \mu, \beta, \delta, C_2) \ , \end{aligned}$$

with

$$R(n, p, \sigma^2, x, \beta, \delta, C) = C x^2 \sum_{i=1}^{n} \frac{i^{4\beta - 2\delta}}{(1 + x i^{2\beta})^2} + \frac{\sigma^2}{np} \sum_{i=1}^{n} \frac{1}{(1 + x i^{2\beta})^2} \ . \tag{4.8}$$

Remark 4.8. *It is to be noted that the function R corresponds to the risk of a single-task ridge estimator when the decomposition of the input function on the eigenbasis of K has $i^{-2\delta}$ for coefficients and when $p = 1$. It has two terms, which corresponds to the bias-variance decomposition performed in Section 4.3.2, page 91. Thus, studying R will allow us to derive both single-task and multi-task oracle rates.*

4.4.1 Study of the optimum of $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$

We just showed that the function $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$ was suited to derive both single-task and multi-task oracle risk. Bach [Bac13] showed how to obtain a majoration on the function $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$, so that its infimum was showed to match the minimax rates under Assumption ($\mathbf{H_M}(\beta, \delta)$).

In this section, we first propose a slightly more precise upper bound of this risk function. We then show how to obtain a lower bound on this infimum that matches the aforementioned upper bound. This will be done by precisely localizing the parameter minimizing $R(n, p, \sigma^2, \cdot, \beta, \delta, C)$.

Let us first introduce the following notation:

Definition 4.1.

$$R^\star(n,p,\sigma^2,\beta,\delta,C) = \inf_{\lambda\in\mathbb{R}_+} \left\{R(n,p,\sigma^2,\lambda,\beta,\delta,C)\right\} \ .$$

We now give the upper bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$. For simplicity, we will denote by $\kappa(\beta,\delta)$ a constant, defined in Equation (4.23), which only depends on β and δ.

Property 4.1. *Let n and p be positive integers, σ, β and δ positive real numbers such that $(\mathbf{H_M}(\beta,\delta))$, $(\mathbf{H_K}(\beta))$ and $(\mathbf{H_{AV}}(\delta,C_1,C_2))$ hold. Then,*

$$R^\star(n,p,\sigma^2,\beta,\delta,C) \leq \left(2^{1/2\delta}\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} C^{1/2\delta}\kappa(\beta,\delta)\right) \wedge \frac{\sigma^2}{p} \ . \tag{4.9}$$

Proof. Property 4.1 is proved in Section 4.C of the appendix. □

In the course of showing Property 4.1, we obtained an upper bound on the risk function R that holds uniformly on $\mathbb{R}_+$. Obtaining a similar (up to multiplicative constants) lower bound that also holds uniformly on $\mathbb{R}_+$ is unrealistic. However, we will be able to lower bound $R^\star$ by showing that R is minimized by an optimal parameter $\lambda^\star$ that goes to 0 as n goes to $+\infty$.

Property 4.2. *If Assumption $(\mathbf{H_M}(\beta,\delta))$ holds, the risk $R(n,p,\sigma^2,\cdot,\beta,\delta,C)$ attains its global minimum over $\mathbb{R}_+$ on $[0,\varepsilon\left(\frac{np}{\sigma^2}\right)]$, with*

$$\varepsilon\left(\frac{np}{\sigma^2}\right) = \sqrt{C^{(1/2\delta)-1}2^{1/2\delta}\kappa(\beta,\delta)} \times \frac{1}{\left(\frac{np}{\sigma^2}\right)^{1/2-(1/4\delta)}}\left(1+\eta\left(\frac{np}{\sigma^2}\right)\right) \ ,$$

where $\eta(x)$ goes to 0 as x goes to $+\infty$.

Proof. Property 4.2 is shown in Section 4.D of the appendix. □

Remark 4.9. *Thanks to the assumption made on δ, $\frac{1}{2\delta}-1<0$ so that $\left(\frac{np}{\sigma^2}\right)^{\frac{1}{2\delta}-1}$ goes to 0 as $\frac{np}{\sigma^2}$ goes to $+\infty$. This allows us to state that, if the other parameters are constant, $\lambda^\star$ goes to 0 as the quantity $\frac{np}{\sigma^2}$ goes to $+\infty$.*

We can now give a lower bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$. We will give two versions of this lower bound. First, we state a general result.

Property 4.3. *For every (C,β,δ) such that $1<2\delta<4\beta$ holds, there exits an integer N and a constant $\alpha\in(0,1)$ such that, for every for every (n,p,σ^2) verifying $\frac{np}{\sigma^2}\geq N$, we have*

$$R^\star(n,p,\sigma^2,\beta,\delta,C) \geq \left(\alpha\left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} C^{1/2\delta}\kappa(\beta,\delta)\right) \wedge \frac{\sigma^2}{4p} \ . \tag{4.10}$$

Proof. Property 4.3 is proved in Section 4.E.3 of the appendix. □

Remark 4.10. *It is to be noted that N and α only depend on β and δ. We can also remark that α can be taken arbitrarily close to*

$$\frac{\int_0^1 \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2}du} \wedge \frac{\int_0^1 \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du} \quad .$$

Numerical computations show that, by taking $\beta = \delta = 2$, this constant is larger than 0.33*.*

Remark 4.11. *The assumption made on β and δ is slighlty more restrictive than $(\mathbf{H_M}(\beta, \delta))$, under which the upper bound is shown to hold and under which the single-task estimator is shown to be minimax optimal.*

We are now ensured that R attains its global minimum on $\mathbb{R}_+$, thus we can give the following definition.

Definition 4.2. For every n, p, σ^2, δ, β and C, under the assumption of Property 4.2, we introduce

$$\lambda_R^\star \in \operatorname*{argmin}_{\lambda \in \mathbb{R}_+} \left\{ R(n, p, \sigma^2, \lambda, \beta, \delta, C) \right\} \quad .$$

We now give a slightly refined version of Property 4.3, by discussing whether this optimal parameter $\lambda_R^\star$ is larger or lower than the threshold $n^{-2\beta}$. This allows us to better understand the effect of regularizarion on the oracle risk $R^\star$.

Property 4.4. *For every (β, δ) such that $4\beta > 2\delta > 1$, integers N_1 and N_2 exist such that*

1. *for every (n, p, σ^2) verifying $\frac{np}{\sigma^2} \geq N_1$ and $n^{1-2\delta} \times \frac{p}{\sigma^2} \leq \frac{1}{N_2}$, then*

$$\lambda_R^\star \geq \frac{1}{n^{2\beta}}$$

and

$$R^\star(n, p, \sigma^2, \beta, \delta, C) \asymp \left(\frac{\sigma^2}{np}\right)^{1-1/2\delta} \quad .$$

2. *for every (n, p, σ^2) verifying $\frac{np}{\sigma^2} \geq N_1$ and $n^{1-2\delta} \times \frac{p}{\sigma^2} \geq N_2$, then*

$$\lambda_R^\star \leq \frac{1}{n^{2\beta}}$$

and

$$R^\star(n, p, \sigma^2, \beta, \delta, C) \asymp R(n, p, \sigma^2, 0, \beta, \delta, C) \asymp \frac{\sigma^2}{p} \ ;$$

Proof. Property 4.4 is proved in Section 4.E.4 of the appendix. □

Remark 4.12. *If $p \leq n\sigma^2$ and $\delta > 1$ then we are in the first case, for a large enough n. This is a case where regularization has to be employed in order to obtain optimal convergence rates. This also comes as a simple consequence of Properties*

Remark 4.13. *If σ^2 and n are fixed and p goes to $+\infty$ then we are in the second case. It is then useless to regularize the risk, since the risk can only be lowered by a factor* 4*, which comes from Properties 4.3 and 4.8. This also corresponds to a single-task setting where the noise variance σ^2 is very small and where the estimation problem becomes trivial.*

4.4.2 Multi-task oracle risk

We can now use the upper and lower bounds on $R^\star$ to control the oracle risk of the multi-task estimator. We define

$$\lambda^\star \in \operatorname*{argmin}_{\lambda \in \mathbb{R}_+} \left\{ R(n, p, \sigma^2, \lambda, \beta, \delta, C_1) \right\}$$

and

$$\mu^\star \in \operatorname*{argmin}_{\mu \in \mathbb{R}_+} \left\{ R(n, p, (p-1)\sigma^2, \mu, \beta, \delta, C_2) \right\} \ .$$

Property 4.2 ensures that $\lambda^\star$ and $\mu^\star$ exist, even though they are not necessarily unique. The oracle risk then is

$$\mathfrak{R}^\star_{\mathrm{MT}} = \inf_{(\lambda,\mu)\in\mathbb{R}_+^2} \left\{ \frac{1}{np} \mathbb{E}\left[\left\| \widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)} - f \right\|_2^2 \right] \right\} = \frac{1}{np} \mathbb{E}\left[\left\| \widehat{f}_{M_{\mathrm{AV}}(\lambda^\star,\mu^\star)} - f \right\|_2^2 \right] \ .$$

We now state the main result of this paper, which simply comes from the analysis of $R^\star$ performed above.

Theorem 4.1. *For every n, p, C_1, C_2, σ^2, β and δ such that Assumption $(\mathbf{H_M}(\beta, \delta))$ holds, we have*

$$\mathfrak{R}^\star_{\mathrm{MT}} \leq 2^{1/2\delta} \left(\frac{np}{\sigma^2} \right)^{1/2\delta - 1} \kappa(\beta, \delta) \left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta} \right] \ . \tag{4.11}$$

Furthermore, constants N and $\alpha \in (0,1)$ exist such that, if $n \geq N$, $p/\sigma^2 \leq n$ and $2 < 2\delta < 4\beta$, we have

$$\mathfrak{R}^\star_{\mathrm{MT}} \geq \alpha \left(\frac{np}{\sigma^2} \right)^{1/2\delta - 1} \kappa(\beta, \delta) \left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta} \right] \ . \tag{4.12}$$

Proof. The risk of the multi-task estimator $\widehat{f}_{M_{\mathrm{AV}}(\lambda,\mu)}$ can be written as

$$R(n, p, \sigma^2, \lambda, \beta, \delta, C_1) + R(n, p, (p-1)\sigma^2, \mu, \beta, \delta, C_2) \ .$$

We then apply Properties 4.1 and 4.3, since $p/\sigma^2 \leq n$ implies that $p/(p-1)\sigma^2 \leq n$. The assumption $\delta > 1$ ensures that the first setting of Property 4.4 holds. □

Remark 4.14. *An interesting fact is that the oracle multi-task risk is of the order $(np/\sigma^2)^{1/2\delta-1}$. This corresponds to the risk of a single-task ridge estimator with sample size np.*

Remark 4.15. *As noted before, the assumption under which the lower bound holds is slightly stronger than Assumption $(\mathbf{H_M}(\beta, \delta))$.*

4.5 Single-task oracle risk

In the former section we obtained a precise approximation of the multi-task oracle risk $\mathfrak{R}^\star_{\mathrm{MT}}$. We would now like to obtain a similar approximation for the single-task oracle risk $\mathfrak{R}^\star_{\mathrm{ST}}$. In the light of Section 4.3, the only element we need to obtain the oracle risk of task $j \in \{1, \ldots, p\}$ is the expression of $(h_i^j)_{i=1}^n$, that is, the coordinates of $(f^j(X_i))_{i=1}^n$ on the

eigenbasis of K. Unfortunately, Assumption $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ does not correspond to one set of task functions $(f^1, \dots, f^p)$. Thus, since several single-task settings can lead to the same multi-task oracle risk, we now explicitly define two repartitions of the task functions $(f^1, \dots, f^p)$, for which the single-task oracle risk will be computed.

– "2 points": suppose, for simplicity, that p is even and that

$$f^1 = \cdots = f^{p/2} \quad \text{and} \quad f^{p/2+1} = \cdots = f^p \ . \tag{2Points}$$

– "1 outlier":

$$f^1 = \cdots = f^{p-1} \ . \tag{1Out}$$

Both assumptions correspond to settings in which the multi-task procedure would legitimately be used. Assumption (2Points) models the fact that all the functions lie in a cluster of small radius. It supposes that the functions are split into two groups of equal size, in order to be able to explicitly derive the single-task oracle risk. Assumption (1Out) supposes that all the functions are grouped in one cluster, with one outlier. In order to make the calculations possible, all the functions in one group are assumed to be equal. Since this is not a fully convincing situation to study the behaviour of the multi-task oracle, simulation experiments were also run on less restrictive settings. The results of those experiments are shown in Section 4.8.

Remark 4.16. *The hypotheses* (2Points) *and* (1Out) *made on the functions* f^j *can be expressed on* (h_i^j)*. Assumption* (2Points) *becomes*

$$\forall i \in \{1, \dots, n\}, \ h_i^1 = \cdots = h_i^{p/2} \quad \textit{and} \quad h_i^{p/2+1} = \cdots = h_i^p \ ,$$

while Assumption (1Out) *becomes*

$$\forall i \in \{1, \dots, n\}, \ h_i^1 = \cdots = h_i^{p-1} \ .$$

Under those hypotheses we now want to derive an expression of $(h_i^1, \dots, h_i^p)$ given (μ_i, ς_i) so that we can exactly compute the single-task oracle risk. Remember we defined for every $i \in \{1, \dots, n\}$,

$$\mu_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} h_i^j$$

and

$$\varsigma_i^2 = \frac{1}{p} \sum_{j=1}^{p} \left(h_i^j\right)^2 - \frac{\mu_i^2}{p} = \frac{1}{p} \sum_{j=1}^{p} \left(h_i^j - \frac{\mu_i}{\sqrt{p}}\right)^2 \ .$$

We also re-introduce the single-task oracle risk:

$$\mathfrak{R}_{\mathrm{ST}}^{\star} = \inf_{(\lambda^1, \dots, \lambda^p) \in \mathbb{R}_+^p} \left\{ \frac{1}{np} \sum_{j=1}^{p} \mathbb{E}\left[\left\| \widehat{f}_{\lambda^j}^j - f^j \right\|_2^2 \right] \right\} \ .$$

We now want to closely study this single-task oracle risk, in both settings.

4.5.1 Analysis of the oracle single-task risk for the "2 points" case (2Points)

In this section we write the single-task oracle risk when Assumption (2Points) holds. As shown in Lemma 4.8, the risk of the estimator $\widehat{f}_\lambda^j = A_\lambda y^j$ for the jth task, which we denote by $R^j(\lambda)$, verifies

$$R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}-\sqrt{C_2}\right)^2) \leq R^j(\lambda) \leq R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}+\sqrt{C_2}\right)^2) \ .$$

Both upper and lower parts eventually behave similarly. In order to simplify notations and to avoid having to constantly write two risks, we will assume that half of the tasks have a risk equal to the right-hand side of the later inequality and the other half a risk equal to the left-hand side of this inequality. This leads to the following assumption:

$$\forall i \in \{1,\dots,n\}, \left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta}(\sqrt{C_1}+\sqrt{C_2}) \\ h_i^p & = & \sqrt{n} i^{-\delta}(\sqrt{C_1}-\sqrt{C_2}) \end{array} \right. . \qquad (\mathbf{H_{2Points}})$$

This minor change does not affect the convergence rates of the estimator. Consequently, if $1 \leq j \leq p/2$ the risk for task j is $R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}+\sqrt{C_2}\right)^2)$ so that the oracle risk for task j is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \times \left(\sqrt{C_1}+\sqrt{C_2}\right)^{1/\delta} \ ,$$

and if $p/2+1 \leq j \leq p$ the risk for task j is $R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1}-\sqrt{C_2}\right)^2)$ so that the oracle risk for task j is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta,\delta) \times \left|\sqrt{C_1}-\sqrt{C_2}\right|^{1/\delta} \ ,$$

Remark 4.17. *We can remark that* ($\mathbf{H_{2Points}}$) *implies* (2Points) *and that* ($\mathbf{H_{2Points}}$) *implies* ($\mathbf{H_{AV}}(\delta,C_1,C_2)$)*, as shown in Lemma 4.10. Consequently, if* ($\mathbf{H_{2Points}}$) *holds, we have, for every* $i \in \{1,\dots,n\}$, $h_i^1 = \frac{\mu_i}{\sqrt{p}} + \varsigma_i$ *and* $h_i^p = \frac{\mu_i}{\sqrt{p}} - \varsigma_i$.

Corollary 4.1. *For every* n, p, C_1, C_2, σ^2, β *and* δ *such that* $2 < 2\delta < 4\beta$ *and* $n\sigma^2 > 1$ *and that Assumptions* ($\mathbf{H_{2Points}}$) *and* ($\mathbf{H_K}(\beta)$) *hold, then*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} \frac{\kappa(\beta,\delta)}{2} \times p^{1-1/2\delta} \left[\left(\sqrt{C_1}+\sqrt{C_2}\right)^{1/\delta} + \left|\sqrt{C_1}-\sqrt{C_2}\right|^{1/\delta}\right] \ . \quad (4.13)$$

4.5.2 Analysis of the oracle single-task risk for the "1 outlier" case (1Out)

In this section we suppose that Assumption (1Out) holds. As shown in Lemma 4.9, we can lower and upper bound the risks of the single-tasks estimators by functions of the shape $R(n,p,\sigma^2,\lambda,\beta,\delta,C)$. As in the latter section, to avoid the burden of writing two long risk terms at every step, and since all those risks have the same convergence rates, we suppose from now on the new assumption:

$$\forall i \in \{1,\dots,n\} \left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1}+\frac{1}{\sqrt{p-1}}\sqrt{C_2}\right) \\ h_i^p & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1}-\sqrt{p-1}\sqrt{C_2}\right) \end{array} \right. . \qquad (\mathbf{H_{1Out}})$$

This minor change does not affect the convergence rates of the estimator. Consequently, if $1 \leq j \leq p-1$ the risk for task j is $R(n, 1, \sigma^2, \lambda, \beta, \delta, \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^2)$ so that the oracle risk for task j is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta, \delta) \times \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^{1/\delta} ,$$

while the risk for task p is $R(n, 1, \sigma^2, \lambda, \beta, \delta, \left(\sqrt{C_1} - \sqrt{(p-1)C_2}\right)^2)$ so that the oracle risk for task p is, given that $n\sigma^2 \geq 1$,

$$\asymp \left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta, \delta) \times \left|\sqrt{C_1} - \sqrt{(p-1)C_2}\right|^{1/\delta} .$$

Remark 4.18. *We can also remark here that* $(\mathbf{H_{1Out}})$ *implies* (1Out) *and that* $(\mathbf{H_{1Out}})$ *implies* $(\mathbf{H_{AV}}(\delta, C_1, C_2))$*, as shown in Lemma 4.9. Consequently, if* $(\mathbf{H_{1Out}})$ *holds, we have, for every* $i \in \{1, \ldots, n\}$*,* $h_i^1 = \frac{\mu_i}{\sqrt{p}} + \frac{1}{\sqrt{p-1}}\varsigma_i$ *and* $h_i^p = \frac{\mu_i}{\sqrt{p}} - \sqrt{p-1}\varsigma_i$*.*

Corollary 4.2. *For every* n*,* p*,* C_1*,* C_2*,* σ^2*,* β *and* δ *such that* $2 < 2\delta < 4\beta$ *and* $n\sigma^2 > 1$ *and that Assumptions* $(\mathbf{H_{1Out}})$ *and* $(\mathbf{H_K}(\beta))$ *hold, then*

$$\mathfrak{R}^\star_{\mathrm{ST}} \asymp \left(\frac{np}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta, \delta) \times p^{1-1/2\delta} \left[\frac{p-1}{p}\left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|\sqrt{C_1} - \sqrt{(p-1)C_2}\right|^{1/\delta}\right] . \quad (4.14)$$

4.6 Comparison of multi-task and single-task

In the two latter section we obtained precise approximations of the multi-task oracle risk, $\mathfrak{R}^\star_{\mathrm{MT}}$, and of the single-task oracle risk, $\mathfrak{R}^\star_{\mathrm{ST}}$, under either Assumption $(\mathbf{H_{2Points}})$ or $(\mathbf{H_{1Out}})$. We can now compare both risks in either setting, by studying their ratio

$$\rho = \frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} .$$

We will express the quantity ρ as a factor of

$$r = \frac{C_2}{C_1} .$$

The parameter r controls the amount of the signal which is contained in the mean of the functions. When r is small, the mean of the tasks contains much more signal than the variance of the tasks, so that the tasks should be "similar". This is a case where the multi-task oracle is expected to perform better than the single-task oracle. On the contrary, when r is large, the variance of the tasks is more important than the mean of the tasks. This is a case where the tasks would be described as "non-similar". It is then harder to conjecture whether the single-task oracle performs better than the multi-task oracle and, as we will see later, the answer to this greatly depends on the setting.

4.6.1 Analysis of the oracle multi-task improvement for the "2 points" case (2Points)

We now express ρ as a function of r when the tasks are split in two groups.

Corollary 4.3. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > p$ and that Assumptions ($\mathbf{H_{2Points}}$) and ($\mathbf{H_K}(\beta)$) hold, then*

$$\rho \asymp \frac{p^{1/2\delta-1} + (\frac{p-1}{p})^{1-(1/2\delta)} r^{1/2\delta}}{(1+\sqrt{r})^{1/\delta} + |1-\sqrt{r}|^{1/\delta}} \ . \tag{4.15}$$

Remark 4.19. *The right-hand side of Equation (4.15) is always smaller than $\frac{1}{2}$. Thus, under the assumptions of Corollary 4.3, the multi-task oracle risk can never be arbitrarily worse than the single-task oracle risk.*

We can first see that, under the assumptions of Corollary 4.3, $\rho = \Theta\left(p^{1/2\delta-1}\right)$ as r goes to 0. This is the same improvement that we get we multiplying the sample-size by p. We also have $\rho = \Theta\left(\left(\frac{p-1}{p}\right)^{1-(1/2\delta)}\right)$ as r goes to $+\infty$, so that the multi-task oracle and the single-task oracle behave similarly. Finally, $\rho = \Theta\left(\frac{r^{1/2\delta}}{\left(1+\sqrt{r}\right)^{1/\delta}+\left|1-\sqrt{r}\right|^{1/\delta}}\right)$ as p goes to $+\infty$, so that the behaviours we just discussed are still valid with a large number of tasks.

4.6.2 Analysis of the oracle multi-task improvement for the "1 outlier" case (1Out)

We now express ρ as a function of r when the tasks are grouped in one group, with one outlier.

Corollary 4.4. *For every n, p, C_1, C_2, σ^2, β and δ such that $2 < 2\delta < 4\beta$ and $n\sigma^2 > p$ and that Assumptions ($\mathbf{H_{1Out}}$) and ($\mathbf{H_K}(\beta)$) hold, then*

$$\rho \asymp \frac{p^{1/2\delta-1} + \left(\frac{p-1}{p}\right)^{1-(1/2\delta)} r^{1/2\delta}}{\frac{p-1}{p}\left(1+\sqrt{\frac{r}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|1-\sqrt{r(p-1)}\right|^{1/\delta}} \ . \tag{4.16}$$

We can see that, under the assumptions of Corollary 4.4, $\rho = \Theta\left(p^{1/2\delta-1}\right)$ as r goes to 0. As in the latter section, this is the same improvement that we get we multiplying the sample-size by p. However, $\rho = \Theta\left(\left(\frac{p-1}{p}\right)^{1-1/2\delta} \times \frac{p(p-1)^{-1/2\delta}}{1+(p-1)^{1-1/\delta}}\right)$ as r goes to $+\infty$. This quantity goes to $+\infty$ as $p \longrightarrow +\infty$, so that the multi-task oracle performs arbitrarily worse than the single-task one in this asymptotic setting. Finally, $\rho = \Theta\left(r^{1/2\delta}\right)$ as p goes to $+\infty$. This quantity goes to $+\infty$ as r goes to $+\infty$, so that the behaviours we just mentioned stay valid with a large number of tasks.

4.6.3 Discussion

When r is small, either under Assumption (2Points) or (1Out), the mean of the signal is much stronger than the variance. Thus, the multi-task procedure performs better than the single-task one.

Example 4.2. *If $r = 0$, then all the tasks are equal. The improvement of the multi-task procedure over the single-task one then is $p^{1/2\delta-1}$. This was expected: it corresponds to the risk of a ridge regression with a np-sample.*

As r goes to 0, the multi-task oracle outperforms its single-task counterpart by a factor $p^{1/2\delta-1}$. When p is large (but, remember, this only holds when $p/\sigma^2 \leq n$, so n also has to be large), this leads to a substantial improvement. It is easily seen that, for any constant $C > 1$, if $r \leq (C-1)^{2\delta}(p-1)^{1-2\delta}$, then the right-hand side of Equation (4.15) becomes smaller than $Cp^{1/2\delta-1}$. Thus, if the tasks are similar enough, the multi-task oracle performs as well as the oracle for a np-sample, up to a constant.

On the contrary, when r is large, the variance carries most of the signal, so that the tasks differ one from another. As r goes to $+\infty$, the two settings have different behaviours:

- under Assumption (2Points) (that is, when we are faced to two equally-sized groups), the oracle risks of the multi-task and of the single-task estimators are of the same order: they can only differ by a multiplicative constant;
- under Assumption (1Out) (that is, when we are faced to one cluster and one outlier), the single-task oracle outperforms the multi-task one, by a factor which is approximatly $p^{1/\delta}$.

Finally, Assumption (2Points) presents no drawback for the multi-task oracle, since under those hypotheses its performance cannot be worse than the single-task oracle's one. On the contrary, Assumption (1Out) presents a case where the use of a multi-task technique greatly increases the oracle risk, when the variance between the tasks is important, while it gives an advantage to the multi-task oracle when this variance is small. The location where the multi-task improvement stops corresponds to the barrier $\rho = 1$. Studying this object seems difficult, since we only know ρ up to a multiplicative constant. Also, finding the contour lines of the righ-hand side of Equation (4.16) does not seem to be an easy task. In Section 4.8, we will run simulations in situations where the oracle risk can no longer be explicitly derived. We will show that the behaviours found in these two examples still appear in the simulated examples.

4.7 Risk of a multi-task estimator

Solnon et al. [SAB12] introduced an entirely data-driven estimator to calibrate $M_{\mathrm{AV}}(\lambda, \mu)$ over $\mathbb{R}_+^2$. One of their main results is an oracle inequality, that compares the risk of this estimator to the oracle risk. Thus, $\mathfrak{R}^\star_{\mathrm{MT}}$ is attainable by a fully data-driven estimator. We now show that our estimation of the multi-task oracle risk is precise enough so that we can use it in the mentionned oracle inequality and still have a lower risk than the single-task oracle one.

The following assumption will be used, with $\mathrm{df}(\lambda) = \mathrm{tr}(A_\lambda)$ and $A_\lambda = K(K + n\lambda I_n)^{-1}$:

$$\left.\begin{array}{l} \forall j \in \{1, \dots, p\}\,, \exists \lambda_{0,j} \in (0, +\infty)\,, \\ \qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{and} \quad \dfrac{1}{n}\left\|(A_{\lambda_{0,j}} - I_n) f^j\right\|_2^2 \leq \sigma^2 \sqrt{\dfrac{\ln n}{n}} \end{array}\right\} \qquad \textbf{(Hdf)}$$

We will also denote $\mathcal{M} = \left\{M_{\mathrm{AV}}(\lambda, \mu),\ (\lambda, \mu) \in \mathbb{R}_+^2\right\}$ and $\widehat{M}_{\mathrm{HM}}$ the estimator introduced in Solnon et al. [SAB12], which belongs to $\mathcal{M}$. Theorem 29 of Solnon et al. [SAB12] thus states:

Theorem 4.2. *Let $\alpha = 2$, $\theta \geq 2$, $p \in \mathbb{N}^\star$ and assume* (**Hdf**) *holds true. An absolute constant $L > 0$ and a constant $n_1(\theta)$ exist such that the following holds as soon as $n \geq n_1(\theta)$.*

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}_{HM}} - f\right\|_2^2\right] \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \mathbb{E}\left[\inf_{M \in \mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2\right\}\right] \\ + L\sigma^2(2+\theta)^2 p \frac{\ln(n)^3}{n} + \frac{p}{n^{\theta/2}}\frac{\|f\|_2^2}{np}\ . \end{aligned} \tag{4.17}$$

We first remark that

$$\mathbb{E}\left[\inf_{M \in \mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2\right\}\right] \leq \mathfrak{R}^\star_{\mathrm{MT}}\ .$$

We can now plug the oracle risk in the oracle inequality (4.17). Then, if we suppose that, for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$, $(h_i^j)^2 = nC^j i^{-2\delta}$, we have that

$$\|f\|_2^2 = \sum_{j=1}^p \sum_{i=1}^n (h_i^j)^2 = n \sum_{j=1}^p C^j \sum_{i=1}^n i^{-2\delta} \leq n\zeta(2\delta) \sum_{j=1}^p C^j\ .$$

Remark 4.20. *Assumption* (2Points) *means that for every $i \in \{1, \dots, n\}$, if $1 \leq j \leq p/2$,*

$$C^j = \left(\sqrt{C_1} + \sqrt{C_2}\right)^2$$

and if $p/2 + 1 \leq j \leq p$,

$$C^j = \left(\sqrt{C_1} - \sqrt{C_2}\right)^2\ .$$

Assumption (1Out) *means that for every $i \in \{1, \dots, n\}$, if $1 \leq j \leq p - 1$,*

$$C^j = \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^2$$

while

$$C^p = \left(\sqrt{C_1} - \sqrt{(p-1)C_2}\right)^2\ .$$

Property 4.5. *Under Assumptions* ($\mathbf{H_K}(\beta)$) *and* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$) *with $2\delta > 2$, there exists a constant N_1 such that for every $n \geq N_1$, Assumption* (**Hdf**) *holds.*

Proof. We can see that Assumption (**Hdf**) is made independently on every task. Thus we can suppose that $p = 1$. Let us denote $b(\lambda) = n^{-1} \|(A_\lambda - I_n)f\|_2^2$. We can see that if there exists constants $c > 0$ and $d > 1$ such that for every $\lambda \in \mathbb{R}_+$ $b(\lambda) \leq c\sigma^2 \operatorname{df}(\lambda)^{-d}$, then Assumption (**Hdf**) holds for n large enough. Indeed, let $\lambda \in \mathbb{R}_+$ such that $\operatorname{df}(\lambda) \leq \sqrt{n}$. Then, if $b(\lambda) \leq c\sigma^2 \operatorname{df}(\lambda)^{-d}$, $b(\lambda) \leq \sigma^2 c(\sqrt{n})^{-d} \leq \sigma^2 c \frac{n^{(-d+1)/2}}{\sqrt{n}}$. It just suffices to see that, for n large enough, $cn^{-d+1} \leq \ln(n)$.

Using Lemmas 4.6 and 4.5 we can see that, for every $\lambda \in \mathbb{R}_+$,

$$b(\lambda) \leq \frac{\lambda^{\frac{2\delta-1}{2\beta}}}{\beta} I_1(\beta, \delta)$$

and, for n large enough, there exists a constant α such that, for every $\lambda \in \mathbb{R}_+$,

$$\operatorname{df}(\lambda) = \operatorname{tr} A_\lambda \geq \alpha \frac{\lambda^{\frac{-1}{2\beta}}}{2\beta} I_2(\beta)$$

Thus, for n large enough, there exists a constant c (depending on σ^2, β and δ) such that, for every $\lambda \in \mathbb{R}_+$,

$$b(\lambda) \leq c\sigma^2 \operatorname{tr}(A_\lambda)^{-(2\delta-1)} \quad .$$

Hence, if $2\delta > 2$, there exists a constant N_1 such that for every $n \geq N_1$, Assumption (**Hdf**) holds. □

Thus, we can apply Theorem 4.2 to the estimator $\widehat{f}_{\widehat{M}_{\mathrm{HM}}}$ under either Assumption (2Points) or (1Out) (and we denote by ρ either $\rho_{2Points}$ or ρ_{1Out}).

Property 4.6. *For every positive numbers $(\beta, \delta, \theta, C_1, C_2)$ verifying $4\beta > 2\delta > 2$ and $\theta > 1$, there exists positive constants $(N(\beta, \delta, \theta), L)$ such that, for every (n, p, σ^2) verifying $n \geq N$ and $\frac{p}{\sigma^2} \leq n$, if Assumption ($\mathbf{H_K}(\beta)$) and if either Assumption ($\mathbf{H_{2Points}}$) or Assumption ($\mathbf{H_{1Out}}$) hold, the ratio between the risk of the estimator $\widehat{f}_{\widehat{M}_{HM}}$ and the single-task oracle risk verifies*

$$\frac{\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}_{HM}} - f\right\|_2^2\right]}{\mathfrak{R}^\star_{\mathrm{ST}}} \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \rho + Cst \times \frac{L\sigma^2(2+\theta)^2 p \frac{\ln(n)^3}{n} + \frac{p\zeta(2\delta)}{n^{\theta/2}} \frac{1}{p}\sum_{j=1}^p C^j}{\left(\frac{n}{\sigma^2}\right)^{1/2\delta-1} \kappa(\beta, \delta) \times \frac{1}{p}\sum_{j=1}^p (C^j)^{1/2\delta}} \quad .$$

Proof. This is a straightforward application of the preceding results. □

We now show that the latter fully data-driven multi-task ridge estimator achieves a lower risk than the single-task ridge oracle, in both settings (2Points) and (1Out).

Corollary 4.5. *For every positive numbers $(\beta, \delta, \theta, \sigma^2, \varepsilon)$ verifying $4\beta > 2\delta > 2$ and $\theta > 2$, there exists positive constants (N, r) such that, for every (n, p, C_1, C_2) verifying $n \geq N$, $\frac{p}{\sigma^2} \leq n^{1/4\delta}$ and $\frac{C_2}{C_1} \leq r$, if Assumptions ($\mathbf{H_K}(\beta)$) holds and if either Assumption ($\mathbf{H_{2Points}}$)*

or Assumption ($\mathbf{H_{1Out}}$) *hold, the ratio between the risk of the estimator* $\widehat{f}_{\widehat{M}_{HM}}$ *and the single-task oracle risk verifies*

$$\frac{\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}_{HM}} - f\right\|_2^2\right]}{\mathfrak{R}^\star_{\mathrm{ST}}} < \varepsilon \ .$$

Proof. First, we can see that under either Assumption (2Points) or Assumption (1Out), both $\frac{1}{p}\sum_{j=1}^p C^j$ and $\frac{1}{p}\sum_{j=1}^p (C^j)^{1/2\delta}$ converge, as p goes to $+\infty$, to quantities only depending on C_1, C_2 and δ and are thus bounded with respect to p. Then, as it was shown in the previous section, both $\rho_{2Points}$ and ρ_{1Out} go to 0 as $\frac{C_1}{C_2}$ goes to 0. Finally, we can see that $\frac{p}{\sigma^2} \leq n^{1/4\delta}$ implies that $\frac{p}{\sigma^2} \leq n$ and that

$$\frac{\sigma^2 p \frac{\ln(n)^3}{n}}{\left(\frac{n}{\sigma^2}\right)^{1/2\delta-1}} = \left(\sigma^2\right)^{1/2\delta} \times p \times \frac{\ln(n)^3}{n^{1/2\delta}} \leq \left(\sigma^2\right)^{1+1/2\delta} \times \frac{\ln(n)^3}{n^{1/4\delta}} \underset{n\to+\infty}{\longrightarrow} 0$$

together with

$$\frac{\frac{p}{n^{\theta/2}}}{\left(\frac{n}{\sigma^2}\right)^{1/2\delta-1}} \leq \left(\sigma^2\right)^{1/2\delta} \times n^{1-\theta/2-1/4\delta} \underset{n\to+\infty}{\longrightarrow} 0 \ .$$

□

Remark 4.21. *The result shown in Corollary* (4.5) *establishes that a fully data-driven multi-task estimator outperforms an oracle single-task estimator, which is minimax optimal* .

4.8 Numerical experiments

The hypotheses we used in the former sections, although sufficient to precisely derive the risk of the estimator, do not reflect realistic situations. In this section we study less restrictive settings. However, we are no longer able to obtain simple formulas for the oracle risk as we did before. Thus, we resort to numerical simulations to illustrate the behaviour of both single-task and multi-task oracles.

4.8.1 Setting A: relaxation of Assumptions ($\mathbf{H_{AV}}(\delta, C_1, C_2)$) and (2Points) in order to get one general group of tasks

In the latter sections we modeled the fact that the p target functions are close. However, due to technical constraints we were only able to deal with cases where the functions are split into two groups and are then equal inside each group, thus introducing Assumptions (2Points) and (1Out). We propose here to extend this setting by simulating a more general group of tasks. Those tasks should all be at a comparable distance from a centroid function.

We suppose that $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$ is a sequence of i.i.d. random variables, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution (that is, such that $\mathbb{P}(\varepsilon_i^j = 1) =$

$\mathbb{P}(\varepsilon_i^j = -1) = 1/2)$. The target functions are then defined by

$$\forall i \in \{1, \dots, n\},\ \forall j \in \{1, \dots, p\},\ h_i^j = \sqrt{n} i^{-\delta} \left(\sqrt{C_1} + \varepsilon_i^j \sqrt{C_2} \right) \ . \tag{4.18}$$

Thus, all the p target functions are "close" to a centroid function, whose coordinates on the eigenvectors of the kernel matrix are $\sqrt{n} i^{-\delta} \sqrt{C_1}$, with a "dispersion factor" $\sqrt{C_2}$. In this setting, we can easily express the key elements for the analysis of this risk :

$$\frac{\mu_i^2}{p} = n i^{-2\delta} \left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p} \sqrt{C_2} \right)^2$$

and

$$\varsigma_i^2 = n i^{-2\delta} \left(\frac{1}{p} \sum_{j=1}^p \left(\sqrt{C_1} + \varepsilon_i^j \sqrt{C_2} \right)^2 - \left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p} \sqrt{C_2} \right)^2 \right) \ .$$

Remark 4.22. *The theoretical analysis developed previously cannot be applied here, due to the presence of random terms, which depend on i, in front of the decay term $n i^{-2\delta}$.*

4.8.2 Setting B: random drawing of the input points and functions

Assumptions $(\mathbf{H_K}(\beta))$ and $(\mathbf{H_{AV}}(\delta, C_1, C_2))$ model the behaviour of the spectral elements of f and K as if they exactly follow the spectral elements of the kernel operator and the input function. Although convenient for the analysis, this setting is unlikely to hold in practice and we propose here to draw the input points $(X_i)_{i=1}^n$ and compute the risk using the eigenvalues of the kernel matrix.

We suppose here that $(X_i)_{i=1}^n$ is a sequence of i.i.d. random variables uniformly drawn on $[-\pi, \pi]$. As in the latter section, we also suppose that we have an i.i.d. sequence of random variables $(\varepsilon_i^j)_{i \in \{1,\dots,n\}, j \in \{1,\dots,p\}}$, independent of $(X_i)_{i \in \{1,\dots,n\}}$, following a Rademacher distribution. The target functions are then defined by

$$\forall i \in \{1, \dots, n\},\ \forall j \in \{1, \dots, p\},\ f^j(X_i) = \left(\sqrt{C_1} + \varepsilon_i^j \sqrt{C_2} \right) |X_i| \ . \tag{4.19}$$

As stated in Wahba [Wah90] and in Gu [Gu02], a natural kernel to use is, with $m \in \mathbb{N}^\star$,

$$R(x, y) = 2 \sum_{i=1}^{+\infty} \frac{\cos\left(i(x-y)\right)}{i^{2m}} \ .$$

In this setting, the coefficients of the decomposition of $f : x \mapsto |x|$ on the Fourier basis are known to be asymptotically equivalent to i^{-2}. Thus, this setting is a natural extension of Assumptions $(\mathbf{H_K}(\beta))$ and $(\mathbf{H_{AV}}(\delta, C_1, C_2))$, with $\beta = m$—since the eigenvalues of the kernel R are i^{-2m}—and $\delta = 2$.

4.8.3 Setting C: further relaxation of Assumptions ($\mathbf{H_{AV}}(\delta, C_1, C_2)$) and (2Points) in one group of tasks

We consider the same tasks than in Setting A, but also allow the regularity of the variance to vary. This gives the following model, supposing that $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$ is a sequence of i.i.d. random variables, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution:

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\},\ h_i^j = \sqrt{n}\left(\sqrt{C_1}i^{-\delta_1} + \varepsilon_i^j\sqrt{C_2}i^{-\delta_2}\right) \ .$$

We allow the variance to have a varying regularity and intensity by changing C_2 and δ_2. This gives us the following quantities of interest: for every $i \in \{1,\dots,n\}$,

$$\frac{\mu_i^2}{p} = ni^{-2\delta_1}\left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p}\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2$$

and

$$\varsigma_i^2 = ni^{-2\delta_1}\left(\frac{1}{p}\sum_{j=1}^p\left(\sqrt{C_1} + \varepsilon_i^j\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2 - \left(\sqrt{C_1} + \frac{\sum_{j=1}^p \varepsilon_i^j}{p}\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2\right) \ .$$

4.8.4 Setting D: relaxation of Assumptions (1Out) and ($\mathbf{H_{AV}}(\delta, C_1, C_2)$)

Assumption (1Out) states that we have one of $p-1$ identical tasks and one outlier. We now simulate a slightly more general setting by having one cluster of $p-1$ around 0 and an outlier. This gives the following model, supposing that $(\varepsilon_i^j)_{i\in\{1,\dots,n\},j\in\{1,\dots,p\}}$ is a sequence of i.i.d. random variables, independent of $(X_i)_{i\in\{1,\dots,n\}}$, following a Rademacher distribution:

$$\forall i \in \{1,\dots,n\}\,,\ \forall j \in \{1,\dots,p-1\}\,,\ h_i^j = \sqrt{n}\varepsilon_i^j i^{-2}$$

and

$$\forall i \in \{1,\dots,n\}\,,\quad h_i^p = \sqrt{nC_2}\varepsilon_i^p i^{-\delta_2} \ .$$

We allow the outlier to have a varying regularity and intensity by changing C_2 and δ_2. This gives us the following quantities of interest: for every $i \in \{1,\dots,n\}$,

$$\frac{\mu_i^2}{p} = ni^{-\delta_1}\left(\frac{\sqrt{C_1}}{p}\sum_{j=1}^{p-1}\varepsilon_i^j + \frac{\varepsilon_i^p}{p}\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2$$

and

$$\varsigma_i^2 = ni^{-\delta_1}\left(\frac{p-1}{p}C_1 + \frac{1}{p}C_2 i^{-2(\delta_2-\delta_1)} - \left(\frac{1}{p}\sum_{j=1}^{p-1}\varepsilon_i^j + \frac{\varepsilon_i^p}{p}\sqrt{C_2}i^{-(\delta_2-\delta_1)}\right)^2\right) \ .$$

4.8.5 Methodology

In every setting, we computed the oracle risks of both the multi-task estimator and the single-task one. As shown before, for instance in Equation (4.5), both the multi-task risk (which has two hyper-parameters, λ and μ) and the single-task risk (which has p hyper-parameters, λ^1 to λ^p) can be decomposed as a sum of several functions, each depending on a unique hyper-parameter. We used Newton's method to optimize each of those $p+2$ functions over, respectively, λ, μ, $\lambda^1, \ldots, \lambda^p$. Our stopping criterion was that the derivative of the function being optimized was inferior to 10^{-5}, in absolute value. We replicated each experiment $N = 100$ times. This gives us N independent realisations of $(\mathfrak{R}^\star_{\mathrm{MT}}, \mathfrak{R}^\star_{\mathrm{ST}})$, the randomness coming from the repartition of the tasks and, in Setting B, from the drawing of the input points $(X_i)_{i=1}^n$.

In Settings A and B, we first test the hypothesis $\mathbb{H}_0 = \{\mathbb{P}(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}) < 0.5\}$ against $\mathbb{H}_1 = \{\mathbb{P}(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}) \geq 0.5\}$. This amounts to testing whether the median of $\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}$ is larger than one. For every iteration $i \in \{1, \ldots, N\}$, we observe $B_i = \mathbf{1}_{\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}}}$. Since the random variables $(B_i)_{i\in\{1,\ldots,N\}}$ follow a Bernouilli distribution of parameter $\mathbb{P}(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}})$, we can apply Hoeffding's inequality [Mas07] and see that, for every $\varepsilon > 0$, $[\bar{B}_N - \varepsilon, 1]$ is a confidence interval of level $1 - e^{-2N\varepsilon^2}$ for $\mathbb{P}(\mathfrak{R}^\star_{\mathrm{MT}} < \mathfrak{R}^\star_{\mathrm{ST}})$. This leads to the following p-value:

$$\pi_1 = \begin{cases} e^{-2N\left(\bar{B}_N - 0.5\right)^2} & \text{if } \bar{B}_N \geq 0.5 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

In those two settings, we also test the hypothesis $\mathbb{H}_0 = \left\{\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] > 1\right\}$ against $\mathbb{H}_1 = \left\{\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] \leq 1\right\}$. Let us denote by $\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$ the empirical mean of the random variables $\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}$, $\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$ the resulting standard deviation and Φ the cumulative distribution function of a standard gaussian distribution. Then, a classical use of the central limit theorem and of Slutsky's Lemma gives that

$$\left[0, \widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] + \frac{\varepsilon}{\sqrt{n}}\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]\right]$$

is an asymptotic confidence interval of level $\Phi(\varepsilon)$ for $\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$. This leads to the following asymptotic p-value:

$$\pi_2 = \Phi\left[\sqrt{n}\left(\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] - 1\right)\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]^{-1}\right] \ .$$

The results of those tests are shown in Table 4.1 for Setting A and in Table 4.2 for Setting B.

In Settings C and D, we use the same asymptotic framework and show error bars corresponding to the asymptotic confidence interval

$$\left[\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] - \frac{z_{0.975}}{\sqrt{n}}\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right], \widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right] + \frac{z_{0.975}}{\sqrt{n}}\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]\right]$$

of level 95%, where z_α denotes the quantile of order α of the standard gaussian distribution. The results of those simulations are shown in Figure 4.1 for Setting C and in Figure 4.2 for Setting D.

We used the following values for the parameters: $n = 50$, $p = 5$, $\sigma^2 = 1$ and $C_1 = 1$. We finally settled $\delta = 2$ in Settings A and B and $\delta_1 = 2$ in Settings C and D.

C_2	$r = \frac{C_2}{C_1}$	β	$\bar{B}_{100}$	π_1	$\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	$\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	π_2
0.01	0.01	2	1	$< 10^{-15}$	0.434	0.0324	$< 10^{-15}$
0.1	0.1	2	1	$< 10^{-15}$	0.672	0.0747	$< 10^{-15}$
0.5	0.5	2	0.94	$< 10^{-15}$	0.898	0.0913	$< 10^{-15}$
1	1	2	0.51	9.80×10^{-1}	1.01	0.129	0.773
5	5	2	0.38	1	0.998	0.0292	0.302
10	10	2	0.42	1	0.996	0.0172	9.90×10^{-3}
100	100	2	0.76	1.35×10^{-6}	0.997	5.44×10^{-3}	5.97×10^{-10}
0.01	0.01	4	1	$< 10^{-15}$	0.426	0.0310	$< 10^{-15}$
0.1	0.1	4	1	$< 10^{-15}$	0.703	0.0737	$< 10^{-15}$
0.5	0.5	4	0.75	3.73×10^{-6}	0.934	0.113	1.80×10^{-9}
1	1	4	0.31	1	1.08	0.163	1.00
5	5	4	0.38	1	1.01	0.0439	0.965
10	10	4	0.43	1	0.993	0.0304	0.0113
100	100	4	0.83	3.48×10^{-10}	0.992	0.0103	1.22×10^{-14}

Table 4.1: Comparison of the multi-task oracle risk to the single-task oracle risk in Setting A.

4.8.6 Interpretation

When all the tasks are grouped in one cluster (Settings A, B and C), the same phenomenon as under Assumption (2Points) appears. In situations where the mean component of the signal has more weight than the variance component (in Settings A and B, that is when r is small, in Setting C, this occurs when δ_2 is large and C_2 is small) then the multi-task oracle seems to outperform the single-task one. On the contrary, when the mean component of the signal is negligible compared to the variance component (likewise, this occurs in Settings A and B when r is large and in Setting C when δ_2 is small or when C_2 large), then both oracles seem to perform similarly.

Adversary settings to the multi-task oracle appear when one task is added outside of a cluster (Setting D). When this outlier is less regular than the tasks belonging to the cluster (that is, when δ_2 is large), the single-task oracle performs better than the multi-task one, which confirms the theoretical analysis performed in Section 4.6.2.

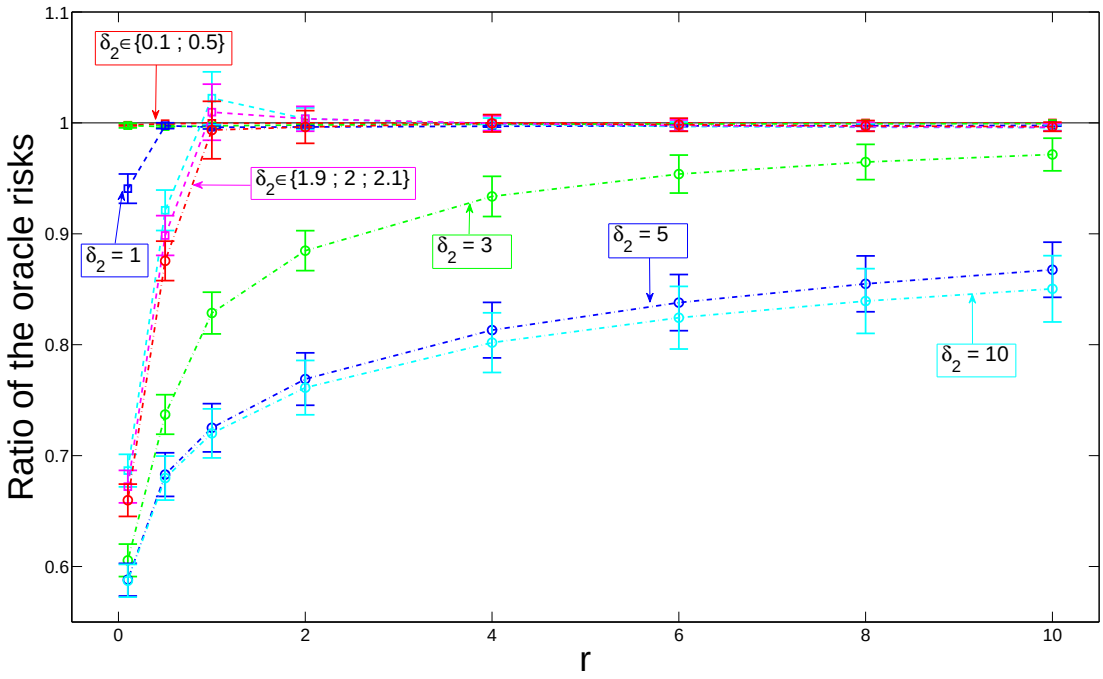

Figure 4.1: Further relaxation of Assumption (2Points) (Experiment C), improvement of multi-task compared to single-task: $\mathbb{E}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$. Best seen in colour.

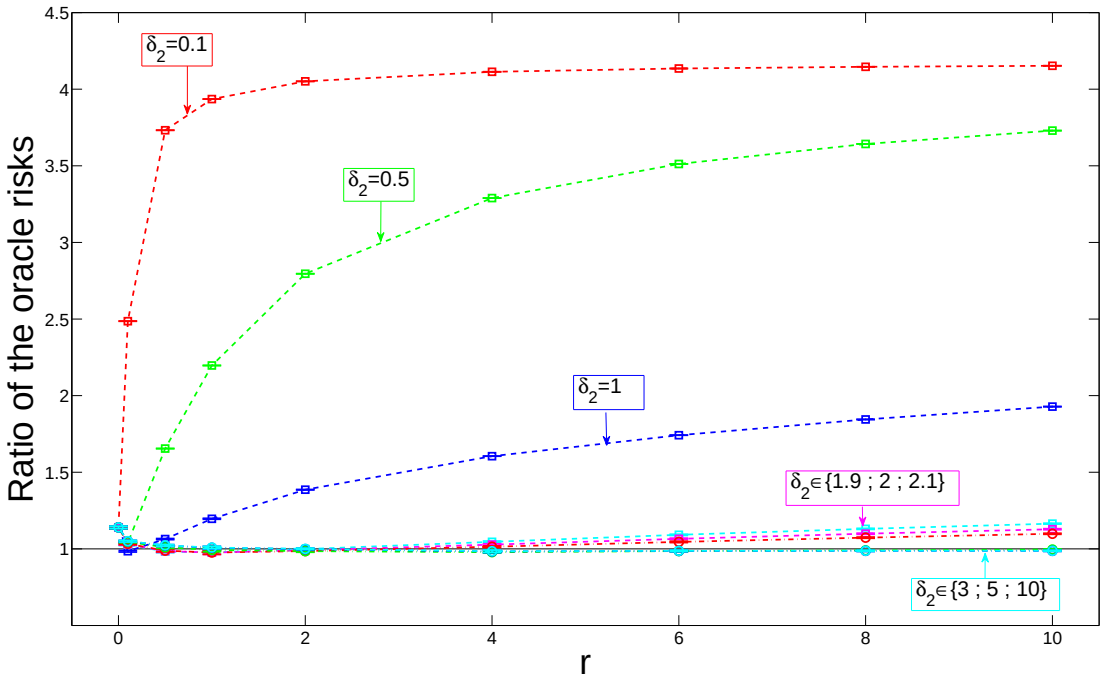

Figure 4.2: Relaxation of Assumption (1Out) (Experiment D), improvement of multi-task compared to single-task: $\mathbb{E}\left[\frac{\mathfrak{R}^{\star}_{\mathrm{MT}}}{\mathfrak{R}^{\star}_{\mathrm{ST}}}\right]$. Best seen in colour.

C_2	$r = \frac{C_2}{C_1}$	m	$\bar{B}_{100}$	π_1	$\widehat{\mathbb{E}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	$\widehat{\mathrm{Std}}\left[\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}}\right]$	π_2
0.01	0.01	2	1	$< 10^{-15}$	0.570	0.0409	$< 10^{-15}$
0.1	0.1	2	1	$< 10^{-15}$	0.745	0.0333	$< 10^{-15}$
0.5	0.5	2	0.99	$< 10^{-15}$	0.907	0.0406	$< 10^{-15}$
1	1	2	0.80	1.52×10^{-8}	0.961	0.0459	$< 10^{-15}$
5	5	2	0.55	0.607	0.995	0.205	2.59×10^{-3}
10	10	2	0.53	0.835	0.996	0.114	6.23×10^{-4}
100	100	2	0.81	4.50×10^{-9}	0.996	6.35×10^{-3}	1.03×10^{-11}
0.01	0.01	4	1	$< 10^{-15}$	0.527	0.0409	$< 10^{-15}$
0.1	0.1	4	1	$< 10^{-15}$	0.756	0.0534	$< 10^{-15}$
0.5	0.5	4	0.93	$< 10^{-15}$	0.917	0.0650	$< 10^{-15}$
1	1	4	0.49	1	1.01	0.0896	0.855
5	5	4	0.40	1	0.997	0.0295	0.170
10	10	4	0.41	1	0.998	0.0179	0.114
100	100	4	0.84	9.10×10^{-11}	0.994	8.71×10^{-3}	7.36×10^{-14}

Table 4.2: Comparison of the multi-task oracle risk to the single-task oracle risk in Setting B.

4.9 Conclusion

This paper shows the existence of situations where the multi-task kernel ridge regression, with a perfect parameter calibration, can perform better than the single-task one. This happens when the tasks are distributed given simple specifications, which are studied both theoretically and on simulated examples.

The analysis performed here allows us to have a precise estimation of the risk of the multi-task oracle (Theorem 4.1), this result holding under a few hypotheses on the regularity of the kernel, of the mean of the tasks and of its resulting variance. Several simple single-task settings are then investigated, with the constraint that they respect the latter assumptions. This theoretical grounding, backed-up by our simulated examples, allows us to understand better when and where the multi-task procedure outperforms the single-task one.

- The situation where all the regression functions are close in the RKHS (that is, their differences are extremely regular) is favorable to the multi-task procedure, when using the matrices $\mathcal{M} = \left\{ M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}_+^2 \right\}$. In this setting, the multi-task procedure can do much better than the single-task one (as if it had p times more input points). It is also shown to never do worse (up to a multiplicative constant) !
- On the contrary, when one outlier lies far apart from this cluster, this multi-task procedure suddenly performs badly, that is, arbitrarily worse than the single-task one. This comes as no surprise, since the addition of a far less regular task naturally destroys the joint learning of a group of tasks. In this case, the use of a multi-task procedure which clusters the tasks together (because of the choice of $\mathcal{M}$) is inadapted to the situation.

4.9. CONCLUSION

Our analysis can easily be adapted to a slightly wider set of assumptions on the tasks than the one presented here (all the tasks are grouped together, in one cluster). It is for instance possible to treat the case where the tasks are grouped in two (or more) clusters—when the allocation of each task to its cluster is known to the statistician, at the price of introducing more hyperparameters. We are still limited, though, to certain cases of hypotheses, reflected on the set of matricial hyperparameters $\mathcal{M}$. The failure of the multi-task oracle on the case where one outlier stays outside of one group of tasks can be seen, not as the impossibility to use multi-task techniques in this situation, but rather as the fact the set of matrices used here, $\mathcal{M} = \left\{ M_{\mathrm{AV}}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}_+^2 \right\}$, is inadapted to the situation. We can at least see two different solutions to this kind of inadaptation. First, the use of prior knowledge can help the statistician to craft an *ad hoc* set $\mathcal{M}$. Second, we could seek to automatically adapt to the situation in order to learn a good set $\mathcal{M}$ from data.

Learning more complex sets $\mathcal{M}$ is an important—but complex—challenge, that we want to address in the future. This question can at least be split into three (not necessarily independent) problems, that call for the elaboration of new tools:

- a careful study of the risk, to find a set $\mathcal{M}^\star \subset \mathcal{S}_p^{++}(\mathbb{R})$ of candidate matrices;
- optimization tools, to derive an algorithm able to select a matrix in this set $\mathcal{M}^\star$;
- new concentration of measure results, to be able to show oracle inequalities that control the risk of the output of the algorithm.

Our estimation of the multi-task oracle risk is also shown to be precise enough so that we can plug it in an oracle inequality, hereby showing the existence of a multi-task estimator that has a lower risk than the single-task oracle (under the same favorable circumstances as before).

Finally, it would be intereting to extend the analysis developped here to the random-design setting. This could be done, for instance, by using the tools brought by Hsu et al. [HKZ11], that link random-design convergence rates to fixed-design ones.

Appendices

4.A Decomposition of the matrices $M_{\mathrm{SD}}(\alpha, \beta)$ and $M_{\mathrm{AV}}(\lambda, \mu)$

We now give a few technical results that were used in the former sections.

Lemma 4.1. *The penalty used in Equation* (4.3) *can be obtained by using in Equation* (4.2) *the matrix* $M_{\mathrm{SD}}(\alpha, \beta)$*, such that*

$$M_{\mathrm{SD}}(\alpha, \beta) = \frac{\alpha}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\alpha + p\beta}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) \ . \tag{4.20}$$

The penalty used in Equation (4.4) *can be obtained by using in Equation* (4.2) *the matrix* $M_{\mathrm{AV}}(\alpha, \beta)$*, such that*

$$M_{\mathrm{AV}}(\lambda, \mu) = \frac{\lambda}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) \ . \tag{4.21}$$

Proof. For the first part, since

$$\begin{aligned}\sum_{j=1}^{p}\sum_{k=1}^{p}\left\|g^j - g^k\right\|_{\mathcal{F}}^2 &= \sum_{j,k}\langle g^j, g^j\rangle_{\mathcal{F}} - 2\langle g^j, g^k\rangle_{\mathcal{F}} + \langle g^k, g^k\rangle_{\mathcal{F}} \\ &= 2p\sum_{j=1}^{p}\langle g^j, g^j\rangle_{\mathcal{F}} - 2\sum_{j,k}\langle g^j, g^k\rangle_{\mathcal{F}} \ ,\end{aligned}$$

the penalty term of Equation (4.3) can be written as

$$\frac{\alpha}{p}\sum_{j=1}^{p}\langle g^j, g^j\rangle_{\mathcal{F}} + \beta\sum_{j=1}^{p}\langle g^j, g^j\rangle_{\mathcal{F}} - \frac{\beta}{p}\sum_{j,k}\langle g^j, g^k\rangle_{\mathcal{F}} \ ,$$

leading to the matrix

$$\frac{\alpha + p\beta}{p}I_p - \frac{\beta}{p}\mathbf{1}\mathbf{1}^\top = \frac{\alpha}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\alpha + p\beta}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) = M_{\mathrm{SD}}(\alpha, \beta) \ .$$

For the second part, since

$$\left\|\sum_{j=1}^{p} g^j\right\|_{\mathcal{F}}^2 = \sum_{j,k}\langle g^j, g^k\rangle_{\mathcal{F}} \ ,$$

the penalty term of Equation (4.4) can be written as

$$\frac{\lambda}{p^2}\sum_{j,k}\langle g^j, g^k\rangle_{\mathcal{F}} + \frac{\mu}{p}\sum_{j=1}^{p}\langle g^j, g^j\rangle_{\mathcal{F}} - \frac{\mu}{p^2}\sum_{j,k}\langle g^j, g^k\rangle_{\mathcal{F}} \ ,$$

leading to the matrix

$$\frac{\lambda-\mu}{p^2}\mathbf{1}\mathbf{1}^\top + \frac{\mu}{p}I_p = \frac{\lambda}{p}\frac{\mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p}\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}\right) = M_{\mathrm{AV}}(\lambda,\mu) \ .$$

□

4.B Useful control of some sums

Let us introduce the following integrals :

$$I_1 = I_1(\beta,\delta) = \int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du \ ,$$
$$I_2 = I_2(\beta) = \int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2}du = I_1(\beta,0) \ .$$

Under Assumption ($\mathbf{H_M}(\beta,\delta)$), both integrals converge. We also introduce their discrete counterparts. For every $n \in \mathbb{N}^\star$ and every $\lambda \in \mathbb{R}_+$:

$$S_1(n,\lambda) = \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \ ,$$
$$S_2(n,\lambda) = \sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} \ .$$

We here give a first elementary technical result.

Lemma 4.2. *The map defined on $\mathbb{R}_+$ by*

$$t \mapsto \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}$$

is positive, increasing on $[0,t^\star]$ and decreasing on $[t^\star,+\infty)$ to 0, with

$$t^\star = \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}$$

Proof. This map is nonnegative and converges to 0 in 0 and $+\infty$. Furthermore

$$\begin{aligned}\frac{d}{dt}\left(\frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}\right) &= (4\beta-2\delta)\frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^2} - 4\beta\lambda t^{2\beta-1}\frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^3} \\ &= \frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^3}\left[(4\beta-2\delta)(1+\lambda t^{2\beta}) - 4\beta\lambda t^{2\beta}\right] \\ &= \frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^3}\left[4\beta+4\beta\lambda t^{2\beta}-2\delta-2\delta\lambda t^{2\beta}-4\beta\lambda t^{2\beta}\right] \\ &= \frac{t^{4\beta-2\delta-1}}{(1+\lambda t^{2\beta})^3}\left[(4\beta-2\delta)-2\delta\lambda t^{2\beta}\right] \ .\end{aligned}$$

The only parameter $t^\star$ that cancels out this equation is

$$t^\star = \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \ .$$

□

We now give a serie of technical results to control I_1, I_2, S_1 and S_2, which will be useful in the following sections.

Lemma 4.3.

$$\int_0^{+\infty}\frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}dt = \frac{\lambda^{(2\delta-1)/2\beta}}{2\beta\lambda^2}\int_0^{+\infty}\frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2}du = \frac{\lambda^{(2\delta-1)/2\beta}}{2\beta\lambda^2}I_1 \ .$$

Proof. Apply the change of variables $u=\lambda t^{2\beta}$ see [Bac13] for more details. □

Lemma 4.4.

$$\int_0^{+\infty}\frac{1}{(1+\lambda t^{2\beta})^2}dt = \frac{\lambda^{-1/2\beta}}{2\beta}\int_0^{+\infty}\frac{u^{\frac{1-2\beta}{2\beta}}}{(1+u)^2}du = \frac{\lambda^{-1/2\beta}}{2\beta}I_2 \ .$$

Proof. Apply the change of variables $u=\lambda t^{2\beta}$ see [Bac13] for more details. □

Lemma 4.5. *We have the following bounds S_2. For every $n\in\mathbb{N}^\star$ and every $\lambda\in\mathbb{R}_+^\star$,*

– $$S_2(n,\lambda) \le \frac{\lambda^{-1/2\beta}}{2\beta}I_2 \ .$$

– $$S_2(n,\lambda) \ge \int_1^{n+1}\frac{1}{(1+\lambda t^{2\beta})^2}dt \ .$$

Proof. To show the first point we just remark that

$$S_2(n,\lambda) = \sum_{i=1}^n\frac{1}{\left(1+\lambda i^{2\beta}\right)^2} \le \int_0^n\frac{1}{(1+\lambda t^{2\beta})^2}dt \le \int_0^{+\infty}\frac{1}{(1+\lambda t^{2\beta})^2}dt \ .$$

The second point is likewise straightforward. □

Lemma 4.6. *We have the following bounds on S_1 : for every $n \in \mathbb{N}^\star$, every $(\beta, \delta) \in \mathbb{R}_+^2$ such that $4\beta > 2\delta$ and every $\lambda \in \mathbb{R}_+^\star$,*

$$S_1(n,\lambda) \leq \frac{\lambda^{(2\delta-1)/2\beta}}{\beta\lambda^2} I_1 \ ,$$

Furthermore, let

$$t^\star = \left(\frac{4\beta - 2\delta}{2\delta\lambda}\right)^{1/2\beta}$$

and $n^\star = \lfloor t^\star \rfloor$.

– *If $n^\star < n-1$*

$$S_1(n,\lambda) \geq \int_0^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - \int_{n^\star}^{n^\star+2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ ;$$

– *while if $n^\star \geq n$*

$$S_1(n,\lambda) \geq \int_0^{n} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

Proof. Lemma 4.2 shows that $t \mapsto \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2}$ is increasing on $[0, t^\star]$ and decreasing on $[t^\star, +\infty($. Thus we have the following comparisons :

$$\int_0^{n^\star} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \leq \sum_{i=1}^{n^\star} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \leq \int_1^{n^\star+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt$$

and

$$\int_{n^\star+2}^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \leq \sum_{i=n^\star+1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \leq \int_{n^\star}^{n} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

By adding those two lines we get

$$\begin{aligned} S_1(n,\lambda) = \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} &\leq \int_1^{n^\star+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt + \int_{n^\star}^{n} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \\ &\leq 2\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ , \end{aligned}$$

which shows the first point. We also get, if $n^\star < n-1$

$$\begin{aligned} S_1(n,\lambda) &\geq \int_0^{n^\star} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt + \int_{n^\star+2}^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \\ &\geq \int_0^{n+1} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - \int_{n^\star}^{n^\star+2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ . \end{aligned}$$

The last point is evident, since if $n^\star \geq n$ the integrand is increasing on $[0, n]$. □

4.C Proof of Property 4.1

Let n and p be integers, σ, β and δ real numbers such that $(\mathbf{H_M}(\beta,\delta))$ hold. We want to study the value and the location of the infimum on $\mathbb{R}_+$ of

$$\lambda \longmapsto R(n,p,\sigma^2,\lambda,\beta,\delta,C) = C\lambda^2 \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} + \frac{\sigma^2}{np}\sum_{i=1}^{n}\frac{1}{(1+\lambda i^{2\beta})^2}$$

Property 4.7. *For every λ in $\mathbb{R}_+$, we have*

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \leq \frac{CI_1}{\beta}\lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np}\lambda^{-1/2\beta} \quad . \tag{4.22}$$

Proof. This is a straightforward application of the majorations of the finite sums by integrals given in Lemmas 4.5 and 4.6, together with the change of variables done in Lemmas 4.3 and 4.4. □

Lemma 4.7. *Let $A \in \mathbb{R}_+$, the minimum over $\mathbb{R}_+^\star$ of $\lambda \mapsto \lambda^{(2\delta-1)/2\beta} + A\lambda^{-1/2\beta}$ is attained for*

$$\lambda^\star = \left(\frac{A}{2\delta-1}\right)^{\beta/\delta}$$

and has for value

$$A^{1-(1/2\delta)}\frac{2\delta}{(2\delta-1)^{1-(1/2\delta)}} \quad .$$

Proof. This mapping is differentiable and has $+\infty$ for limit in 0 and in $+\infty$. Then

$$\frac{d}{d\lambda}\left(\lambda^{2\delta/(2\delta-1)} + A\lambda^{-1/2\beta}\right) = \frac{1}{\lambda}\left(\frac{2\delta-1}{2\beta}\lambda^{(2\delta-1)/2\beta} - \frac{A}{2\beta}\lambda^{-1/2\beta}\right) \quad .$$

We see there is only one minimizer $\lambda^\star$ verifying

$$\begin{array}{rrcl} & \frac{2\delta-1}{2\beta}(\lambda^\star)^{(2\delta-1)/2\beta} & = & \frac{A}{2\beta}(\lambda^\star)^{-1/2\beta} \\ \Leftrightarrow & (2\delta-1)^{2\beta}(\lambda^\star)^{2\delta-1} & = & A^{2\beta}(\lambda^\star)^{-1} \\ \Leftrightarrow & (\lambda^\star)^{2\delta} & = & \frac{A^{2\beta}}{(2\delta-1)^{2\beta}} \\ \Leftrightarrow & \lambda^\star & = & \left(\frac{A}{2\delta-1}\right)^{\beta/\delta} \quad . \end{array}$$

Pluging-in the value of $\lambda^\star$ leads to the optimal value

$$\begin{aligned} \left(\frac{A}{2\delta-1}\right)^{(2\delta-1)/2\delta} + A\left(\frac{A}{2\delta-1}\right)^{-1/2\delta} &= A^{(2\delta-1)/2\delta}\left((2\delta-1)^{(1/2\delta)-1} + (2\delta-1)^{1/2\delta}\right) \\ &= A^{(2\delta-1)/2\delta}(2\delta-1)^{1/2\delta}\left(\frac{1}{2\delta-1}+1\right) \\ &= A^{(2\delta-1)/2\delta}(2\delta-1)^{1/2\delta}\left(\frac{2\delta}{2\delta-1}\right) \\ &= A^{(2\delta-1)/2\delta}\frac{2\delta}{(2\delta-1)^{1-(1/2\delta)}} \quad . \end{aligned}$$

□

Definition 4.3. To simplify notations, since this quantity depends only on β and δ and appears throughout the paper, we will use the following notation :

$$\kappa(\beta,\delta) = I_1(\beta,\delta)^{1/2\delta} I_2(\beta)^{1-(1/2\delta)} (2\delta-1)^{1/2\delta} \frac{\delta}{\beta(2\delta-1)} \ . \tag{4.23}$$

We now prove Property 4.1

Proof. First $R(n,p,\sigma^2,0,\beta,\delta,C) = \frac{\sigma^2}{p}$, so that $R^\star(n,p,\sigma^2,\beta,\delta,C) \leq \frac{\sigma^2}{p}$. Then, the right-hand side of Equation (4.22) can be written as

$$\frac{CI_1}{\beta}\left[\lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2npCI_1}\lambda^{-1/2\beta}\right] \ .$$

Consequently, Lemma 4.7 implies that the optimal value of this upper bound with respect to λ is

$$\frac{CI_1}{\beta}\left(\frac{\sigma^2 I_2}{2npCI_1}\right)^{1-(1/2\delta)} \frac{2\delta}{(2\delta-1)^{1-(1/2\delta)}} \ ,$$

which is exactly the right-hand side of Equation (4.9). □

4.D Proof of Property 3.2

In order to perform this analysis we observe that R is composed of two factors :

– a bias factor $C\lambda^2 \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2}$, which is an increasing function of λ;

– a variance factor $\frac{\sigma^2}{np}\sum_{i=1}^{n}\frac{1}{(1+\lambda i^{2\beta})^2}$, which is a convex, decreasing function of λ.

We show that, if λ is too large, then the bias term exceeds the upper bound on $R^\star(n,p,\sigma^2,\beta,\delta,C)$ given in Equation (4.9).

Proof. We see that, using Equation (4.8), for every $\lambda \in \mathbb{R}_+$,

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \geq C\frac{\lambda^2}{(1+\lambda)^2} \ .$$

The right-hand side of this equation is increasing. Thus, if a real number ε matches this bound with the upper bound of $R^\star$, that is,

$$C\frac{\varepsilon^2}{(1+\varepsilon)^2} = \frac{1}{np} \times (np)^{1/2\delta} C^{(1/2\delta)} 2^{1/2\delta} \kappa(\beta,\delta) \ ,$$

we can state that the infimum of R is attained by a parameter $\lambda^\star \in [0,\varepsilon]$. The latter equation is equivalent to

$$\varepsilon^2 = A\left(\frac{np}{\sigma^2}\right)^{(1/2\delta)-1}(1+\varepsilon)^2 \ ,$$

with

$$A = C^{(1/2\delta)-1} 2^{1/2\delta} \kappa(\beta,\delta) \ .$$

This leads to

$$\varepsilon\left(1-\sqrt{A}\left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}\right)=\sqrt{A}\left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-2} \quad ,$$

so that if $\sqrt{A}\left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}<1$ that is, if

$$\frac{np}{\sigma^2}>\frac{1}{C}\times 2^{\frac{1}{2\delta-1}}\times\kappa(\beta,\delta)^{\frac{2\delta}{2\delta-1}} \quad ,$$

then

$$\varepsilon=\frac{\sqrt{A}\left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}}{1-\sqrt{A}\left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}}=\sqrt{A}\left(\frac{np}{\sigma^2}\right)^{(1/4\delta)-1/2}\left(1+\eta\left(\frac{np}{\sigma^2}\right)\right) \quad , \tag{4.24}$$

where $\eta(x)$ goes to 0 as x goes to $+\infty$. □

4.E On the way to showing Property 3.3

The proof of Property 4.3 uses two results that we give here.

4.E.1 Control of the risk on $\left[0,n^{-2\beta}\right]$

Property 4.8. *For every n, p, σ^2, C, δ and β, we have*

$$\inf_{\lambda\in\left[0,n^{-2\beta}\right]}\left\{R(n,p,\sigma^2,\lambda,\beta,\delta,C)\right\}\geq\frac{\sigma^2}{4p} \quad .$$

Proof. For every $\lambda\in\left[0,n^{-2\beta}\right]$ we have

$$\begin{aligned}R(n,p,\sigma^2,\lambda,\beta,\delta,C)&\geq\frac{\sigma^2}{np}\sum_{i=1}^{n}\frac{1}{\left(1+\lambda i^{2\beta}\right)^2}\\&\geq\frac{\sigma^2}{p}\times\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left(1+\left(\frac{i}{n}\right)^{2\beta}\right)^2}\\&\geq\frac{\sigma^2}{4p} \quad .\end{aligned}$$

□

4.E.2 Control of the risk on $\left[n^{-2\beta},\varepsilon\left(\frac{np}{\sigma^2}\right)\right]$

Property 4.9. *There exists an integer N and a constant $\alpha\in(0,1)$ such that for every (n,p,σ^2) such that $np/\sigma^2\geq N$, every $(\beta,\delta)\in\mathbb{R}_+^2$ such that $4\beta>2\delta>1$ and every $\lambda\in[n^{-2\beta},\varepsilon\left(\frac{np}{\sigma^2}\right)]$ we have*

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C)\geq\alpha\left(\frac{CI_1}{\beta}\lambda^{(2\delta-1)/2\beta}+\frac{\sigma^2I_2}{2\beta np}\lambda^{-1/2\beta}\right) \quad .$$

Proof. We seek to minor the two sums composing R, which was definded in Equation (4.8), by their integral counterparts, uniformly on $[n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$. The technical details are exposed in Lemmas 4.5 and 4.6.

For the first sum, using Lemma 4.5, we have that

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} &\geq \int_0^{n+1} \frac{1}{(1+\lambda t^{2\beta})^2} dt - \int_0^1 \frac{1}{(1+\lambda t^{2\beta})^2} dt \\ &\geq \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt - \int_{n+1}^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt - \int_0^1 \frac{1}{(1+\lambda t^{2\beta})^2} dt \ .\end{aligned}$$

First, with the change of variables $u = \lambda t^{2\beta}$ [Bac13],

$$\begin{aligned}\int_{n+1}^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{n+1}^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt}{\int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt} \\ &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{\lambda(n+1)^{2\beta}}^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \\ &\leq \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_1^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \ ,\end{aligned}$$

since $\lambda \geq n^{-2\beta}$.

We also have , with the change of variables $u = \lambda t^{2\beta}$ [Bac13],

$$\begin{aligned}\int_0^{1} \frac{1}{(1+\lambda t^{2\beta})^2} dt &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_0^{1} \frac{1}{(1+\lambda t^{2\beta})^2} dt}{\int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt} \\ &= \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_0^{\lambda} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \\ &\leq \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \frac{\int_0^{\varepsilon} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \ .\end{aligned}$$

Since ε, which was defined in Equation (4.24), verifies $\varepsilon(x) \longrightarrow 0$ as $x \longrightarrow +\infty$, we get

$$\frac{\int_0^{\varepsilon(x)} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1}{2\beta}-1}}{(1+u)^2} du} \underset{x\to+\infty}{\longrightarrow} 0 \ .$$

All those arguments imply that there exists an integer n_1 and real number $c_1 \in (0,1)$ such that, for every (n, p, σ^2) such that $np/\sigma^2 \geq n_3$ and for every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$,

$$\sum_{i=1}^{n} \frac{1}{(1+\lambda i^{2\beta})^2} \geq c_1 \int_0^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt \ .$$

For the second sum we carry a similar analysis, using Lemma 4.6 instead of Lemma 4.5. First, supposing that $4\beta > 2\delta$, we know that

$$\frac{\left\lfloor \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \right\rfloor}{\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}} \underset{\lambda\to 0}{\longrightarrow} 1 \ .$$

Since $\varepsilon(np/\sigma^2)$ goes to 0 as np/σ^2 goes to $+\infty$. Consequently, let $\zeta > 0$ and n_3 be an integer such that for every (n, p, σ^2) such that $np/\sigma^2 \geq n_3$, and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$, we have

$$\left| \frac{\left\lfloor \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \right\rfloor}{\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}} - 1 \right| < \zeta \quad \text{and} \quad \left| \frac{\left\lfloor \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} \right\rfloor + 2}{\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}} - 1 \right| < \zeta \ .$$

Consequently, for every (n, p, σ^2) such that $np/\sigma^2 \geq n_3$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$, we have (with $t^\star = \left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta}$ and $n^\star = \lfloor t^\star \rfloor$) :

$$n^\star \geq (1-\zeta)\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} = z_1 \quad \text{and} \quad n^\star + 2 \leq (1+\zeta)\left(\frac{4\beta-2\delta}{2\delta\lambda}\right)^{1/2\beta} = z_2 \ .$$

We can remark that $\lambda z_1^{2\beta}$ and $\lambda z_2^{2\beta}$ do not depend on λ. Consequently, for every (n, p, σ^2) such that $np/\sigma^2 \geq n_3$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$, we get

$$\int_{n^\star}^{n^\star+2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \leq \int_{z_1}^{z_2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

We finally see that

$$\begin{aligned}
\int_{z_1}^{z_2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{z_1}^{z_2} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt}{\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt} \\
&= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{\lambda z_1^{2\beta}}^{\lambda z_2^{2\beta}} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du} \\
&= c_3 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ ,
\end{aligned}$$

with

$$c_3 = \frac{\int_{\lambda z_1^{2\beta}}^{\lambda z_2^{2\beta}} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du} \in (0,1)$$

being independent of λ and arbitrarily close to 0. Thus, we have that, using Lemma 4.6,

– if $n^\star \geq n-1$:

$$\sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \geq \int_0^n \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \\ \geq \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - \int_n^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ ;$$

– if $n^\star < n-1$ and $np/\sigma^2 \geq n_3$:

$$\begin{aligned} & \sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \\ \geq & \int_0^n \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - c_3 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \\ \geq & \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - \int_n^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt - c_3 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ . \end{aligned}$$

With the change of variables $u = \lambda t^{2\beta}$ [Bac13],

$$\begin{aligned} \int_n^{+\infty} \frac{1}{(1+\lambda t^{2\beta})^2} dt &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \frac{\int_n^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt}{\int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt} \\ &= \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \frac{\int_{\lambda n^{2\beta}}^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du} \\ &\leq \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \frac{\int_1^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du}{\int_0^{+\infty} \frac{u^{\frac{1-2\delta}{2\beta}+1}}{(1+u)^2} du} \ , \end{aligned}$$

since $\lambda \geq n^{-2\beta}$. This implies that there exists an integer n_2 and real number $c_2 \in (0,1)$ such that, for every (n,p,σ^2) such that $np/\sigma^2 \geq n_2$ and for every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$,

$$\sum_{i=1}^{n} \frac{i^{4\beta-2\delta}}{(1+\lambda i^{2\beta})^2} \geq c_2 \int_0^{+\infty} \frac{t^{4\beta-2\delta}}{(1+\lambda t^{2\beta})^2} dt \ .$$

By taking $N = \max(n_1, n_2)$ and $\alpha = \min(c_1, c_2)$, we have that for every (n,p,σ^2) such that $np/\sigma^2 \geq N$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$

$$R(n,p,\sigma^2,\lambda,\beta,\delta,C) \geq \alpha \left(\frac{CI_1}{2\beta} \lambda^{(2\delta-1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np} \lambda^{-1/2\beta} \right) \ .$$

□

4.E.3 Proof of Property 4.3

Proof. This proof uses two results proved in Sections 4.E.1 and 4.E.2 of the appendix. Property 4.2 shows that R attains its minimum on $\left[0, \varepsilon\left(\frac{np}{\sigma^2}\right)\right]$, where $\varepsilon(x)$ goes to 0 as x goes to 0. First, Property 4.8 shows that

$$\inf_{\lambda \in \left[0, n^{-2\beta}\right]} \left\{ R(n, p, \sigma^2, \lambda, \beta, \delta, C) \right\} \geq \frac{\sigma^2}{4p} \ .$$

Then, using Property 4.9 shows that there exists an integer N and a constant $\alpha \in (0,1)$ such that for every (n, p, σ^2) such that $\frac{np}{\sigma^2} \geq N$, every $(\beta, \delta) \in \mathbb{R}_+^2$ such that $4\beta > 2\delta > 1$ and every $\lambda \in [n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)]$ we have

$$R(n, p, \sigma^2, \lambda, \beta, \delta, C) \geq \alpha \left(\frac{C I_1}{\beta} \lambda^{(2\delta - 1)/2\beta} + \frac{\sigma^2 I_2}{2\beta np} \lambda^{-1/2\beta} \right) \ .$$

Thus, using the same analysis than for Property 4.1, we get

$$\inf_{\lambda \in \left[n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)\right]} \left\{ R(n, p, \sigma^2, \lambda, \beta, \delta, C) \right\} \geq \alpha \left(\frac{np}{\sigma^2} \right)^{1/2\delta - 1} C^{1/2\delta} \kappa(\beta, \delta) \ .$$

□

4.E.4 Proof of Property 4.4

The proof of Property 4.1 clearly shows two regimes :
– when $\lambda_R^\star \leq n^{-2\beta}$, the multi-task risk is $\asymp \frac{\sigma^2}{p}$;
– when $\lambda_R^\star \geq n^{-2\beta}$, the multi-task risk is $\asymp \left(\frac{\sigma^2}{np}\right)^{1-1/2\delta}$.

We now show that if λ is too close to zero then the variance term exceeds the upper bound on $R^\star(n, p, \sigma^2, \beta, \delta, C)$ given in Equation (4.9).

Proof. Let us denote

$$m_1 = \inf_{\lambda \in \left[0, n^{-2\beta}\right]} \left\{ R(n, p, \sigma^2, \lambda, \beta, \delta, C) \right\}$$

and

$$m_2 = \inf_{\lambda \in \left[n^{-2\beta}, \varepsilon\left(\frac{np}{\sigma^2}\right)\right]} \left\{ R(n, p, \sigma^2, \lambda, \beta, \delta, C) \right\} \ .$$

If $m_1 < m_2$ then $\lambda_R^\star \leq \frac{1}{n^{2\beta}}$, else $\lambda_R^\star \geq \frac{1}{n^{2\beta}}$. Under the present assumptions, we can use the proof Property 4.3 and state that there exists an integer N_1 and a constant $\alpha \in (0,1)$ such that

$$\frac{\sigma^2}{p} \geq m_1 \geq \frac{\sigma^2}{4p} \ ,$$

and

$$2^{1/2\delta} \left(\frac{np}{\sigma^2} \right)^{1/2\delta - 1} C^{1/2\delta} \kappa(\beta, \delta) \geq m_2 \geq \alpha \left(\frac{np}{\sigma^2} \right)^{1/2\delta - 1} C^{1/2\delta} \kappa(\beta, \delta) \ .$$

Both assumptions $n^{2\delta - 1} \times \frac{\sigma^2}{p} \longrightarrow 0$ and $n^{2\delta - 1} \times \frac{\sigma^2}{p} \longrightarrow +\infty$ ensure that either $m_1 > m_2$ or $m_2 < m_1$ asymptotically hold. □

4.F Study of the different multi-task hypotheses

Lemma 4.8. *Under Assupmtion* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$)*, Assumption* (2Points) *is equivalent to*

$$\exists(\varepsilon_i)_{i\in\mathbb{N}} \in \{-1,1\}^{\mathbb{N}},\ \forall i \in \{1,\dots,n\},$$
$$\left\{ \begin{array}{rcl} \forall j \in \left\{1,\dots,\frac{p}{2}\right\},\ h_i^j &=& \sqrt{n} i^{-\delta}(\sqrt{C_1} + \varepsilon_i \sqrt{C_2}) \\ \forall j \in \left\{\frac{p}{2}+1,\dots,p\right\},\ h_i^j &=& \sqrt{n} i^{-\delta}(\sqrt{C_1} - \varepsilon_i \sqrt{C_2}) \end{array} \right. .$$

The risk of the estimator $\widehat{f}_\lambda^j = A_\lambda y^j$ for the jth task, which we denote by $R^j(\lambda)$, verifies

$$R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} - \sqrt{C_2}\right)^2) \leq R^j(\lambda)$$

and

$$R^j(\lambda) \leq R(n,1,\sigma^2,\lambda,\beta,\delta,\left(\sqrt{C_1} + \sqrt{C_2}\right)^2)\ .$$

Proof. We have that, for every $i \in \{1,\dots,n\}$

$$\left\{ \begin{array}{rcl} \frac{\mu_i}{\sqrt{p}} &=& \frac{1}{2}h_i^1 + \frac{1}{2}h_i^p \\ \varsigma_i^2 &=& \frac{1}{2}\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 + \frac{1}{2}\left(h_i^p - \frac{\mu_i}{\sqrt{p}}\right)^2 \end{array} \right.$$
$$\Leftrightarrow \left\{ \begin{array}{rcl} h_i^p &=& 2\frac{\mu_i}{\sqrt{p}} - h_i^1 \\ \varsigma_i^2 &=& \frac{1}{2}\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 + \frac{1}{2}\left(2\frac{\mu_i}{\sqrt{p}} - h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 \end{array} \right.$$
$$\Leftrightarrow \left\{ \begin{array}{rcl} h_i^p &=& 2\mu_i - h_i^1 \\ \varsigma_i^2 &=& \left(h_i^1 - \mu_i\right)^2 \end{array} \right.$$

This is equivalent to $h_i^1 = \frac{\mu_i}{\sqrt{p}} + \varsigma_i$ and $h_i^p = \frac{\mu_i}{\sqrt{p}} - \varsigma_i$. Thus, the first point is proved. For the second point, let $j \in \{1,\dots,p\}$. There exists $(\varepsilon_i)_{i\in\mathbb{N}} \in \{-1,1\}^{\mathbb{N}}$ such that $(h_i^j)^2 = ni^{-2\delta}\left(\sqrt{C_1} + \varepsilon_i\sqrt{C_2}\right)^2$. The risk of $\widehat{f}_\lambda^j$ then is

$$\lambda^2 \sum_{i=1}^n \frac{i^{4\beta-2\delta}\left(\sqrt{C_1} + \varepsilon_i\sqrt{C_2}\right)^2}{(1+\lambda i^{2\beta})^2} + \frac{\sigma^2}{n}\sum_{i=1}^n \frac{1}{\left(1+\lambda i^{2\beta}\right)^2}\ .$$

We conclude by seeing that, for every $\varepsilon \in \{-1,1\}$, we have $\left(\sqrt{C_1} - \sqrt{C_2}\right)^2 \leq \left(\sqrt{C_1} + \varepsilon\sqrt{C_2}\right)^2 \leq \left(\sqrt{C_1} + \sqrt{C_2}\right)^2$ □

Lemma 4.9. *Under Assupmtion* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$)*, Assumption* (1Out) *is equivalent to*

$$\exists(\varepsilon_i) i \in \mathbb{N} \in \{-1,1\}^{\mathbb{N}},\ \forall i \in \{1,\dots,n\},$$
$$\left\{ \begin{array}{rcl} \forall j \in \{1,\dots,p-1\},\ h_i^j &=& \sqrt{n} i^{-\delta}\left(\sqrt{C_1} + \varepsilon_i\sqrt{\frac{C_2}{p-1}}\right) \\ h_i^p &=& \sqrt{n} i^{-\delta}(\sqrt{C_1} - \varepsilon_i\sqrt{(p-1)C_2}) \end{array} \right. .$$

If $j \in \{1, \dots, p-1\}$, the risk of the estimator $\widehat{f}_\lambda^j = A_\lambda y^j$ for the jth task, which we denote by $R^j(\lambda)$, verifies

$$R(n, 1, \sigma^2, \lambda, \beta, \delta, \left(\sqrt{C_1} - \sqrt{\frac{C_2}{p-1}}\right)^2) \leq R^j(\lambda)$$

and

$$R^j(\lambda) \leq R(n, 1, \sigma^2, \lambda, \beta, \delta, \left(\sqrt{C_1} + \sqrt{\frac{C_2}{p-1}}\right)^2) \ ,$$

while the risk of the estimator $\widehat{f}_\lambda^p = A_\lambda y^p$ for the pth task, which is denoted by $R^p(\lambda)$, verifies

$$R(n, 1, \sigma^2, \lambda, \beta, \delta, \left(\sqrt{C_1} - \sqrt{(p-1)C_2}\right)^2) \leq R^p(\lambda)$$

and

$$R^p(\lambda) \leq R(n, 1, \sigma^2, \lambda, \beta, \delta, \left(\sqrt{C_1} + \sqrt{(p-1)C_2}\right)^2) \ .$$

Proof. For the first part, we have that, for every $i \in \{1, \dots, n\}$

$$\begin{aligned}
&\left\{ \begin{array}{rcl} \frac{\mu_i}{\sqrt{p}} & = & \frac{p-1}{p} h_i^1 + \frac{1}{p} h_i^p \\ \varsigma_i^2 & = & \frac{p-1}{p}\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 + \frac{1}{p}\left(h_i^p - \frac{\mu_i}{\sqrt{p}}\right)^2 \end{array} \right. \\
\Leftrightarrow &\left\{ \begin{array}{rcl} h_i^p & = & p\frac{\mu_i}{\sqrt{p}} - (p-1)h_i^1 \\ \varsigma_i^2 & = & \frac{p-1}{p}\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 + \frac{1}{p}\left(p\frac{\mu_i}{\sqrt{p}} - (p-1)h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 \end{array} \right. \\
\Leftrightarrow &\left\{ \begin{array}{rcl} h_i^p & = & p\frac{\mu_i}{\sqrt{p}} - (p-1)h_i^1 \\ \varsigma_i^2 & = & \frac{p-1}{p}\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 + \frac{(p-1)^2}{p}\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 \end{array} \right. \\
\Leftrightarrow &\left\{ \begin{array}{rcl} h_i^p & = & p\frac{\mu_i}{\sqrt{p}} - (p-1)h_i^1 \\ \varsigma_i^2 & = & (p-1)\left(h_i^1 - \frac{\mu_i}{\sqrt{p}}\right)^2 \end{array} \right.
\end{aligned}$$

This is equivalent to saying that there exists $(\varepsilon_i)_{i \in \mathbb{N}} \in \{-1, 1\}^{\mathbb{N}}$ such that

$$\begin{aligned}
&\left\{ \begin{array}{rcl} h_i^p & = & p\frac{\mu_i}{\sqrt{p}} - (p-1)h_i^1 \\ h_i^1 & = & \frac{\mu_i}{\sqrt{p}} + \frac{\varepsilon_i}{\sqrt{p-1}}\varsigma_i \end{array} \right. \\
\Leftrightarrow &\left\{ \begin{array}{rcl} h_i^1 & = & \frac{\mu_i}{\sqrt{p}} + \frac{\varepsilon_i}{\sqrt{p-1}}\varsigma_i \\ h_i^p & = & \frac{\mu_i}{\sqrt{p}} - \varepsilon_i\sqrt{p-1}\varsigma_i \end{array} \right.
\end{aligned}$$

□

Lemma 4.10. *Assumption* ($\mathbf{H_{2Points}}$) *implies Assumption* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$).

Proof. For every $i \in \{1, \dots, n\}$, we suppose we have

$$\left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta} (\sqrt{C_1} + \sqrt{C_2}) \\ h_i^p & = & \sqrt{n} i^{-\delta} (\sqrt{C_1} - \sqrt{C_2}) \end{array} \right. .$$

Thus,

$$\mu_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} h_i^j = \frac{\sqrt{p}}{2} \left(h_i^1 + h_i^p \right) = \sqrt{p} \times \sqrt{n} i^{-\delta} \sqrt{C_1} \ ,$$

so that $\mu_i^2 = p C_1 n i^{-2\delta}$. Furthermore,

$$\varsigma_i^2 = \frac{1}{p} \sum_{j=1}^{p} \left(h_i^j - \frac{\mu_i}{\sqrt{p}} \right)^2 = \frac{1}{p} \sum_{j=1}^{p} \left(\sqrt{n} i^{-\delta} \sqrt{C_2} \right)^2 = C_2 n i^{-2\delta} \ .$$

□

Lemma 4.11. *Assumption* ($\mathbf{H_{1Out}}$) *implies Assumption* ($\mathbf{H_{AV}}(\delta, C_1, C_2)$).

Proof. For every $i \in \{1, \dots, n\}$, we suppose we have

$$\left\{ \begin{array}{rcl} h_i^1 & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1} + \frac{1}{\sqrt{p-1}} \sqrt{C_2} \right) \\ h_i^p & = & \sqrt{n} i^{-\delta} \left(\sqrt{C_1} - \sqrt{p-1} \sqrt{C_2} \right) \end{array} \right. .$$

Thus,

$$\mu_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} h_i^j = \frac{1}{\sqrt{p}} \left((p-1) h_i^1 + h_i^p \right) = \sqrt{p} \times \sqrt{n} i^{-\delta} \sqrt{C_1} \ ,$$

so that $\mu_i^2 = p C_1 n i^{-2\delta}$. Furthermore,

$$\begin{aligned} \varsigma_i^2 &= \frac{1}{p} \sum_{j=1}^{p} \left(h_i^j - \frac{\mu_i}{\sqrt{p}} \right)^2 \\ &= \frac{1}{p} \left[(p-1) \left(\sqrt{n} i^{-\delta} \frac{\sqrt{C_2}}{\sqrt{p-1}} \right)^2 + \left(\sqrt{p-1} \sqrt{n} i^{-\delta} \sqrt{C_2} \right)^2 \right] = C_2 n i^{-2\delta} \ . \end{aligned}$$

□

Chapitre 5

Conclusion and open questions

Résumé. Nous résumons dans cette partie les principales avancées apportées par cette thèse et envisageons quelques pistes pouvant étendre ce travail.

In this section, we recapitulate the main results brought by this thesis and their main consequences. We also formulate a few questions brought up by this work which, if solved, could extend it.

While constructing our multi-task procedure, we introduced a matricial regularization parameter M, which is meant to encode the similarity between the tasks. The main question of Chapter 3 was to find how to correctly calibrate this parameter and to investigate the properties of the resulting estimator.

Estimation of Σ

Our procedure is mostly based on the estimation of the covariance matrix Σ of the noise between the tasks. As for now, this estimation is naïve but gives acceptable estimation rates. It relies on one-dimensionnal projections of the p tasks, performed on $p(p+1)/2$ different directions, which are then used to build the estimator $\widehat{\Sigma}$. The estimator is then shown to approximate Σ well, with a multiplicative error term of the form

$$(1-\eta_{n,p})\Sigma \preceq \widehat{\Sigma} \preceq (1+\eta_{n,p})\Sigma$$

(Theorem 3.2, page 56), where $A \preceq B$ if $B - A$ is symetric positive semi-definite.

Open Question 1. *Can we obtain minimax rates on $\eta_{n,p}$? The results found in the litterature consider $f = 0$ and use heavy assumptions on Σ, while giving rates that are hardly comparable to ours, see Cai et al. [CZZ10] for instance.*

Open Question 2. *Can we build $\widehat{\Sigma}$ with less projections, while keeping a reasonable rate $\eta_{n,p}$?*

A more general multi-task hypothesis

We showed oracle inequalities that control the risk of the estimator (Theorems 3.3 and 3.4, page 57), thus ensuring that the selected parameter is optimal in a given collection of parameters $\mathcal{M}$. One of the key aspects is the distinction between two settings.

1. A setting where we have a strong assumption on the similarity between the tasks, which is reflected on the shape of $\mathcal{M}$ (that is, matrices that are jointly diagonalizable on an orthonormal basis). It mostly covers the cases where the different regression functions are split into several clusters, the allocation of each function to its cluster being known.
2. A setting where such a strong assumption is not available. If a weaker one gives a larger set $\mathcal{M}$, or a more difficult one to treat, one way to deal with it is to discretize it.

The first setting allows for a simpler estimation of Σ and a more precise oracle inequality compared to the discrete case. Practically speaking, it is also easy to optimize the criterion used here to select the regularization parameter M, since this can be done separately on one-dimensional variables.

Open Question 3. *Which optimization tools can be used to solve the minimization problems given in Definition 3.5 (page 57) over larger or more complex sets $\mathcal{M}$?*

Theoretically speaking, the main difficulty is to concentrate the quantities $\delta_1(M)$, $\delta_2(M)$, $\delta_3(M)$, $\delta_4(M)$ (Definition 3.10, page 77) around their means uniformly over $\mathcal{M}$. With the mentioned hypotheses, this can easily be done by doing uniform controls over either real parameters or discrete sets (Lemma 3.4, page 77).

Open Question 4. *Can similar concentration inequalities be obtained uniformly on larger or more complex sets $\mathcal{M}$?*

In Chapter 4, we studied the oracle risks in the first setting, by using a bias-variance decomposition and when using a set of matrices $\mathcal{M}$ tailored for this assumption. By doing so, we were able to compare single-task and multi-task risks, both theoretically and on simulated examples. We noticed two main facts. First, if the tasks are extremely similar, the multi-task procedure outperforms the single-task one by a large amount. However, if the similarity between the tasks was wrongly modeled and if the tasks are not very similar, then the multi-task estimator can perform awfully compared to the single-task one.

Open Question 5. *Given a repartition of the task-wise regression functions, what would be the best possible matrix $M^\star$ or, at least, can we obtain a small set $\mathcal{M}^\star$ which contains $M^\star$?*

Open Question 6. *Is it possible to design a more robust procedure for which the estimators can still be computed and analysed ? For instance, given sets $\mathcal{M}_1 \subset \mathcal{M}_2$ of matricial parameters, where $\mathcal{M}_1$ is much smaller than $\mathcal{M}_2$, is it possible to select one of those sets before calibrating the estimator over them ?*

One of the positive effects of solving those last questions would be to loosen the assumptions on the model, one of which is that the tasks are observed on the same input points. This assumption is mostly needed to preserve the structure of $\mathcal{M}$ that we impose (the matrices are jointly diagonalizable in an orthonormal basis), and could thus disappear if this structure is no longer needed.

Bibliographie

[AB11] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, 2011. arXiv :0909.1884v2.

[AC10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4 :40–79, 2010.

[AEP08] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3) :243–272, 2008.

[Aka70] Hirotogu Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22 :203–217, 1970.

[AM09] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10 :245–279 (electronic), 2009.

[Arl07] Sylvain Arlot. *Rééchantillonnage et sélection de modèles*. PhD thesis, Université Paris 11, 2007.

[Arl09] Sylvain Arlot. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3 :557–624, 06 2009. extended version of http ://hal.archives-ouvertes.fr/hal-00125455, with a technical appendix AMS 62G09 ; 62G08 ; 62M20.

[Aro50] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–404, May 1950.

[AZ05] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6 :1817–1853, December 2005.

[Bac13] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. *International Conference on Learning Theory*, 26, 2013.

[Bax00] Jonathan Baxter. A model of inductive bias learning. *Journal Of Artificial Intelligence Research*, 12 :149–198, 2000.

[BDS03] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

[BH03] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4 :83–99, December 2003.

BIBLIOGRAPHIE

[BL08] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6) :2577–2604, 2008.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138 :33–73, 2007.

[BV11] Kevin Bleakley and Jean-Philippe Vert. The group fused Lasso for multiple change-point detection. 2011.

[BZ80] Philip J. Brown and James V. Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1) :pp. 64–74, 1980.

[Car97] Rich Caruana. Multitask learning. *Machine Learning*, 28 :41–75, July 1997.

[CDV07] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3) :331–368, 2007.

[CZZ10] T Tony Cai, Cun-Hui Zhang, and Harrison H Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4) :2118–2144, 2010.

[DVN79] Norman R. Draper and Craig R. Van Norstrand. Ridge regression and jamesstein estimation : Review and comments. *Technometric*, 21, 1979.

[EM73] Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68, 1973.

[EM77] Bradley Efron and Carl N. Morris. Stein's paradox in statistics. *Scientific American*, 236 :119–127, 1977.

[EMP05] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6 :615–637, 2005.

[FGF12] Sergey Feldman, Maya R. Gupta, and Bela A. Frigyik. Multi-task averaging. *Advances in Neural Information Processing Systems 25*, pages 1178–1186, 2012.

[Gau09] Karl Friedrich Gauss. *Theoria motus corporum cœlestium in sectionibus conicis solem ambientium.* 1809.

[Gir11] Christophe Giraud. Low rank multivariate regression. *Electronic Journal of Statistics*, 5 :775–799, 2011.

[GRC09] Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with non-convex penalties and dc programming. *IEEE Trans. Signal Processing*, 57(12) :4686–4698, 2009.

[Gu02] Chong Gu. *Smoothing spline ANOVA models.* Springer, 2002.

[HJ91] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 1970.

[HKZ11] Daniel Hsu, Sham M. Kakade, and Tong Zhang. An analysis of random design linear regression. *arXiv preprint arXiv :1106.2363*, 2011.

[Hoe59] Arthur E. Hoerl. Optimum solution of many variables equation. *Chemical Engineering Progress*, 55, 1959.

[Hoe62] Arthur E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58, 1962.

[Hoe85] Roger W. Hoerl. Ridge analysis 25 years later. *The American Statistician*, 39(3), 1985.

[JBV08] Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning : A convex formulation. *Computing Research Repository*, pages –1–1, 2008.

[Joh94] Iain M. Johnstone. Minimax bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics V*, pages 303–326. Springer New York, 1994.

[JS61] William James and Charles Stein. Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, 1(1961) :361–379, 1961.

[Kar08] Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008.

[KY08] Vladimir Koltchinskii and Ming Yuan. Sparse recovery in large ensembles of kernel machines. *Conference of Learning Theory*, (21), 2008.

[KY10] Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6) :3660–3695, 2010.

[LBBJ10] Percy Liang, Francis Bach, Guillaume Bouchard, and Michael I. Jordan. Asymptotically optimal regularization in smooth parametric models. In *Advances in Neural Information Processing Systems*, 2010.

[Leg05] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. 1805.

[Ler11] Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 2011.

[LF09] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B) :4254, 2009.

[LPTvdG09] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sarah A. van de Geer. Taking advantage of sparsity in multi-task learning. *Conference On Learning Theory*, 2009.

[LPTvdG11] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4) :2164–2204, 2011.

[LST11] Joseph J. Lim, Ruslan Salakhutdinov, and Antonio Torralba. Transfer learning by borrowing examples for multiclass object detection. *Advances in Neural Information Processing Systems 24*, 2011.

[Mal73] Colin L. Mallows. Some comments on C_P. *Technometrics*, pages 661–675, 1973.

BIBLIOGRAPHIE

[Mas07] Pascal Massart. *Concentration Inequalities and Model Selection.* École d'Été de Probabilités de Saint Flour XXXIII - 2003. Springer, 2007.

[OWJ11] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1) :1–17, 2011.

[RT11] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 2011.

[RW06] Carl E. Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[SAB12] Matthieu Solnon, Sylvain Arlot, and Francis Bach. Multi-task Regression using Minimal Penalties. *Journal of Machine Learning Research*, 13 :2773–2812, September 2012.

[SS02] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 12 2002.

[Ste56] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, 1(399) :197–206, 1956.

[Ste81] Charles Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1981.

[TO96] Sebastian Thrun and Joseph O'Sullivan. Discovering structure in multiple learning tasks : The TC algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 1996.

[Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2008.

[Wah90] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Was06] Larry Wasserman. *All of nonparametric statistics*, volume 4. Springer, 2006.

[Zha05] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9) :2077–2098, 2005.

Matthieu SOLNON
Apprentissage statistique multi-tâches

Résumé : Cette thèse a pour objet la construction, la calibration et l'étude d'estimateurs multi-tâches, dans un cadre fréquentiste non paramétrique et non asymptotique.

Nous nous plaçons dans le cadre de la régression *ridge* à noyau et y étendons les méthodes existantes de régression multi-tâches. La question clef est la calibration d'un paramètre de régularisation matriciel, qui encode la similarité entre les tâches. Nous proposons une méthode de calibration de ce paramètre, fondée sur l'estimation de la matrice de covariance du bruit entre les tâches. Nous donnons ensuite pour l'estimateur obtenu des garanties d'optimalité, via une inégalité oracle, puis vérifions son comportement sur des exemples simulés.

Nous obtenons par ailleurs un encadrement précis des risques des estimateurs oracles multi-tâches et mono-tâche dans certains cas. Cela nous permet de dégager plusieurs situations intéressantes, où l'oracle multi-tâches est plus efficace que l'oracle mono-tâche, ou *vice versa*. Cela nous permet aussi de nous assurer que l'inégalité oracle force l'estimateur multi-tâches à avoir un risque inférieur à l'estimateur mono-tâche dans les cas étudiés. Le comportement des oracles multi-tâches et mono-tâche est vérifié sur des exemples simulés.
Mots-clefs : Calibration de paramètres ; Inégalité oracle ; Méthodes à noyau ; Multi-tâches ; Régression *ridge* ; Statistique fréquentiste ; Statistique non asymptotique ; Statistique non paramétrique

Abstract : This thesis aims at constructing, calibrating and studying multi-task estimators, in a frequentist non-parametric and non-asymptotic framework.

We consider here kernel ridge regression and extend the existing multi-task regression methods in this setting. The main question is the calibration of a matricial regularization parameter, which encodes the similarity between the tasks. We propose a method to calibrate this parameter, based on the estimation of the covariance matrix of the noise between tasks. We then show optimality guarantees for the estimator thus obtained, via an oracle inequality. We also check its behaviour on simulated examples.

We carefully bound the risks of both multi-task and single-task oracle estimators in some specific settings. This allows us to discern several interesting situations, whether the multi-task oracle outperforms the single-task one or not. This ensure the oracle inequality enforces the multi-task oracle to have a lower risk than the single-task one in the studied settings. Finally, we check the behaviour of the oracles on simulated examples.
Keywords : Frequentist statistics ; Kernel methods ; Multi-task ; Non-asymptotic statistics ; Non-parametric statistics ; Oracle inequality ; Parameter calibration ; Ridge regression