Model Averaging in Large Scale Learning

Edwin Grappin

NNT : 2018SACLG001

Thèse de doctorat

de l'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Etablissement d'inscription : ENSAE ParisTech

Etablissement d'accueil : CREST (UMR CNRS 9194) - Laboratoire de Statistiques

Spécialité de doctorat : Mathématiques fondamentales

Edwin Grappin

Estimateur par agrégat en apprentissage statistique en grande dimension

Soutenue le 6 mars 2018 à l'ENSAE ParisTech, Palaiseau.

Après avis des rapporteurs :	Jalal Fadili	(GREYC CNRS, ENSICAEN)
	Karim Lounici	(Georgia Institute of Technology)

Jury de soutenance :	Cristina Butucea	(Université Paris-Est)	Examinateur
	Ismaël Castillo	(Université Sorbonne)	Examinateur
	Arnak Dalalyan	(CREST - ENSAE - GENES)	Directeur de thèse
	Jalal Fadili	(GREYC CNRS, ENSICAEN)	Rapporteur
	Mohamed Hebiri	(Université Paris-Est)	Examinateur
	Alexandre Tsybakov	(CREST - ENSAE - GENES)	Président de jury

NNT : 2018SACLG001

Thesis presented for the title of Doctor of Philosophy

at Université Paris-Saclay.

Doctoral School of Mathematics Hadamard (EDMH, ED 574)

University : ENSAE ParisTech

Hosting research center: CREST (UMR CNRS 9194) - Laboratoire de Statistiques

Doctoral specialty: Fundamental mathematics

Edwin Grappin

Model Averaging in Large Scale Learning

6th March 2018 at ENSAE ParisTech, Palaiseau.

Reviewing committee :	Jalal Fadili	(GREYC CNRS, ENSICAEN)
	Karim Lounici	(Georgia Institute of Technology)

Ph.D. committee :	Cristina Butucea	(Université Paris-Est)
	Ismaël Castillo	(Université Sorbonne)
	Arnak Dalalyan	(CREST - ENSAE - GENES) - Ph.D. Supervisor
	Jalal Fadili	(GREYC CNRS, ENSICAEN)
	Mohamed Hebiri	(Université Paris-Est)
	Alexandre Tsybakov	(CREST - ENSAE - GENES)

Model Averaging in Large Scale Learning

Edwin Grappin

Submitted for the degree of Doctor of Philosophy at Université Paris-Saclay
February 2018

Abstract

This thesis explores both statistical and computational properties of estimations procedures closely related to aggregation in the problem of high-dimensional regression in a sparse setting. The exponentially weighted aggregate is well studied in the machine learning and statistical literature. It benefits from strong results in fixed and random design with a PAC-Bayesian approach. However, little is known about the properties of the exponentially weighted aggregate with Laplace prior. In Chapter 2 we study the statistical behaviour of the prediction loss of the exponentially weighted aggregate with Laplace prior in the fixed design setting. We establish sharp oracle inequalities which generalize the properties of the Lasso to a larger family of estimators. These results also bridge the gap from the Lasso to the Bayesian Lasso as these estimators belong to the class of estimators we consider. Moreover, the method of the proof can be easily applied to other estimators. Oracle inequalities are proven for the matrix regression setting with the nuclear norm prior. In Chapter 3 we introduce an adjusted Langevin Monte Carlo sampling method that approximates the exponentially weighted aggregate with Laplace prior in an explicit finite number of iterations for any targeted accuracy. These works generalize the results proved in Dalalyan (2017) in order to apply theoretical guarantees for non-smooth priors such as the Laplace prior. In Chapter 4, we study a complementary subject, namely the statisctical behaviour of adjusted versions of the Lasso for the transductive and semi-supervised learning task in the random design setting. Upperbound on the prediction risk, in both deviation and expectation, are proved and we point out that unlabeled features can substantially improve bounds on the prediction loss.

Estimateur par agrégat en apprentissage statistique en grande dimension

Edwin Grappin

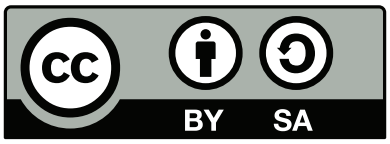
Thèse présentée dans le cadre d'un doctorat à l'Université Paris-Saclay
Février 2018

Abstract

Les travaux de cette thèse explorent les propriétés statistiques et computationnelles de procédures d'estimation par agrégation appliquées aux problèmes de régression en grande dimension dans un context parcimonieux (ou *sparse*). Les estimateurs par agrégation à poids exponentiels font l'objet d'une abondante littérature dans les communautés de la statistique et de l'apprentissage automatisé. Ces méthodes bénéficient de résultats théoriques optimaux sous une approche PAC-Bayésienne dans le cadre de données aléatoires ou fixes. Cependant, le comportement théorique de l'agrégat avec *prior* de Laplace n'est guère connu. Ce dernier représente pourtant un intérêt important puisqu'il est l'analogue du Lasso dans le cadre pseudo-bayésien. Le Chapitre 2 explicite une borne du risque de prédiction de cet estimateur, généralisant ainsi les résultats du Lasso. De ce fait, nous montrons aussi que pour certains niveaux faibles de la température, l'estimateur bénéficie de bornes optimales. Le Chapitre 3 prouve qu'une méthode de simulation s'appuyant sur un processus de Langevin Monte Carlo permet de choisir explicitement le nombre d'itérations nécessaire pour garantir une qualité d'approximation souhaitée. Le Chapitre 4 introduit des variantes du Lasso pour améliorer les performances de prédiction dans des contextes partiellement labélisés.

Declaration

The work in this thesis is based on research carried out at ENSAE with the Statistics department of CREST, France.



Dedicated to

Bérengère, Pierre and my parents without whom this thesis would have never been possible. Had they not pretended to find interesting what I was doing for such a long time, this journey would have seemed very arid to me.

Acknowledgements

First and foremost I want to thank Arnak, my Ph.D. advisor. Not only Arnak is one of the most passionate and pedagogic professors I have ever met, he also proves himself to be an understanding, supportive and benevolent advisor. Arnak introduced me to incredibly stimulating statistical questions and I appreciate all his contributions of support, time and ideas to make my Ph.D. experience the best it could have been. His guidance was invaluable and I could not imagine having a better advisor and mentor for my Ph.D journey.

I am also very grateful to the examiners of this thesis, namely Karim Lounici and Jalal Fadili for their time and their kind feedbacks on this work. And I want to thank the members of my dissertation committee: Cristina Butucea, Alexandre Tsybakov, Jalal Fadili, Mohamed Hebiri, Ismaël Castillo, and Arnak Dalalyan.

The members of the Statistical laboratory have substantially contributed to my personal and professional time at CREST. They have all been very supportive and have provided me with numerous advice and some very thoughtful insights. This laboratory is a great place to both explore Statistics and enjoy a tremendous environment. Obviously, I am grateful to Alexandre Tsybakov, who manages and has gathered so many talented and enthusiastic professors in the Statistics laboratory. I am also grateful for the excellent example he has provided as a great Statistician and professor.

Among all the members of the laboratory, I have a special thought for Pierre Alquier who brings life to the coffee room (and the working groups) and who always has had a spare dose of coffee when the most needed. I think Pierre has helped every Ph.D. student I know from the Statistics laboratory to make our thesis as enjoyable as possible. On a personal note, I consider Pierre as a friend. He has always been there during both the difficult times as well as the most exciting periods of this three-year journey. Looking backward, I doubt my Ph.D. would have been possible without Pierre. I am also very grateful to Nicolas Chopin, Judith Rousseau, Cristina Butucea, Marco Cuturi, Nicolas Chopin, Massimiliano Pontil, Anna Simoni, Olivier Catoni and Guillaume Lecué who have offered their points of view and their knowledge with a

tremendous enthusiasm.

It goes without saying that I am thankful to all the Ph.D. students of the laboratory. I could only start with James, with whom I shared the same office for two years and who taught me the do's and dont's of the Ph.D. student life. I am grateful to Alexander, the Ph.D. successor of James (they were both under the supervision of Nicolas Chopin) for his enthusiasm and all the great time we spent together. I am grateful to Mehdi for the incredible human being he is and all the support he brings me at some important times. All the PhD students I had the chance to meet during my Ph.D. study brought this very contagious enthusiasm that I dearly appreciate: Pierre Bellec, Mohamed, The Tien, Léna, Gautier, Lionel, Philippe and Vincent.

I gratefully acknowledge the financial fundings that made my Ph.D. possible. I have been funded by the Labex Ecodec and the CREST. I am glad I had the opportunity to study within the Ph.D. program in Mathematics "Ecole Doctorale de Mathématiques Hadamard" (EDMH) of the Université Paris-Saclay.

Lastly, I would like to thank my family and my friends for all their love and encouragement. My Ph.D. time was made enjoyable thanks to my dearest friends Rudy, Arthur, Alexis, Guillaume *Moby* and Nicolas who have all been very supportive and have shared some great and unforgetable moments with me. I want to thank Pierre for his incredible support and all the things we have lived together the last three years. Pierre is a very unique person that has proven his loyalty and his determination by helping me through my Ph.D. journey. The last 10 years, I had the chance to be part of the Ultimate Frisbee community and I am grateful to my teammates and all the players, coaches, mentors who I crossed onto fields around the world and are inspiring examples of the values I cherish the most.

Obviously, I am thankful to my parents who raised me with a love for science and supported me in all my pursuits, whatever they are. Last but not the least, I would like to thank Bérengère for her loving support and her incredible patience she showed during this long journey. Thank you.

Edwin Grappin

March 2018

Estimateur par agrégat en apprentissage statistique en grande dimension

Résumé Substantiel

Soient n et p des entiers strictement positifs. Considérons le couple $(\boldsymbol{X}, \boldsymbol{y}) \in (\mathbb{R}^{n\times p} \times \mathbb{R}^n)$ tiré d'une distribution P sur l'espace $\mathcal{X} \times \mathcal{Y}$. Dans le cadre d'un problème de régression, l'objectif est de prédire le vecteur $\boldsymbol{y}$ à partir d'un jeu de données $\boldsymbol{X}$. Une approche possible consiste à estimer une fonction $f : \mathcal{X} \to \mathcal{Y}$ qui minimise le risque

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} l(y, f(\boldsymbol{x}))\, P(\mathrm{d}\boldsymbol{x}, \mathrm{dy}), \tag{0.0.1}$$

où l est une fonction de perte arbitrairement choisie. Nous appelons $f^\star$ la fonction qui minimise Equation 0.0.1. Dans ce cas, le problème de régression peut s'écrire sous la forme

$$\boldsymbol{y} = f^\star(\boldsymbol{X}) + \boldsymbol{\xi},$$

où $\boldsymbol{\xi} \in \mathbb{R}^p$ est un vecteur de variables aléatoires.

Etude théorique de l'EWA avec *prior* de Laplace

L'apport principal de ce manuscrit est l'étude du comportement théorique de l'estimateur par agrégation avec prior de Laplace lorsqu'il existe une représentation quasi-parcimonieuse[1] de la relation fonctionnelle qui lie $\boldsymbol{y}$ au jeu de données $\mathbf{X}$.

L'estimation par agrégation à poids exponentiels est une méthode efficace pour inférer un signal dans un cadre quasi-parcimonieux (Dalalyan and Tsybakov, 2012a,b). Différents priors ont été étudiés dans la litérature de l'apprentissage statistique. Il est intéressant de remarquer que le *prior* de Laplace n'a jamais été utilisé efficacement.

L'estimateur par agrégat avec *prior* de Laplace représente un intérêt théorique puisqu'il est l'analogue pseudo-bayésien de l'estimateur Lasso. L'estimateur Lasso est certainement l'estimateur le plus largement étudié (et utilisé) parmi les méthodes de régression pénalisée dans un contexte quasi-parcimonieux en grande dimension. En dépit de ses avantages computationnels et

[1] Le lecteur est invité à lire la suite de ce manuscrit pour obtenir une compréhension plus complète de la notion de quasi-parcimonie abordée sous le terme de *nearly-sparse*.

théoriques, les garanties d'inégalités oracle optimales pour le Lasso nécessitent des hypothèses restrictives sur le jeu de données $\mathbf{X}$.

A contrario, les estimateurs par aggrégation à poids exponentiels (EWA) bénéficient de résultats théoriques optimaux sous une approche PAC-Bayésienne dans le cadre de données aléatoires ou fixes avec des hypothèses moins contraignantes. A ce jour, l'EWA avec *prior* de Laplace est peu étudié. Le Chapitre 2 explicite une borne du risque de prédiction de cet estimateur, généralisant ainsi les résultats du Lasso. De ce fait, nous montrons aussi que pour de faibles niveaux du paramètre de température, l'estimateur bénéficie de bornes optimales.

Le principal objet d'étude de cette thèse est la régression linéaire où l'on cherche à prédire $\boldsymbol{y}$ par une relation linéaire entre le jeu de données $\mathbf{X}$ et un vecteur $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{0.0.2}$$

où l'on cherche à estimer $\boldsymbol{\beta}$ de sorte à minimiser une fonction de perte.

Dans ce contexte, le Théorème 2.3.1 du Chapitre2 est une version simplifiée des résultats. Il permet de mettre en évidence l'impact du choix du paramètre de température τ sur la borne du risque de prédiction.

Le Théorème 2.4.1 explicite des résultats de type concentration du *pseudo-posterior*. Ces résultats sont généralisable à d'autres *prior* et à d'autres contextes tels que la régression matricielle. Ainsi les Théorèmes 2.5.1 et 2.5.2 du Chapitre 2 étendent ces résultats au cas matriciel.

Etude théorique d'une méthode de simulation de l'EWA avec *prior* de Laplace

Si le Chapitre 2 permet de regrouper les résultats théoriques du Lasso et de son analogue pseudo-bayésien, le Chapitre 3 en étudie l'aspect computationnel. Garantir l'existence d'une méthode qui approche efficacement cet estimateur s'avère être un défi plus difficile. Cela reste cependant essentiel pour que l'EWA avec *prior* de Laplace soit utilisable en pratique. En nous appuyant sur les travaux de Dalalyan (2016), Durmus and Moulines (2016) et Dalalyan (2017) nous étudions le comportement d'une méthode de simulation par Langevin Monte Carlo pour approcher cet estimateur.

Une application directe d'un processus de Langevin Monte Carlo comme présenté dans (Dalalyan, 2016, 2017; Durmus and Moulines, 2016) ne garantirait pas nécessairement l'obtention d'une précision souhaitée après un nombre fini d'itérations. En effet, ces résultats nécessitent que le

log-posterior soit fortement convexe et lisse alors que dans le cas du *prior* de Laplace le *log-posterior* n'est pas différentiable. De même, la forte convexité n'est pas respectée pour tout jeu de données. Dans le Chapitre 3, nous résolvons partiellement cette question. Nous étudions le comportement d'une simulation de la discrétisation d'Euler d'un processus de Langevin Monte Carlo. Plus particulièrement, nous étudions la qualité de la simulation au sens de la distance de Wasserstein par rapport à l'agrégat à poids exponentiels avec *prior* de Laplace ciblé. L'approche consiste à adapter le travail de Dalalyan (2016) afin de contourner la non différentiabilité du *pseudo posterior*. Nous explicitons un nombre d'itérations K du même ordre de grandeur que le nombre d'itérations nécessaires dans Dalalyan (2016) en vue d'une tolérance à l'erreur ϵ et de la dimension p. Il s'agit de noter que cette ébauche de résultat ne résout pas entièrement la question computationnelle. En effet, ces résultats sont garantis sous l'hypothèse de forte convexité. Cela requiert notamment des hypothèses trop restrictives sur la matrice de Gram. En particulier, ces hypothèses ne sont pas réalistes dans un problème en grande dimension. En effet, dans le Chapitre 3, nous supposons que la plus petite valeur propre de la matrice de Gram est strictement positive. Malgrés ces limites cette étude définie une méthode computationnelle qui garantie une approximation précise d'une densité ciblée dans une situation légérement plus généralisée que la littérature existante.

Apprentissage transductif et semi-supervisé

Le Chapitre 4 est une étude de l'estimateur Lasso dans des contextes semi-supervisés ou transductifs. Il peut être lu indépendemment du reste de ce manuscrit bien qu'il complète et peut se voir complété par les résultats des autres chapitres de cette thèse. Nous montrons que des données non labélisées devraient être utilisées dans le calcul de l'estimateur afin d'inférer la matrice de variance-covariance. Ainsi, nous présentons deux adaptations de l'estimateur Lasso afin d'améliorer les performances de prédiction dans un cadre d'apprentissage transductif ou partiellement labélisé. Sous certaines hypothèses, nous démontrons des inégalités oracle optimales dans le cadre de *designs* aléatoires.

Contents

List of Figures

Chapter 1

Introduction

Contents

1.1 Context

1.1.1 The rise of Statistics

Over the last decades, Statistics has been at the center of attention, in a wide variety of ways. Hardly a day goes by without one hearing about Statistics, Artificial Intelligence, Machine Learning or Big Data. Large companies such as Google, Apple, Facebook and Amazon (also known as GAFA) or Baidu, Alibaba, Tencent and Xiaomi (sometimes called BATX) play an important role in the mainstream status of all these technical terms. What can be done with data and the computer resources that are recently accessible causes a lot of ink to flow and is subject to a great deal of thoughts and speculations. While some consider artificial intelligence as a threat for the future of humanity[1], others see their applications for the greater good, and are very optimistic about the impact of artificial intelligence applications, such as automated or assisted medical diagnoses for early detection, or robots that substitute for humans in laborious chores. If the impact of artificial intelligence on the future is not well understood, it seems to be a great consensus that its applications are going to be a key changer of the day-to-day life. Recent improvements, such as the first weak artificial intelligence algorithm, AlphaGo, that can outperform the best Go masters in the world, have reinforced the common belief that artificial intelligence is intended to a bright future. As discussed in McCarthy and Hayes (1969), with such unclear questions at stakes, it is clear that philosophical and ethical guidelines are to be questioned and that national and international regulations will be necessary to ensure a positive impact of artificial intelligence applications.

The rise of challenges and breakthroughs put under the name of artificial intelligence has been made possible with technological improvements. The most important being the increase of computer performance and the soar of sharing data capacity with the Internet democratization. As pointed out by Chen (2016), the number of possible floating-point operations per second in CPU and GPU has dramatically increased over the last ten years (see Chen (2016)[Figure 4]). The Moore's law, introduced for the first time in 1975 Moore (1975), predicted that the number of components for each chip would double every single year. This early prediction has been proven to be true until now. From 1991 to 2011, the microprocessors performance has grown

[1]For example, the open letter Hawking et al. (2015) has been signed by researchers in artificial intelligence and robotics as well as other non-scientist notorieties, such as Elon Musk, CEO of SpaceX and Tesla Inc., who stated that artificial intelligence is one of today's "biggest existential threats" (Crawford (2016); Markoff (2015)).

1000-fold. This dramatic increase mentioned in Borkar and Chien (2011) is supposed to face new challenges. The energy is becoming the limit and will curb the increase of microprocessor frequency. As a result, large-scale parallelism is one of the promising paths to push the increase of performance. Storage capacity has dramatically increased while the price of storage has curbed. This is why it is possible to store more and more data. And with the increase of computer performance, this large amount of data can be processed at large scale[2]. The last needed improvement was the ability to share data and computer resources. With the Internet speed increase, this has been made possible.

On the one hand, it is now very easy to send large amounts of data through the Internet. Data can be more easily shared, mutualised and used. One agent can produce data while another agent can process or use the data. For example, a dramatic comparison that can be made is the first hard drive disk commercialized by IBM in 1956 (the RAMAC 350) and the last serie of hard drive built by IBM in 2002 (the star serie) (Wikipedia (2017)). While the RAMAC 350 had a storage capacity of 5 Megabytes, its actual volume was approximately two cubic meters; it required a power of 625 watt per Megabyte and its cost was $\$9,200$ per Megabyte. To get an idea of how big the hardware was at that time, Figure 1-1 shows the transportation of a RAMAC 350 in 1956. On the other hand, the Travelstar 80GN had a 80-Gigabyte capacity, it only requires a power of 0.02 watt per Megabytes and would only cost \$0.0053 per Megabyte[3]. Additionally, the fact that data can be relatively easily shared implies that computing resources can be outsourced and used when needed. Cloud computing solutions offer on demand computing resources. This has only been made possible by the ease of sending data over the net.

By facilitating the capacity of producing, sharing, storing and processing data, the aforementioned technological improvements reshape our economical environment. Data, and information extracted from it, become very valuable and strategical assets in our economy. Some companies, such as Alphabet (Google's parent company), offer free services in order to gather user data. Smartphones and the very common use of the Internet produce considerable amounts of personal data. These data are valuable for many purposes, such as getting insights on social trends, improving marketing strategies by using customer data, or even training and reinforcing predictive artificial intelligence algorithms which require important amounts of data to be trained. As data become valuable assets with very strong potential due to computer performance and artificial intelligence progress, it is clear that institutions are urged to regulate the use of data and define some ethics guidelines.

[2]Of course, the complexity of computational algorithms to process data is a key limit in the capacity of current processes. It will be one of the topics we will take into consideration in this thesis.

[3]And the Travelstar 80GN was produced 15 years ago from the time we write this thesis.

Figure 1-1: A 5-Megabyte IBM hard drive transported in 1956. *Photo credit: IBM Company.*

If everyone talks about artificial intelligence, we should be aware that the biggest breakthroughs in artificial intelligence are algorithms based on statistical methods with great computing implementations. The algorithm AlphaGo, described in Silver et al. (2016), has been developed by Google and uses tree search and neural networks. Autonomous car innovations mainly rely on image recognition, object detection and trajectory decision. The state of the art algorithms to achieve these operations use Machine Learning algorithms such as support vector machine, as in Levinson et al. (2011), and/or neural networks, as in Pomerleau (1991). Machine and statistical learning are subareas of artificial intelligence. Arguably, not every method used in artificial intelligence comes from Statistics or Machine Learning. For example, the study Olmstadt (2000) describes the early expert systems only which used human knowledge, in which there were no learning steps in the process. However, the most intricate decisions and challenging operations are based on learning new representations of the environment and detecting patterns in order to take decisions, which are tasks achieved by statistical methods. This is the goal of Machine Learning methods, that we will not differentiate from statistical learning in this thesis. In order to provide the reader with a better understanding of what is Statistics and of the context of this thesis, we will define some concepts that will be used throughout this thesis.

1.1.2 Definition of Statistics

According to Donoho (2015), the term and the use of Statistics were introduced 200 years ago along with the need to collect census data about the inhabitants of a given country. The statistical tools have been limited for a long time by the size of the data and the capacity to store and process the information, no computer being available. The introduction of the first automated systems, such as punch card tabulators, was the beginning of the capacity of scaling the amount of data that could be stored and eventually processed. From this point of view, Statistics is a boundless field that could be summarized in a very large sense as the definition of Statistics given by Agresti and Finlay (1997).

Definition 1.1.1 (Statistics)**.** *Statistics consists of a body of methods for collecting and analyzing data.*

Defined as such, some questions have been treated by statisticians over time, from both theoretical and empirical points of view. With no claim of being exhaustive, we can mention:

Data collection and storage This subarea tackles some questions such as the type of data that should be collected, and the way the data should be stored, referenced and organized. Polling has been a very deeply studied subject, asking some questions such as how many observations we need or how we can collect a survey with a small bias. On a larger extent, some questions about compressing data to limit storage costs, while loosing as little as possible information can be seen as part of this area.

Inference Statistical inference is a set of data analysis methods that help interpreting empirical observations. The purpose of estimation is often to understand a phenomenon by evaluating parameters that could explain the behaviour of a studied sample, or even a larger set that is supposed to be well represented by this given sample. The term of causal inference is used when the goal is to explain the causal interactions in a given model. The book Pearl et al. (2016) provides an excellent explanation of the difference between simple association and causal relationships in Statistics.

Prediction According to Shmueli (2010), predictive modeling is the process of applying statistical methods in order to predict the observation of a new individual. We will go into further details later in this section. From Donoho (2015), statistical prediction is defined as *predicting what responses are going to be for future inputs.*

Quality assessment The question of measuring the quality of statistical methods has been of paramount concern throughout the history of Statistics. How accurate is a prediction? How representative is a modeling inference to the real observations? Such questions are at the core of the statistical theory discipline.

This list, far from being exhaustive, comes under a plethoric literature. These different fields of study have very strongly developed from some simple results to very complex and subtle results. Some estimation methods have been thoroughly studied and the literature guarantees that the quality of these estimations are well understood in a given context.

If such a definition of Statistics is very large, it seems to conflict with other disciplines such as Machine Learning. In the following section, we will point out some similarities and differences between Statistics and Machine Learning. As it is not the topic of this thesis, we discuss it very briefly. For any reader who wishes a deeper understanding of the definition of these fields, we could only recommend the papers Donoho (2015), Breiman (2001), and the book Wasserman (2013), that are great food for thoughts on these matters.

Machine learning is very similar to Statistics. In view of Definition 1.1.1, both are studying methods to collect and analyze data. Since Statistics is a much older discipline than the invention of computers, statisticians could claim that Machine Learning is a mere clone of Statistics. However, the origins of Machine Learning differ from those of Statistics. This is why these two communities have distinct terms and sometimes different purposes. According to Wasserman (2013), Machine Learning comes from computer science departments. In Wasserman (2013)[Preface], a table of vocabulary equivalence between Machine Learning and Statistics is presented. It is worth remarking that there is a trend in both fields to share more and more terms. As an example, the term *learning* arised from Machine Learning is now very commonly used in the area of statistical learning. Originally, as statisticians did not have access to the calculus capacities of computers, low-dimensional problems were considered. Moreover, as pointed out in Breiman (2001), the issue of inference was initially prioritized over the question of prediction. This is explained by a long tradition of Statistics as being the inference from a data sample of an unknown underlying generative models. On the contrary, the Machine Learning community often worked without probability assumptions on the model; consequently, the field has been mainly focused on the prediction issue and, of course, computational challenges. As a consequence, there are differences between Machine Learning and Statistics. However, as there exist communities within Statistics and Machine Learning, it seems fair to admit that these two disciplines are also two communities working on same challenges with different backgrounds.

The focus and vocabulary of these disciplines tend to converge over time.

1.1.3 Examples of applications

There is a tremendous amount of current and potential applications of Statistics in our environment. While some of these applications are very well-known, some others are used in our daily life without us being aware of it. Earlier, we mentioned Google AlphaGo (Silver et al. (2016)), the first algorithm which managed to outperform any living human at playing the board game Go. The paper Bouzy and Cazenave (2001) provides an analysis of Go from a statistical point of view. Other games are being, or have been, learned by algorithms and are strongly advertized. Of course, our first thought goes to the famous IBM Deep Blue algorithm that defeated the best chess players in the world, as explained in Campbell et al. (2002) and Sutton and Barto (1998). The book Hsu (2002) provides interesting insights on Deep Blue story. It is worth noting that building a world-class algorithm playing Go has been more challenging than developing a Chess computer. The main reason of this difficulty gap is mainly the much higher number of possible combinations in Go, as explained in Burmeister and Wiles (1995).

Other applications are of paramount importance in our society. The healthcare industry has strongly benefited from statistical algorithms. The diagnosis of various diseases can be automated or assisted. The review in Kononenko (2001) mentions several statistical methods such as naive and semi-naive Bayesian classifiers, k-nearest neighbors, neural networks or decision trees. The article Dreiseitl et al. (2001) compares algorithms such as logistic regression, artificial neural networks, decision trees, and support vector machines on the task of diagnosing pigmented skin lesions, in order to distinguish common nevi from dysplastic nevi or melanoma. The authors of Shipp et al. (2002) describe a method to ease the detection of blood cancer.

The discovery of new drugs in the pharmaceutical industry is a growing challenge. The more active compounds are discovered, the less likely it is to discover a new drug with positive impact. In order to keep innovating, the pharmaceutical industry needs to increase the capacity of screening active compounds. Standard high-throughput screening methods become more and more costly as the number of active compounds already tested increases. As in ranking Agarwal et al. (2010), Machine Learning solutions enable to screen million of active compounds to rank them according to their likelihood to match a given response target. Statisticians and computer scientists developed digital high-throughput screening solutions based on support vector machine methods Burbidge et al. (2001) and neural networks Byvatov et al. (2003). The domain of active learning (Warmuth et al., 2003) plays an important role in the early phases

of drug discovery.

Image analysis plays a central role in radiology and medical imaging. Numerous examples are developed in the literature. Methods to detect microcalcifications from mammograms are compared in Wei et al. (2005). These methods include support vector machine, kernel Fisher discriminant and ensemble averaging. A review of various radiology applications is made in Wang and Summers (2012). Image segmentation, computer-aided diagnosis, neurological diagnosis are among the most astonishing applications of Machine Learning. Artificial intelligence can support the field of medicine with many other applications, such as health monitoring devices that can analyze data from patients (Bacci, 2017; Boukhebouze et al., 2016; Graham, 2014; Roux et al., 2017). On a larger scale, the detection of epidemiological outbreaks (Aramaki et al., 2011; Culotta, 2010) can be done with Bayesian network modeling, as in Wong et al. (2003), or with support vector machine and logistic classification as in Adar and Adamic (2005). As mentioned earlier, the field of autonomous robotics (Thrun et al., 2001) is one of the applications of Machine Learning and the last decade has seen the invention of many autonomous vehicles such as automated drone, unmanned aircrafts (Austin, 2011) or autonomous cars. This field of applications relies on statistical methods such as Monte Carlo simulations (Thrun et al., 2001), neural networks (Pomerleau, 1991) or fuzzy logic (Driankov and Saffiotti, 2013). The aeronautic and defense industries have also strongly benefited from signal processing and classification as in Zhang et al. (2004) and Zhao and Principe (2001), where wavelet support vector machines are used to automate recognition from radar data. Security is not the only major concern in the air. Facial recognition (Jain and Li, 2011) has been used for security purposes to grant access (Liu et al., 2005) or for surveillance by national authorities (Gilliom, 2001; Haque, 2015), which of course raises some ethical and philosophical questions about citizen freedom (Introna and Wood, 2004).

On a very different focus, recommendation engines have been a pretext for deep researches in Statistics. The Machine Learning discipline has benefited from the famous Netflix challenge (Bell and Koren, 2007; Zhou et al., 2008), which consisted in developing an algorithm to recommend to a user a list of movies he or she may enjoy. Some early applications occurred in the industry of online music radios. The two leaders were Pandora and Last.fm, but they had very different approaches. On the one hand, Pandora used content-based filtering as in Mooney and Roy (2000). The content-based approach consists in using features of songs[4] and some feedback from each user. If one user likes a given song, the content-based algorithm aims at recommending similar song. On the other hand, Last.fm used collaborative filtering

[4]See the Music Genome Project for more detailed information about such features (John, 2006).

as in Breese et al. (1998). Collaborative filtering is a very different approach that relies on the analysis of the behaviour of a community of users. Such methods analyze the behaviour of one user (such as a list of regularly listened songs) and compare it against the habits of other users. Collaborative filtering algorithms would recommend songs that other users with similar behaviours listen to on a regular basis. More recent applications, such as SoundCloud, combine the two approaches. Youtube (Davidson et al., 2010) recommends to its users the next video they could watch and Amazon (Linden et al., 2003) increases its sales by proposing items that a customer may wish to acquire.

Amazon is one of the most active companies in the Machine Learning area. Thus, Amazon has developed other applications, such as personal assistants, that use speech recognition in order to comply with the request of a user. Other important companies have developed personal assistants, including Apple with SIRI, Microsoft with Cortana and Google with Google Assistant. Extended applications occurred with the development of home assistant devices such as Amazon Echo (which is based on Alexa technology), Apple Homepod and Google Home (Nijholt, 2008; Clauser, 2016). These applications rely on speech recognition and natural language processing (Cambria and White, 2014; Kumar et al., 2016; Bowden et al., 2017; Earley, 2015). The list of current and potential applications of statistical methods is very long, if not infinite. We have mentioned here some disruptive applications. Of course, other areas benefit from Machine Learning algorithms. In the financial industry, quantitative funds use and keep exploring Machine Learning algorithms. Google Ad system relies on ranking estimations to rank and manage bid allocation of advertising content. Media apply filtering algorithms to rank (Rusmevichientong and Williamson, 2006) and eventually broadcast contents. The Edge Rank algorithm has been developed by Facebook to evaluate the potential interest of a post with respect to a specific user (Pennock et al., 2000; Chen et al., 2010). Our email box benefits from spam detectors (Jindal and Liu, 2007). The use of chatbots to automatically assist customers, or new analytics brought to sport in order to entertain the audience and to increase athletes' performance, seem to bring promising applications as well.

The intention behind mentioning this ridiculous number of applications is to show that Statistics, in its broad meaning as defined in 1.1.1, has a strong impact on artificial intelligence and all the applications that have been so much publicized lately. Statistical (or Machine Learning) methods such as support vector machine, neural networks, naive Bayesian classifiers, penalized regressions, make these innovations possible.

We have listed examples of statistical applications and we described the relationship between Statistics and other names such as artificial intelligence and Machine Learning. In order to give

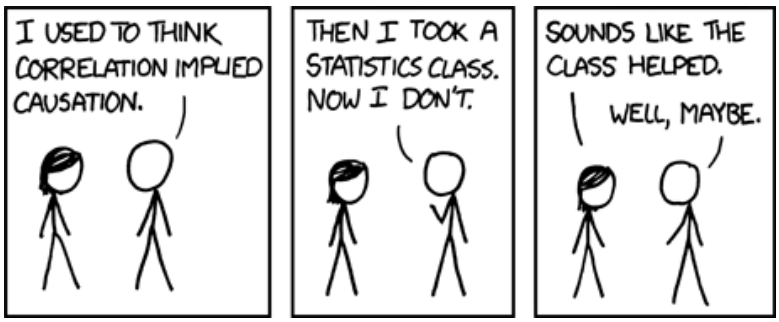


Figure 1-2: The risk of spurious correlation. *Credit:* xkcd.com

a more specific context to the study of this thesis, and close this non-quantitative introduction, we will introduce some important notions and concepts. Among them are the notions of unsupervised, supervised and semi-supervised learning, which describe the structure of the data with respect to the output.

1.1.4 Supervised, unsupervised and partially labeled learning

A great description of these three settings is provided in Wang and Summers (2012). In the supervised settings, there are two different types of data: inputs and outputs. The inputs are often named observations in Statistics and features in Machine Learning. The outputs are called outcomes in Statistics and labels in Machine Learning. The goal of Statistics in a supervised setting is to estimate a relationship between the inputs and the outputs. The relationship does not need to be a causal effect; it can be fortuitous (c.f. Figure 1-2). Most famous examples of supervised problems are the linear and non-linear regressions as well as the classification. Examples of supervised studies can be found in (Garcia et al., 2013; Bates et al., 2014; Aphinyanaphongs et al., 2014; Tibshirani, 1996b).

The second setting is the unsupervised learning. In that case, there is no output in the sample, only inputs. The goal of unsupervised learning is to infer some relationships within the inputs. It is often assumed that there is an underlying latent variable that explains the relationship and behaviour of the observations. Well studied examples of unlabeled learning are density estimation, clustering and anomaly detection (Zeng et al., 2014; Le, 2013; Cheriyadat, 2014; Zimek et al., 2014; Dinh et al., 2016; Costa et al., 2015; Gupta et al., 2014).

The partially labeled learning setting is an intermediate form of supervised learning. In that setting, some observations are associated with labels while other inputs have no corresponding outputs. The purpose of partially labeled learning is similar to the one of supervised learning,

which is to estimate a relationship between the inputs and the outputs. However, one may want to improve the quality of the estimation by using structural information of additional unlabeled features. Most of the time, semi-supervised learning is used when there are a few labeled data and a large amount of unlabeled data available. Within the partially labeled setting, two types of tasks might be considered. The semi-supervised learning tasks consist in estimating a predictor that minimizes a risk as in Equation 1.2.3. On the other hand, the task of transductive learning is to predict the unknown outcomes of the unlabeled data that are within the original dataset. These settings will be further discussed in Chapter 4.

It is important to note that partially labeled learning differs from active learning. Active learning tasks consist of estimating a predictor from unlabeled data where an algorithm can request interactively the desired outputs of new data points in order to increase the quality of the estimate.

We have described the notions of supervised, non-supervised, and partially labeled learning, as well as semi-supervised, transductive and active learning tasks. In Chapter 4, we will consider the case of partially labeled setting and address some theoretical questions in the context of penalized regression.

In the next section, we will introduce some fields of statistical learning which are related to the work presented in this thesis. The goal of this section is to provide the reader with a global view of the literature related to our study. As this thesis is about theoretical statistics, we will mainly focus on papers that present an interest from the theoretical point of view. Besides, as it is at the intersection of many concepts, we believe it is worth going through a certain number of theories such as high-dimensional statistics, PAC-Bayesian estimation, aggregation, oracle paradigm, regularization and penalized regression. We will also discuss some theoretical results from the literature that address Monte Carlo computational challenges. It will nurture the last chapter of this thesis (Chapter 4).

1.2 Challenges in high-dimensional statistics

1.2.1 High-dimensional statistics

There has been a crucial shift of paradigm in the last decades. Let $n \in \mathbb{N}$ be the number of observations and $p \in \mathbb{N}$ be the number of features. The standard statistical framework considers the case where n is relatively large and p substantially smaller than n. As pointed out in Giraud (2014), the technological evolution of computing has urged a shift of paradigm

from classical statistical theory to high-dimensional statistics. In the high-dimension settings, a very large number of features p is considered and the number of observations n is of the same order of p, if not smaller.

The practical interest of the high-dimensional theory has come with the increasing number of data gathered by any connected object that can collect thousands to millions of different features. The paper Donoho (2000) gives a very clear overview of the particularities of high-dimensional statistics in comparison with classical statistics. This paper mentions that a very large number p of features is not a blessing but a curse. Indeed, the very large number of features may sound like an opportunity to obtain a very thorough and complete quantity of information in order to infer or predict potentially anything. However, the difficulties in a high-dimensional settings are many. The computational challenge in a high-dimensional settings is of course one of the main concerns and supposes algorithms to be computable in polynomial time with respect to p and n. In high-dimensional statistics, some methods priorize the computational complexity over the estimator optimality. For example, the optimal method to detect sparse principal components of high-dimensional variance-covariance matrix requires to solve a NP-complete problem. The authors of Berthet and Rigollet (2013) propose an alternative method that is nearly optimal in the detection level but guarantees the solution to be computable in polynomial time. In Chapter 4, we propose an algorithm to approximate a certain class of estimates and we prove that any targeted accuracy of the approximation can be reached in an explicit polynomial time.

In some situations, the high-dimensional context induces some heterogeneous collectin of data. A common situation is when a large dataset can be collected automatically at a low cost, while the label is difficult or costly to collect. In that case, it is relevant to collect a very large number of observations with no label and to manually collect labels associated with a few observations. This situation refers to partially labeled learning setting that we already mentioned earlier. This context will be discussed in the Chapter 3 and a review of partially labeled classification can be found in Schwenker and Trentin (2014) for further details on that matter.

On the other hand, some theoretical properties of the high-dimensional settings require to face other challenges. The survey Zimek et al. (2012) explains well the difficulty that a large value of p induces when detecting outliers in a data set. This has strong consequences in high-throughput screening of molecules in the pharmaceutical industry, where the goal is to detect non-zero effect of a given feature as mentioned in the introduction of the book Bühlmann and Van De Geer (2011). Controlling false discovery becomes more difficult. A set of results on concentration inequalities is very useful to circumvent the challenge of controlling extreme

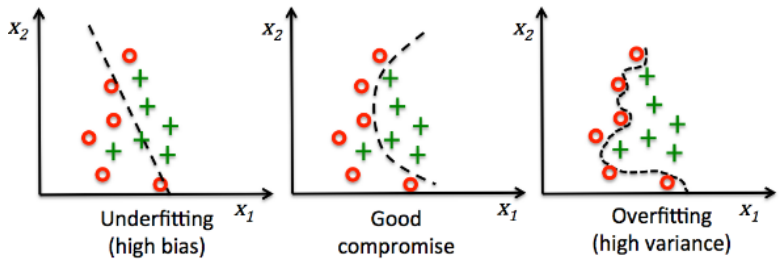


Figure 1-3: A representation (from left to right) of underfitting, balanced and overfitting classification learners. *Credit: Python Machine Learning - Sebastian Raschka*

values, when p is large. The book Ledoux (2005) and the paper Boucheron et al. (2013) provide handy results on this matter.

Another famous phenomenon, when p is relatively large, is the absence of property of convergence of the variance-covariance matrix estimation. For example, in the trivial case where n random variables are sampled from a normalized Gaussian density of dimension p, $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, the empirical covariance matrix does not converge almost surely toward the true covariance matrix which is the identity $\mathbf{I}_p$ if p is of the order of n or larger. In the paper Ravikumar et al. (2011), a method is proved to be an appropriate estimator of the covariance in the high-dimensional settings.

Another theoretical challenge, and arguably the most famous, is the overfitting risk. As many features are available, it is tempting to use too many parameters to build an estimate relatively to the number of observations. However, if such an estimate has a very good fitting quality, the predictive risk on a new observation may be large. In that case the learning algorithm is said to have high variance. An excessively simplistic model has often high bias as it fails to model the actual connexion between the variables and the outcomes.

The overfitting phenomenon is widely discussed in the literature on different learning tasks such as regression (Hawkins, 2004; Hurvich and Tsai, 1989), classification (Khoshgoftaar and Allen, 2001; Hsu et al., 2003) or outlier detection (Abraham and Chuang, 1989; Pell, 2000), and many methods have been developed such as cross-validation (Hsu et al., 2003; Refaeilzadeh et al., 2009; Ng, 1997), regularization (Tibshirani, 1996b; Bogdan et al., 2015; Zhang and Oles, 2001; Bickel et al., 2009) or Bayesian prior (Cawley and Talbot, 2007; Park and Casella, 2008; Wipf and Rao, 2004) or pruning (Bramer, 2002). We will discuss some of these methods later in this thesis.

With the exception of cross-validation, these methods use a key concept of the high-dimensional

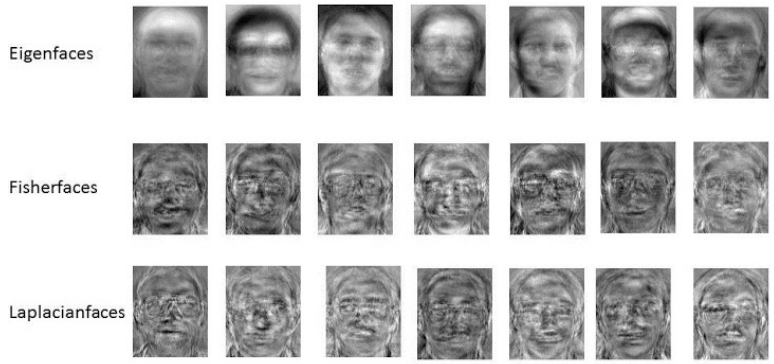


Figure 1-4: Various bases can be used to represent face images in order to approximate faces in a sparse representation. From top to bottom eigenfaces, fisherfaces and laplacianfaces are used. *Credit: The Github repository Face Recognition from Wihoho* - https://github.com/wihoho/FaceRecognition

settings: the sparsity. From a theoretical point of view, with no further assumptions, guarantees on estimation risk would be very difficult to prove. However, in empirical situations, it is very common to observe an underlying model that explains the generation of the data (from an inference point of view) or predict well the labels (from a predictive point of view) from a lower dimensional representation. Indeed, a lot of observed phenomena are in very high-dimensional settings but are actually governed by underlying patterns that can be explained in a much smaller dimension (at least approximately).

In signal processing, some sparse representations are infered using bases such as wavelets (Mallat, 1999).

1.2.2 Sparsity

Conceptually, the sparsity paradigm is far from being new. In the early 14^{th} century, Occam introduced the Occam's razor principle[5]. The Occam's razor is a principle making a recommendation to choose among several possible explanations of a phenomenon. Occam suggests that it is best to choose the simplest possible model among the models that fit the observations with a relative accuracy. In statistical terms, it means that among all the models that fit relatively well the dataset, it may be better to choose the simplest model. In high-dimensional statistics, the sparsity settings is a necessity. In Bühlmann and Van De Geer (2011), there is a rough expression of a condition on the sparsity level with respect to p and n required to achieve

[5]In Latin, Occam's concept is mentioned by the term *lex parsimoniae*, which reads as the law of parsimony.

Figure 1-5: An illustration of learning the empirical average of a face to increase robustness in image classification. *Credit: Deric Bownd, Deric's Mindblog*

good estimation:

$$s \log(p) \ll n, \tag{1.2.1}$$

where s is the sparsity level that will need to be defined. The definition of sparsity can differ according to the settings and the learning task. However, in many situations, there is an equivalence of the notion of sparsity.

For example, in the linear regression settings, the sparsity level is the number of non-zero parameters. Let consider some data that consist of n observations of random outcomes $y_1, \dots, y_n \in \mathbb{R}$ and p fixed covariates $\boldsymbol{x}^1, \dots, \boldsymbol{x}^p \in \mathbb{R}^n$. Let assume there is an unknown vector $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that the residuals $\xi_i = y_i - \beta_1^\star \boldsymbol{x}_i^1 - \ldots - \beta_p^\star \boldsymbol{x}_i^p$ are independent, zero mean random variables. In vector notation, this reads as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi}, \tag{1.2.2}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ is the response vector, $\mathbf{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{\xi}$ is the noise vector. For the sake of simplicity, let assume the noise vector to be distributed according to the Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, with σ a relatively small quantity in comparison to the variance of the elements of $\boldsymbol{y}$. If the quantity $\|\boldsymbol{\beta}^\star\|_0$ of non null element of $\boldsymbol{\beta}^\star$ is equal to s, then the linear model of y admits a s-sparse representation with respect to the features of the design matrix $\mathbf{X}$. If s is substantially smaller than the dimension p, it means that only a few features are explaining (or predicting) the outcome $\boldsymbol{y}$.

Numerous patterns of sparsity exist. In order to describe the concept of sparsity in a general settings, we present the problem of regression in a general settings. Let consider the pair $(\boldsymbol{X}, \boldsymbol{y})$ drawn from a distribution P on a product space $\mathcal{X} \times \mathcal{Y}$, we aim at predicting $\boldsymbol{y}$ as a function of $\boldsymbol{X}$. Mathematically speaking, we want to estimate a function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the

risk,

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} l(y, f(\boldsymbol{x}))\, P(\mathrm{d}\boldsymbol{x}, \mathrm{d}y), \tag{1.2.3}$$

where l is an arbitrary loss function. We define $f^\star$ the function that minimizes Equation 1.2.3. The regression problem can then be written in the form

$$\boldsymbol{y} = f^\star(\boldsymbol{X}) + \boldsymbol{\xi},$$

where $\boldsymbol{\xi} \in \mathbb{R}^p$ is a vector of random variables. For example, if l is the quadratic loss function, then $f^\star$ is the Bayes predictor,

$$f^\star(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}],$$

and the noise $\boldsymbol{\xi}$ is such that,

$$\mathbb{E}[\boldsymbol{\xi}|\boldsymbol{X}] = \mathbf{0_p}. \tag{1.2.4}$$

for any $\boldsymbol{\beta} \in \mathbb{R}^p$. Then, the model admits a s-sparse representation if there exists $\boldsymbol{\beta^\star} \in \mathbb{R}^p$ such that $f^\star = f_{\beta^\star}$ and $\|\boldsymbol{\beta^\star}\|_0 \leq s$. In the sparsity approach theory, the goal is to detect a pattern within the data that can be approached with a sparse representation, but the linear representation may not be the right sparse representation. In the aforementioned case of linear regression, we consider a subset of functions f such that f can be written as a linear relationship betwen the data $\mathbf{X}$ and the label $\boldsymbol{y}$,

$$\boldsymbol{y} = f_{\boldsymbol{\beta}}(\mathbf{X}) + \boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{1.2.5}$$

For example, in the task of piecewise constant regression, the goal is to estimate a piecewise constant function f on a given interval I defined by:

$$f^\star(x) = \sum_{t\in T} c_t \mathbb{1}_{x>t}, \tag{1.2.6}$$

where $T \subset I$ is a finite set of elements. The goal of a piecewise constant regression task is to propose an estimation f of the function $f^\star$ while we observe n discrete random variables $(y_i)_{i\in[n]}$ such that

$$y_i = f^\star(x_i) + \xi_i, \tag{1.2.7}$$

where $x_i \in I$.

If the set of rupture points T is small, then there is a sparse structure in the model in the sense that there are very few variations. However, this is a different type of sparsity than in

the standard sparse linear model, since the support of the function $f^\star$ is not sparse. The sparse regression of piecewise constant time series has been studied in numerous papers including (Bleakley and Vert, 2011; Hocking et al., 2013; Antoch and Jarušková, 2000). In Giraud (2014), several sparsity patterns are listed, including coordinate and variation[6] sparsity, that we just mentioned, but also group, sparse-group and basis sparsity that are important subjects in the literature (Reid, 1982; Elad and Bruckstein, 2002; Friedman et al., 2010; Meier et al., 2008). It is interesting to note that this typology of sparsity patterns can be generalized by remarking that they can all be considered as a sparse representation with respect to a specific representation. Indeed, let consider the noise-free oracle signal $f^\star(\mathbf{X})$, all the aforementioned learning contexts can be recasted as representing $f^\star(\boldsymbol{x}_i) = f_i^\star$ as a scalar product between a given vector $\boldsymbol{\beta}^\star$ and a well suited basis ψ of vectors $(\boldsymbol{\psi_i})_{i\in[n]}$ such that $f_i^\star = \langle \boldsymbol{\beta}^\star, \boldsymbol{\psi_i} \rangle$ for any $i \in [n]$. Using this representation, it is possible to define with consistence the notion of sparsity within a given basis.

Definition 1.2.1 (s-sparsity)**.** *Let $s \in \mathbb{N}$, the signal $\boldsymbol{l} = (l(x_i))_{i\in[n]}$ is said to admit a sparse representation with respect to ψ if there exists a vector $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that $\|\boldsymbol{\beta}^\star\|_0 \leq s$ and*

$$l_i = \langle \boldsymbol{\beta}^\star, \boldsymbol{\psi_i} \rangle, \tag{1.2.8}$$

for any $i \in [n]$.

The parameter s of Definition 1.2.1 may also be referred as the intrinsic dimension of a problem by some authors of the literature such as Guedj and Alquier (2013).
Since many regression problems can be recasted as a linear regression task, with respect to a given basis, we will only consider the linear case. Consequently, $f^\star$ will be represented by the parameter $\boldsymbol{\beta}^\star$ that we name the oracle parameter and we will make no distinction between the data $\mathbf{X}$ and the basis ϕ, with no loss of generality.

1.2.3 Prediction risk and oracle inequality

Unfortunately, the oracle is not known in practice. Indeed, the knowledge of $f^\star$ requires a perfect information over the model. Therefore, one may need to estimate $\widehat{\boldsymbol{\beta}}$ from the given observation $(\boldsymbol{y}, \mathbf{X})$. It would be great to obtain guarantees on the risk of this estimator in the form

$$\mathcal{R}(f_{\widehat{\boldsymbol{\beta}}}) \leq C, \tag{1.2.9}$$

[6] Variation sparsity is also known as fused sparsity.

where $C \in \mathbb{R}$ is a constant. Ideally, for a given estimate, this guarantee would hold in any condition, with very few assumptions with no dependency on the data $\mathbf{X}$ neither the noise $\boldsymbol{\xi}$. Unfortunately, the prediction risk $\mathcal{R}(f_{\widehat{\boldsymbol{\beta}}})$ may differ from one problem to another. Indeed, $f_{\widehat{\boldsymbol{\beta}}}$ is not the only parameter that plays a role on the risk. The level of noise ξ, and the level of information that $\mathbf{X}$ represents with respect to $\boldsymbol{y}$, have both a strong impact on the risk. The smaller the noise to signal ratio, the smaller the risk can potentially be. From this point of view, the difficulty of the regression task differs. As a consequence, it is often impossible to obtain a guarantee of the quality of an estimator independently to the coherence of the model defined by the data $\mathbf{X}$. Even though an absolute guarantee is not reasonable, it is possible, for some specific estimators, to compare the risk of the estimator to the risk of the oracle. This type of comparison is known in the literature as oracle inequalities and enables the guarantees of estimators to be studied with the consideration of the difficulty of a regression problem. The paper Candes (2006) reviews the powerful concept of oracle inequalities.

Definition 1.2.2 (Oracle risk inequality)**.** *Let $f_{\widehat{\boldsymbol{\beta}}}$ be the prediction associated to the estimator $\widehat{\boldsymbol{\beta}}$ and let consider the regression problem defined in Equation 1.2.2, then the estimate $\widehat{\boldsymbol{\beta}}$ is said to admit an oracle inequality with respect to an arbitrary loss function l, if there exists a constant $C_1 \in \mathbb{R}$ and a function C_2, depending on the model defined in Equation 1.2.2, such that:*

$$l(f_{\widehat{\boldsymbol{\beta}}}, f^\star) \leq C_1 \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ l(f_{\boldsymbol{\beta}}, f^\star) + C_2(\mathbf{X}, \boldsymbol{\xi}) \right\}. \tag{1.2.10}$$

Furthermore, when $C_1 = 1$, the oracle inequality is said to be sharp.

It is worth remarking that Inequality 1.2.10 guarantees that there is no estimator $\boldsymbol{\beta} \in \mathbb{R}^p$ that is much better (in the sense of C_2) than $\widehat{\boldsymbol{\beta}}$. In particular, it implies that for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$l(f_{\widehat{\boldsymbol{\beta}}}, f_{\boldsymbol{\beta}}) \leq C_2(\mathbf{X}, \boldsymbol{\xi}). \tag{1.2.11}$$

In particular, Inequality 1.2.11 holds for $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$.
In that case, the inequality links the performance of the estimator $\widehat{\boldsymbol{\beta}}$ to the performance of the oracle $\boldsymbol{\beta}^\star$. The study of the quantity C_2 in Inequality 1.2.10 is of great interest, it is the rate of convergence of the estimator $\widehat{\boldsymbol{\beta}}$. In many cases, C_2 decreases with the number of observations n and increases with the variance σ^2 of the noise $\boldsymbol{\xi}$ and with the dimension p. The optimality of the rate C_2 is a well studied question in the literature of many estimators. Moreover, the question of achieving a fast rate of the order $\sigma^2 \log(p)/n$ is a common benchmark in the linear regression problem.

1.2.4 Oracle inequality in the sparsity context

In the context of the s-sparse assumption, one may consider all the supports with s or less elements. In that case, the goal is to obtain a s-sparse estimator with a small risk. The s-sparse context is deeply studied in the literature (Bickel et al., 2009; Bellec et al., 2016b; Giraud, 2014). It is worth remarking that two types of results may be of interest with respect to the oracle inequalities. On the one hand, the first type of results are guaranteed provided that the true signal $\boldsymbol{\beta}^\star$ is s-sparse. These results are said to consider well-specified learning tasks. On the other hand, some results do not rely on the assumption of a well specified signal $\boldsymbol{\beta}^\star$, this is the notion of ill-specified oracle inequalities. In view of Definition 1.2.2, well-specified oracle inequalities in the well specified case are of the form:

$$l(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \leq C_1 \inf_{\boldsymbol{\beta} \in \mathbb{R}^p_s} \left\{ l(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) + C_2(\mathbf{X}, \boldsymbol{\xi}) \right\}, \tag{1.2.12}$$

where $\mathbb{R}^p_s$ is the set of s-sparse vectors of dimension p. This type of inequality guarantees that the estimator performs nearly as well as any other s-sparse parameter. In the ill-specified case, this inequality has to hold for any $\boldsymbol{\beta} \in \mathbb{R}^p$. Of course, the second case is more difficult to prove. The function C_2 is expected to be larger in the ill-specified case than in the well-specified case. It is also of interest to quantify the degradation of this inequality when it is no longer a well-specified case. In the results of Chapter 2, we propose oracle inequalities that consider the ill-specified case and explicit the degradation of the inequalities due to the relaxation of the sparsity assumption.

In the sparse regression settings, with the belief that an unknown underlying sparse pattern represents the signal, or at least predicts an important portion of the signal, it is interesting to recover the sparsity pattern. Let S be the support of the true parameter $\boldsymbol{\beta}^\star$, the subset of non null elements of $\boldsymbol{\beta}^\star$, and let consider a computable loss criteria such as, for example,

$$\ell_n(\boldsymbol{\beta}) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{1.2.13}$$

for any $\boldsymbol{\beta} \in \mathbb{R}^p$; then, if S was known, a good strategy would be to define the estimate as the vector that minimizes Equation 1.2.13 among all the vectors whose support is S. However, this under constraint minimization problem is not feasible since one does not know S. The task of selecting a model within a set of models $\mathcal{S}$ often refers in the literature to the task of computing an estimate $\widehat{S}$ as an adequate support[7]. The best support among the set of support $\mathcal{S}$ is the

[7]What is considered adequate may be interpreted in different ways depending on the context. In this thesis,

oracle model.

Definition 1.2.3 (Oracle support)**.** *Let consider the problem of regression as defined in Equation 1.2.2 and a loss function ℓ. Let $\mathcal{S}$ be a set of supports in $\mathbb{R}^p$. Then we define the Oracle support with the set $\mathcal{S}$ as*

$$S^\star_{\mathcal{S}} = \arg\min_{\bar{S}\in\mathcal{S}} \mathbb{E}\Big\{\ell\big(\boldsymbol{\beta}^\star, \boldsymbol{\beta}^\star_{\bar{S}}\big)\Big\}, \tag{1.2.14}$$

where $\boldsymbol{\beta}^\star_{\bar{S}}$ is the best estimator with the support $\bar{S}$. The unknown parameter $\boldsymbol{\beta}^\star_{\bar{S}}$ is the oracle parameter on the support $\bar{S}$.

Unfortunately, the oracle support cannot be calculated from the data since it relies on the unknown signal $\boldsymbol{\beta}^\star$ and the unknown distribution P. A possible approach is to estimate the expected risk by an empirical loss function (such as 1.2.13) and define the estimated support such as

$$\widehat{S}_{\mathcal{S}} = \arg\min_{\bar{S}\in\mathcal{S}} \Big\{\ell_n\big(\widehat{\boldsymbol{\beta}}_{\bar{S}}\big)\Big\}, \tag{1.2.15}$$

where $\widehat{\boldsymbol{\beta}}_{\bar{S}}$ is the estimate with support $\bar{S}$ that minimizes the empirical risk $\ell_n(\boldsymbol{\beta})$, as defined in Equation 1.2.13.

After estimating a support $\widehat{S}_{\mathcal{S}}$, an intuitive estimate of the regression problem would be $\widehat{\boldsymbol{\beta}}_{\bar{S}}$ as defined in Equation 1.2.15. However, such methods present a risk of bias. Indeed, the theoretical risk is likely to be underestimated when estimated with the empirical risk in the high-dimensional settings. This phenomenon is known under the name of overfitting. The empirical risk decreases with complex models that do not benefit from good properties on new data. In order to compensate this bias due to overfitting, some unbiased risk estimates have been proposed. These estimations combine an empirical loss function that measures the fitness of the parameter of the model and a penalty function that increases with the dimension of the selected model. Historically, some early penalization estimations of the risk have been proposed, such as the Akaike criterion:

$$\widehat{S}_{AIC} = \arg\min_{\bar{S}\in\mathcal{S}} \big\{\ell_n(\widehat{\boldsymbol{\beta}}_{\bar{S}}) + 2\sigma^2\|\widehat{\boldsymbol{\beta}}_{\bar{S}}\|_0\big\}. \tag{1.2.16}$$

Even though the limits of the AIC estimator have been shown, it has opened the field of penalization methods that benefit from strong properties. More specifically, the AIC estimator can be seen as a maximum a posteriori estimator.

The remaining of this introductory chapter will give a brief overview of the litterature of results on prediction by penalized (Section 1.3) and averaged (Section 1.4) estimators in fixed and

we will focus on prediction tasks.

random designs.

1.3 Maximum a posteriori estimation

Maximum a posteriori estimation is one of the most used and studied class of estimators. This class of estimators is very similar to the maximum likelihood approach. However a prior on the parameter distribution is included in the optimization problem. Hence, considering the prior as a regularization penalty, one may consider the maximum a posteriori estimations the penalized equivalent of the maximum likelihood approach. One of the most studied maximum a posteriori estimator is the Lasso introduced in Tibshirani (1996b) that uses a Laplace prior that is a sparse inducing prior.

1.3.1 Penalized regression and MAP

Penalization is an alternative approach to the question of the estimation under sparsity assumptions. Instead of estimating a support of the sparse pattern and then estimating the estimate within this subset of vectors, penalization consists of estimating both model and estimate in a global optimization problem.

Definition 1.3.1 (Linear penalized regression)**.** *Let consider a matrix* $\mathbf{X}$*, and some random vectors* $\boldsymbol{y}$ *and* $\boldsymbol{\xi}$ *in* $\mathbb{R}^n$ *and a parameter* $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ *such that*

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi},$$

then the parameter $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ *is said to be estimated from a penalized regression if it is the solution of an optimization problem of the form,*

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in E \subset \mathbb{R}^p} \big\{ L(\boldsymbol{\beta}) + \mathcal{P}(\boldsymbol{\beta}) \big\}, \tag{1.3.1}$$

where E *is arbitrarily chosen, and the function* L *is a function corresponding to a fitting criterion (such as* ℓ_n*) and* $\mathcal{P}$ *is a function that penalizes the non sparsity of* $\boldsymbol{\beta}$*.*

In penalized regression, the solution $\widehat{\boldsymbol{\beta}}$ of Equation 1.3.1 can be seen as the maximum a posteriori

(MAP) with respect to a negative log-likelihood L and a prior proportional to $\exp\{-\mathcal{P}\}$. Indeed,

$$\widehat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta} \in E \subset \mathbb{R}^p} \big\{ \exp(-L(\boldsymbol{\beta})) \exp(-\mathcal{P}(\boldsymbol{\beta})) \big\}. \tag{1.3.2}$$

From this point of view, an intuitive approach would be the Bayesian method that consists in averaging with respect to the prior instead of maximizing the posterior function. In that case, the pseudo-Bayesian estimator is of the form:

$$\widehat{\boldsymbol{\beta}}_B = \int_{E \subset \mathbb{R}^p} \boldsymbol{u} \pi(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \tag{1.3.3}$$

where π is the normalized posterior,

$$\pi(\boldsymbol{\beta}) = \frac{\exp(-L(\boldsymbol{\beta}) - \mathcal{P}(\boldsymbol{\beta}))}{\int_E \exp(-L(\boldsymbol{u}) - \mathcal{P}(\boldsymbol{u})) \mathrm{d}\boldsymbol{u}}. \tag{1.3.4}$$

Since a convex combination of sparse vectors is not necessary sparse, the weighted average is often not sparse. However, when the estimate does not require to be sparse, the average estimate may be of interest. In particular, when the quality criterion is the prediction, sparsity is rarely required. In that case, one may want to study the property of such estimates with respect to the pseudo-prior density.

As we will discuss later, Bayesian procedures do not always achieve good results for prediction tasks in the sparsity settings. The choice of the pseudo-prior density is of paramount importance. The literature of weighted aggregate estimation will be further discussed in this chapter. One important take-away is that the guarantees of weighted aggregate estimators are less understood by the statistical community than those of classical MAP estimators. The goal of Chapter 2 is to provide some understandings of the behaviours of a specific family of weighted aggregate estimators.

To sum up, there is a strong analogy between penalized estimators and weighted average ones. However, the state of the art has proven better oracle results in the classical settings of penalized regression than in the weighted average case. Without any mathematical justification, it seems more difficult to understand the behaviour of averages than of maxima. One of the goals of this thesis is to very modestly close this gap.

1.3.2 The Lasso estimator and related estimators

In the vectorial high-dimensional regression settings, the ℓ_1-penalized least squares estimator (Lasso) is very well known and is arguably one of the most studied estimators. The Lasso can be defined as the solution of the following convex problem:

$$\widehat{\boldsymbol{\beta}}_L \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \, \|\boldsymbol{\beta}\|_1 \right\}, \tag{1.3.5}$$

where $\lambda > 0$ is a tuning parameter. From the convexity of the ℓ_1-norm, this problem can be efficiently solved as described in the LARS algorithm Efron et al. (2004) and more recently in Bach et al. (2012). This estimator has been introduced by Tibshirani (1996b) and not only has it been known for its computational convenience. Since its introduction in 1996, it has been a well studied estimator as it benefits from high theoretical accuracy performances in the sparse and nearly-sparse settings. Many papers have studied the theoretical behaviour of the Lasso and it would be foolish to aim at providing a comprehensive review of the literature on this topic. Risk bounds have been proven for the Lasso for both prediction and inference purposes (Bühlmann and van de Geer, 2011; Koltchinskii, 2011; Dalalyan et al., 2014b; Bickel et al., 2009; Koltchinskii et al., 2011b; Bunea et al., 2007a; Zhang, 2009; Wainwright, 2009; Cai et al., 2010; Lounici, 2008; Meinshausen and Yu, 2009; Van De Geer, 2007; Zhang and Huang, 2008; Sun and Zhang, 2012c). The ℓ_1-penalized estimator has been generalized to the matrix settings (Koltchinskii et al., 2011b; Bickel et al., 2009), where the ℓ_1-norm is the nuclear norm. An extension to the total variation norm has been proposed in Dalalyan et al. (2014b). To the best of our knowledge, the first sharp oracle inequality with fast rate has been proven in Koltchinskii et al. (2011b)[Theorem 6.1]. Moreover, Sun and Zhang (2012b)[Theorem 4] provides a sharp oracle inequality with nearly fast rate. This result is also discussed in Dalalyan et al. (2014b)[Theorem 2]. This last result relies on the compatibility factor assumption. The compatibility factor of the design matrix $\mathbf{X}$ is defined, for any $J \subset [p]$ and $c > 0$, by

$$\kappa_{J,c} = \inf_{\substack{\boldsymbol{u}\in\mathbb{R}^p: \\ \|\boldsymbol{u}_{J^c}\|_1 < c\|\boldsymbol{u}_J\|_1}} \frac{c^2|J|\|\mathbf{X}\boldsymbol{u}\|_2^2}{n(c\|\boldsymbol{u}_J\|_1 - \|\boldsymbol{u}_{J^c}\|_1)^2}. \tag{1.3.6}$$

Theorem 1.3.1. *Sun and Zhang (2012b)[Theorem 4] – Dalalyan et al. (2014b)[Theorem 2] Let $\delta \in (0,1)$ and $\gamma > 1$ be arbitrarily chosen and the penalty parameter of the Lasso*

$$\lambda = \gamma\sigma^\star\Big(\frac{2}{n}\log(p/\delta)\Big)^{1/2}, \tag{1.3.7}$$

then with probability $1-\delta$,

$$\ell_n(\widehat{\boldsymbol{\beta}}_L, \boldsymbol{\beta}^\star) \leq \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p \\ J\subset[p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{2(1+\gamma)^2\sigma^{\star 2}|J|\log(p/\delta)}{n\kappa_{J,c}} \right\}, \tag{1.3.8}$$

where $\sigma^{\star 2}$ *is the known variance of the noise,* $c = (\gamma+1)/(\gamma-1)$ *and* ℓ_n *is defined in* (1.2.13).

Some recent works from Bellec et al. (2016b) have been exploring the properties of the Lasso estimator in comparison with the SLOPE estimator. To the best of our knowledge, they are the fastest oracle inequalities rates for the Lasso. In Bellec et al. (2016b)[Theorem 4.2] the rate of convergence is improved from $s/n\log(p)$ to $s/n\log(p/s)$, provided that the sparsity level s is known and that the restricted eigenvalue condition holds. Most of the oracle inequalities that apply to the Lasso require the confidence level δ to be tied to the penalty term λ. For example, 1.3.1 relies on Equation 1.3.7. The authors Bellec et al. (2016b) prove an oracle inequality with a rate $s/n\log(p)$ with a confidence level δ chosen irrespectively to the penalty term λ in Bellec et al. (2016b)[Proposition 3.2].

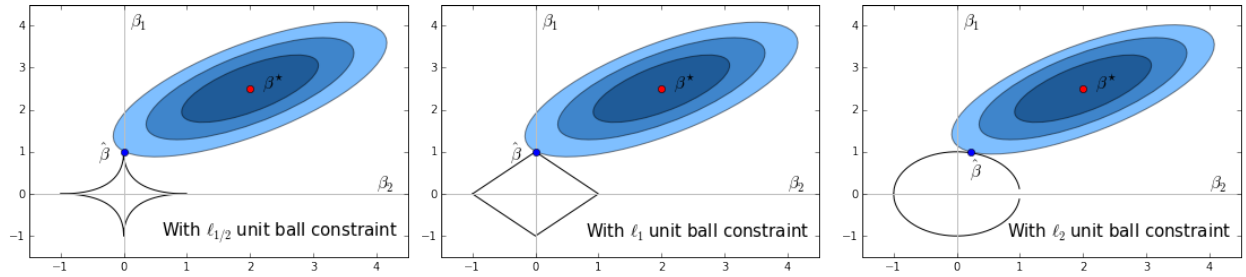
It is worth remarking that some estimators such as the Dantzig selector (Candes and Tao, 2007; Bickel et al., 2009) and the SLOPE (Bogdan et al., 2015; Su and Candes, 2016) benefit from a similar performance than the Lasso.

The Dantzig selector has been introduced by Candes and Tao (2007). The Dantzig selector is not defined by a penalized regression optimization problem but by the minimization of the ℓ_1-norm of the parameter under the Dantzig constraint Bickel et al. (2009). The seminal work of Bickel et al. (2009) compares the Lasso and the Dantzig selector and shows that both estimators have similar behaviour (Bickel et al. (2009)[Theorem 7.1] and additional studies have been carried out by Lounici (2008). Indeed, provided that the restricted eigenvalue condition holds, these estimators benefit from analogous theoretical bounds for both prediction risk and estimation loss in the linear regression[8].

The SLOPE is another interesting estimator that has been recently introduced by Bogdan et al. (2015) and Su and Candes (2016). The SLOPE is the solution of a penalized optimization problem. On the one hand, it differs from the Lasso in that the ℓ_1-norm is being substituted by a specific norm (Bogdan et al. (2015)[Proposition 1.2]) that depends on a vector $\boldsymbol{\lambda} \in \mathbb{R}^p$ of tuning parameters[9]. The very general results studied by Lecué and Mendelson (2016b) enable to prove error bounds on the SLOPE. On the other hand, in Bellec et al. (2016b), the

[8] A similar observation is also made in Bickel et al. (2009) in the nonparametric settings.

[9] See Bellec et al. (2016b)[Equation 2.2] for a definition of the norm and Bellec et al. (2016b)[Equation 2.4] for a definition of the SLOPE.



(a) The $\ell_{1/2}$-norm constraint induces sparse estimates but its non-convexity makes the computational cost of this estimator crippling.

(b) The Lasso induces sparse estimates and benefits from convex constraint.

(c) Even though the ℓ_2-norm penalization offers convex constraints, its geometrical shapes does not induce sparse estimates.

Figure 1-6: Estimation picture of three penalized regression methods in a two-dimensional setting. The red points represent the true parameters $\boldsymbol{\beta}^\star$ and the blue points the estimates with respect to the given penalization norm.

authors exhibit the strong relationship between the Lasso and the SLOPE by showing that oracle inequalities benefit from very similar rates from analogous conditions. In the case of the SLOPE, the conditions are slightly more restrictive (Bellec et al. (2016b)[Strong Restricted Eigenvalue Condition]).

1.3.3 Review of literature

In the sparsity scenario, the ℓ_1-penalized least squares method has been well studied. Some oracle inequalities are achieved when strong assumptions on the data hold. For example, if the restricted eigenvalue (Bickel et al., 2009; Raskutti et al., 2010b) condition holds, oracle inequalities can be achieved in the sparse and nearly-sparse scenarios. The uniform uncertainty isometry principle (Candes and Tao, 2007; Needell and Vershynin, 2009) is sufficient for oracle results as well. These assumptions are quite restrictive and are discussed in Van De Geer and Bühlmann (2009). These assumptions guarantee that the covariates can be distinguished from each other by ensuring the Gram matrix is not too far (in a given sense) from the Identity matrix.

For example, the consequences of correlation have been studied in van de Geer and Lederer (2013) and Hebiri and Lederer (2013). In Dalalyan et al. (2014b), the authors provide insights of the impact of highly correlated and moderatly correlated covariates on the prediction risk. In particular, in Dalalyan et al. (2014b)[Example 2], the authors prove that the prediction from Lasso cannot guarantee a fast rate of convergence for any type of covariates correlation even if the penalty term λ was chosen with oracle information. Therefore, the design of the data

may have a strong impact on the quality of the estimation and the prediction and hampers the theoretical guarantees of the maximum a posteriori estimator.

This requirement makes sense when the purpose of the estimator is to recover the parameter. However, when the prediction risk is the criterion of estimation, these assumptions do not seem reasonable anymore. Thus, in the prediction context, it would be very interesting to study families of estimators that achieve oracle inequalities with less restrictive conditions on the Gram matrix.

Such guarantees exist on other estimators such as the ℓ_0-penalized estimator. As mentioned earlier, the ℓ_0-penalized estimator computation is a NP-hard problem that cannot be approximated accurately enough by convex problem. Therefore, even though the ℓ_0-penalization offers great guarantees, it is not very useful in high-dimensional settings.

In inspiration of the ℓ_0 and ℓ_1-penalized estimator properties, it would be of great interest to study theoretical properties on a family of estimators that can guarantee oracle inequalities in the nearly-sparse context with weak assumptions on the data, while being fast enough to compute.

Obtaining results with weaker conditions on the data is a possible way to start tackling the challenge of obtaining oracle inequalities in the random design settings and in more realistic situations. For these reasons, aggregation of estimates has proven to be of theoretical and empirical interest. In particular, we will focus on exponentially weighted aggregations. They have been well studied in the literature as they benefit from great properties in the PAC-Bayesian settings. As such, they are a good starting point to investigate the properties of oracle inequalities in the context of aggregation methods.

1.4 The PAC-Bayesian settings and aggregation estimators

If the Bayesian Information Criterion (BIC) estimator, introduced in Schwarz (1978), performs well from a theoretical point of view (Bunea et al., 2007b), it is very challenging to compute this estimator in high-dimensional settings and it shows poor results when the dimension p is larger than the number of observations n (Giraud, 2014). On the contrary, as discussed earlier, the Lasso estimator benefits from fast computation algorithms but requires strong and restrictive conditions on the design to offer fast rate guarantees. This is especially restrictive when the data are drawn randomly.

The goal of the PAC-learning framework is to study new estimators with new types of correctness theorems that are convenient for randomly drawn data. The PAC-Bayesian approach has

been inspired from the works of Vapnik and Chervonenkis (Vapnik, 1998; Vapnik and Chervonenkis, 1974) and the term of PAC learnability has been introduced by Valiant (1984).
In the Bayesian settings, correctness theorems apply under the assumption that the data are generated from a given prior distribution as considered in McAllester (1998). PAC-Bayesian learning differs from this approach in the sense that the goal is to obtain correctness results when the data are generated from an unknown i.i.d. density. This difference often involves weaker guarantees in the PAC-Bayesian settings than in the Bayesian context. However, the less restrictive assumptions may blend the frequentist and Bayesian approach as the theoretical guarantees do not rely on a prior assumption even though the bounds often refer to aggregate or averaging estimator (which relies on a posterior). Recently, Germain et al. (2016) focused on the similarity of PAC-Bayesian and Bayesian bounds.

1.4.1 The concept of PAC-learning

PAC stands for Probably Approximately Correct and comes from two main ideas.

Probably correct A PAC-bound is not a deterministic guarantee. It allows a small probability that the estimator does not behave well. Hence, the term *probably correct*.

Approximately correct Not only a PAC-bound allows a small erratic behaviour of the estimator. It is also tolerant to a non exact performance of the estimator. In other words, the estimator is given a margin of error and can be *approximately correct*.

These concepts are very important and make some results possible even when an exact and deterministic recovery of the estimator is not possible. In a practical context, it may be particularly useful provided that the margin of error and the probability of success are known and chosen.
The PAC paradigm is very closely related to the paradigm of the VC dimension introduced in Vapnik and Chervonenkis (1974).
We use one theorem from the monograph Catoni (2007) to illustrate this concept[10]. Let consider n couple $(\boldsymbol{x}_i, y_i)$ i.i.d. randomly drawn from an unknown distribution $\mathbb{P}$. Let consider a prediction function $f_{\boldsymbol{\beta}}$ where $\boldsymbol{\beta} \in \Lambda$. The goal is to minimize the expected loss criterion $R(\boldsymbol{\beta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\{l(y, f_{\boldsymbol{\beta}}(\boldsymbol{x}))\}$ for any bounded loss l, $|l(.,.)| < C$. As the probability $\mathbb{P}$ is unknown, one may not directly minimize the expected loss criterion. Therefore, we consider a

[10] I recommend the brief though very insightful introduction to the PAC-learning paradigm made by Pierre Alquier at the *Institut des hautes études scientifiques* on January 2016. The presentation support can be found on https://indico.math.cnrs.fr/event/921/session/8/contribution/36/material/slides/0.pdf.

proxy of the loss r_n that uses the empirical data set $(\boldsymbol{x}_i, y_i)_{i\in[n]}$,

$$r_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i\in[n]} l(y_i, f_{\boldsymbol{\beta}}(\boldsymbol{x}_i)).$$

Theorem 1.4.1 (Catoni (2007)). *Let consider a family of posteriors* Π *and let* π *be a prior on the parameter space* Λ. *For any* $\tau > 0$ *and for a given posterior estimation* $\widehat{\pi}_\tau$, *the following inequality,*

$$\int_{\boldsymbol{u}\in\Lambda} R(\boldsymbol{u})\widehat{\pi}_\tau(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \leq \inf_{\mu\in\Pi}\left[\int_{\boldsymbol{u}\in\Lambda} \mathrm{R}(\boldsymbol{u})\mu(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + \frac{\tau \mathrm{B}}{\mathrm{n}} + \frac{2}{\tau}\Big\{\mathcal{K}(\mu,\pi) + \log\big(2/\epsilon\big)\Big\}\right], \qquad (1.4.1)$$

holds with probability greater than $1-\epsilon$, *where* $\mathcal{K}$ *is the Kullback Leibler divergence.*

The form of Equation 1.4.1 emphasizes the concept behind the term *probably approximately correct*.

In order to manage a learning task in the PAC paradigm the notion of PAC-learnability has been introduced by Valiant (1984).

Definition 1.4.1 (PAC-learnability (Valiant (1984)[11]). *Let* $\mathcal{F}$ *be a class of signal and let* $\ell(.,.)$ *be a loss function on* $\mathcal{F}^2$. *The class* $\mathcal{F}$ *is said to be PAC-learnable if there exists an estimation procedure* $\widehat{\cdot}$ *and an integer* n *such that for any* $f \in \mathcal{F}$, *for any distribution* $\mathbb{P}$ *of observations* $(\boldsymbol{x}, y)$, *for any error margin* $\epsilon > 0$ *and for any probability* $1-\delta$, *the estimator* $\widehat{f}$ *computed from the empirical data* $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ *will be such that*

$$\ell(\widehat{f}, f) < \epsilon,$$

with probability greater than $1-\delta$.

Moreover, let assume the number of observations n *needed to obtain the PAC-bound is bounded by a polynomial function of* $1/\epsilon$, $1/\delta$ *and of the dimension* p *of* $\mathcal{F}$. *Then, if the complexity of the estimation procedure* $\widehat{\cdot}$ *is also bounded by a polynomial time of the latter parameters, then,* $\mathcal{F}$ *is said to be efficiently PAC-learnable.*

Definition 1.4.1 explains the interest of the PAC-learning paradigm in the statistical learning community. A problem is said to be PAC-learnable if a minimax guarantee (for any element of $\mathcal{F}$ for any distribution $\mathbb{P}$) can be proven. Moreover, PAC-learning tolerates a given error and a given probability of failure that gives the opportunity to solve more difficult problem than an exact estimation framework.

[11] For the sake of clarity, we restrain Definition 1.4.1 to the regression context.

Moreover, the concept of efficient PAC learnability is of interest for practical interest, in particular in the high-dimensional settings where a polynomial time of complexity is required (if not faster).
It is worth noting that, with some additional assumptions, a problem is PAC-learnable if and only if the VC dimension of a problem is upperbounded.

1.4.2 Literature in the PAC-Bayesian community

If Theorem 1.4.1 is an important result in the understanding of the PAC-Bayesian theory, other results have proved to be important as well. The statistical learning community has shown a strong interest into the PAC-Bayesian theory.
Two seminal studies of the PAC-Bayesian theory have been carried out in Shawe-Taylor and Williamson (1997) and McAllester (1998). They have set the PAC-Bayesian paradigm and its first results.

Remark 1.4.1. *In Shawe-Taylor and Williamson (1997), the paper Shawe-Taylor et al. (1998) is referred as a work on PAC-bounds results without the Bayesian approach. Even though the publication of Shawe-Taylor et al. (1998) follows the one of Shawe-Taylor and Williamson (1997), the submission was anterior.*

Later on, some studies have set the formalization of the PAC-Bayesian approach in the classification settings (Catoni, 2003, 2004, 2007; Audibert, 2004b) and for the regression learning task (Audibert, 2004c,a). This settings has been extended in the transductive and inductive settings in Alquier (2006) and generalized in (Alquier, 2008; Audibert and Catoni, 2010, 2011). The general message from the literature is that averaging (or aggregating) estimators with a well chosen weight instead of penalized regression seems to perform well in the PAC-Bayesian frame.
Even though these studies bring new insights on the theoretical behaviours of family of estimators, no solution was brought to address the high-dimensional context with (or without) the sparsity assumption. The results in (Dalalyan and Tsybakov, 2008; Alquier and Lounici, 2011; Dalalyan and Tsybakov, 2012b; Rigollet and Tsybakov, 2012b; Guedj and Alquier, 2013; Ridgway et al., 2014)[12]. In Dalalyan and Tsybakov (2008) the aggregation with exponential weighting is used in order to obtain PAC-Bayes sharp oracle inequalities in the high-dimensional sparse settings. The main results of Dalalyan and Tsybakov (2008) are Dalalyan and Tsybakov

[12]For a French speaking audience, one may want to refer to (Guedj and Robbiano, 2014; Guedj et al.) for additional literature on the PAC-Bayesian subject.

(2008)[Theorems 1 and 2]. The following theorem presents the rationale of these results in a simpler but less general context.

Theorem 1.4.2 (Dalalyan and Tsybakov (2008)(Theorem 1)). *Let consider a family of posteriors* Π *and let consider the regression task where for any* $i \in [n]$,

$$y_i = f(\boldsymbol{x}_i) + \xi_i,$$

we assume[13] $\xi_i \sim \mathcal{N}(0, \sigma^2)$ *with* $\sigma < \infty$. *Let assume that the observations* $(\boldsymbol{x}_i)_{i\in[n]}$ *are deterministic. We consider the regression task within a family* $\mathcal{F}$ *of estimators* $f_{\boldsymbol{\beta}}$, *for any* $\boldsymbol{\beta} \in \Lambda \subset \mathbb{R}^p$. *Let* R *be the risk loss* $R(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\xi}}\{l(f(\boldsymbol{x}), f_{\boldsymbol{\beta}}(\boldsymbol{x}))\}$ *for any loss* l. *As the ground truth signal* f *is unknown, one may not directly minimize the expected loss criterion. Therefore, we consider a proxy of the loss* r_n *that uses the empirical data set* $(\boldsymbol{x}_i, y_i)_{i\in[n]}$,

$$r_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i\in[n]} l(y_i, f_{\boldsymbol{\beta}}(\boldsymbol{x}_i)).$$

Let define for any temperature parameter $1/\tau \leq \frac{n}{4\sigma^2}$ *and for any prior* $\pi(\boldsymbol{\beta})$ *in the parameter space* Λ, *the pseudo-posterior*

$$\widehat{\pi}_\tau(\mathrm{d}\boldsymbol{\beta}) \propto \exp\Big\{-\frac{1}{\tau}\mathrm{r_n}(\boldsymbol{\beta})\Big\}\pi(\mathrm{d}\boldsymbol{\beta}).$$

Then the averaged parameter

$$\widehat{\boldsymbol{\beta}}_\tau = \int_{\boldsymbol{u}\in\Lambda} \boldsymbol{u}\widehat{\pi}_\tau(\mathrm{d}\boldsymbol{u}),$$

is such that the estimator $f_{\widehat{\boldsymbol{\beta}}_\tau}$ *ensures the following bound*

$$\mathbb{E}(R(\widehat{\boldsymbol{\beta}}_\tau)) \leq \inf_{\mu\in\Pi}\Big\{\int_{\boldsymbol{u}\in\Lambda} \boldsymbol{u}\widehat{\mu}(\mathrm{d}\boldsymbol{u}) + \tau\mathcal{K}(\mu,\pi)\Big\},$$

where $\mathcal{K}$ *is the Kullback-Leibler divergence.*

The results in Dalalyan and Tsybakov (2008) are extensions of the work of Leung and Barron (2006), where properties were only proven in the setting of finite set of parameters.
PAC-Bayesian and aggregation methods have been applied to the online learning tasks (Audibert, 2009; Cesa-Bianchi et al., 2004) or for variational bayes approximation (Alquier et al., 2016).

[13]The results in Dalalyan and Tsybakov (2008) work with a less restrictive condition named *Assumption A* in the article.

Finally, the PAC-Bayesian approach has emphasized the theoretical interest of aggregation of estimators in order to achieve optimal minimax bounds. The next section presents the aggregation concept.

1.4.3 Aggregation

The name of weighted aggregate estimators appeared early in the literature. In the book Cesa-Bianchi and Lugosi (2006), the weighted aggregate estimation is introduced in the discrete settings with a finite number of experts advice but can easily be generalized to any measurable space. However, as pointed out in Yang (2001b), the author of Stone (1974) introduced the aggregation concept through the notion of stacking several estimators in 1974. The theoretical study and understanding of aggregated estimators improved in the 1990s. One of the initial goals was to provide a decision in a settings where several experts give their advice (or prediction) and one has to give a final decision based on previous performance. The question is then to determine a procedure to predict. One could for example follow the best expert in the sense of a given cumulative loss or one could average uniformly over every expert. In a binary classification problem, the *weighted majority algorithm* introduced in Littlestone and Warmuth (1994) is a generalization of the halving algorithm. Every expert is assigned with a given weight and the choice is made according to the weighted majority. If this expert is wrong at predicting an outcome, the weight of this expert is deprecated by an arbitrary parameter.

Let consider the regression model[14],

$$y_i = f(\boldsymbol{x}_i) + \xi_i,$$

where ξ_i is a random variable and $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are elements of a set $\mathcal{X}$. $\mathcal{F}$ of estimators $f_{\boldsymbol{\beta}}$, for any $\boldsymbol{\beta} \in \Lambda \subset \mathbb{R}^p$. Let consider a set Λ and a mapping $\boldsymbol{\beta} \to f_{\boldsymbol{\beta}}$ that associates to every element $\boldsymbol{\beta} \in \Lambda$ an application $f_{\boldsymbol{\beta}} : \mathcal{X} \to \mathbb{R}$. In the context where for every $\boldsymbol{\beta} \in \Lambda$, $f_{\boldsymbol{\beta}}$ is an estimator of the outcome in the regression model, we would have access to a various number of prediction proposition for any observation $\boldsymbol{x}_i$. In the literature, $f_{\boldsymbol{\beta}}$ is called a weak learner. A question of interest is to understand how one can build a single estimator $\widehat{f}$ according to the knowledge one has on the weak estimators. This single estimator is referred as an aggregate of the weak learners. There exist different types of aggregates and three categories have been well studied in the literature.

[14]Here, the classification model could be considered as well with no additional difficulty.

Linear Aggregation (L) A linear aggregate can be any weighted combination of the family of learners $(f_{\boldsymbol{\beta}})_{\boldsymbol{\beta}\in\Lambda}$. If Λ is a finite set, then the aggregate $\widehat{f}$ is such that for every $i \in [n]$ and any $\boldsymbol{\beta} \in \Lambda$, there exist some $\omega^i_{\boldsymbol{\beta}} \in \mathbb{R}$ such that

$$\widehat{f}(\boldsymbol{x}_i) = \sum_{\boldsymbol{\beta}\in\Lambda} \omega^i_{\boldsymbol{\beta}} f_{\boldsymbol{\beta}}(\boldsymbol{x}_i).$$

In the linear aggregation settings, the performance of the aggregate is compared to the best possible linear combination.

Convex Aggregation (C) In the convex aggregation settings, $\widehat{f}$ is built from non-negative weights which sum to one. In other words, the aggregate is computed under the constraint of being a convex combination of the learners $f_{\boldsymbol{\beta}}$. Again, if Λ is a finite set, then the aggregate $\widehat{f}$ is such that for every $i \in [n]$ and any $\boldsymbol{\beta} \in \Lambda$, there exist some $\omega^i_{\boldsymbol{\beta}} \in \mathbb{R}$ such that

$$\widehat{f}(\boldsymbol{x}_i) = \sum_{\boldsymbol{\beta}\in\Lambda} \omega^i_{\boldsymbol{\beta}} f_{\boldsymbol{\beta}}(\boldsymbol{x}_i),$$

where the vector $\boldsymbol{\omega^i} \in \Omega$ with

$$\Omega = \left\{ \omega \ : \ \forall \boldsymbol{\beta} \in \Lambda, \omega_{\boldsymbol{\beta}} \geq 0, \ \sum_{\boldsymbol{\beta}\in\Lambda} \omega_{\boldsymbol{\beta}} = 1 \right\}.$$

Similarly to the linear settings, the performance of a convex aggregate is often compared with the best convex combination of learners.

Model Selection Aggregation (MS) The goal of model selection aggregates is to pick a weak learner at every iteration $i \in [n]$ that performs nearly as well as the best weak learner of the set Λ. In this context, the aggregate $\widehat{f}$ is such that for every $i \in [n]$ there exists $\boldsymbol{\beta} \in \Lambda$, such that,

$$\widehat{f}(\boldsymbol{x}_i) = f_{\boldsymbol{\beta}}(\boldsymbol{x}_i).$$

There is a transitive inclusion relationship between linear, convex and model selection aggregations. Indeed, any model selection aggregate is a particular convex aggregate and any convex aggregate is a linear aggregate. Of course, other types of aggregation have been considered in the literature such as the s-sparse or the ℓ_q aggregations (Tsybakov, 2014).

It is common usage to compare the model selection aggregate performance with the one of the best learner. In other word, the goal of aggregation algorithms is to find an estimator $\widehat{f}$ that

guarantees that there exists a small quantity $\Delta_{\mathcal{Z}}$ such that,

$$R(\widehat{f}, f) \leq \inf_{\tilde{f} \in \mathcal{Z}} \left\{ R(\tilde{f}, f) + \Delta_{\mathcal{Z}} \right\},$$

where $\mathcal{Z}$ is the set of aggregation $\mathcal{Z} = \{L, C, MS\}$.

The smallest values of $\Delta_{\mathcal{Z}}$ are given in Tsybakov (2003) and are called optimal rate of aggregation. Let K be the size of the learning set and n the number of observations. Under additional conditions (c.f. Tsybakov (2003) for more details) the optimal rate of aggregation in the model selection aggregation is

$$\Psi_{MS} = \frac{\log(K)}{n}.$$

In the convex setting, the optimal rate is

$$\Psi_C = \begin{cases} K/n, & \text{if } K \leq \sqrt{n} \\ \left(\frac{1}{n} \log \left\{ \frac{K}{\sqrt{n}} + 1 \right\} \right)^{1/2}, & \text{otherwise.} \end{cases}$$

Finally the optimal rate of linear aggregation is

$$\Psi_L = \frac{K}{n}.$$

The context of sequential data with expert prediction has been studied in Cesa-Bianchi et al. (1997) where the binary classification task is studied in worst case scenario, with no assumption on the data. A further study of this settings is developped in the thorough work Cesa-Bianchi and Lugosi (2006). The results of Kivinen and Warmuth (1999) proves that a simplified version of the aggregating algorithm introduced in the early work Vovk (1990) can guarantee a regret loss of the order $c \log(n)$, where n is the number of sequences and c is a parameter depending on the considered loss function. With a different approach, the authors of (Littlestone, 1990; Littlestone and Warmuth, 1994) consider the case where one weak learner is guaranteed to do at most m prediction errors in the n sequences. Let K be the number of experts in the set Λ, then the proposed algorithm called *weighted majority algorithm* guarantees a total number of misclassifications smaller than $c(\log(K) + m)$ with a constant c.

Later on, the regression task has been studied as well and technical challenges were required to be solved to do so. Indeed the misclassification loss is bounded which is rarely the case in the regression task. Moreover, the condition that the set Λ is finite is no longer reasonable in this context.

Not all aggregation methods are alike from a guarantee standpoint. For example, it has been proven that the least squares aggregate is not optimal within the family of model selection aggregate (Tsybakov (2014)). The definition of optimality may differ according to the type of aggregation. For example, in the model selection family of aggregation with K weak-learners and n observations, an aggregate is said to be optimal if the excess risk cannot exceed a quantity upperbounded by $C\log(K)/n$, where $C > 0$. The next section mentions some of the performing aggregation methods. Moreover, we discuss guarantees in the random design setting. Indeed, an important takeaway from the PAC-learning theory and the study of aggregation methods is the proof of oracle inequalities (bounds in high probability) and in expectation in the random design setting. To the best of our knowledge, previous theories and estimators from the maximum a posteriori approach did not benefit from easy and convenient bounds in the random design. For example, the Lasso requires restrictive conditions on the design (Bickel et al., 2009; Koltchinskii et al., 2011a; Bellec et al., 2016b).

In that regard, the understanding of the performance of estimators in the context of random design, whether they rely on penalization or averaging methods, would be of great interest. Chapter 4 aims at providing some results on the prediction quality of some maximum a posteriori estimators in the context of transductive and partially labeled prediction in the random design settings. On the other hand, an appealing characteristics of aggregation methods is that the literature has provided some guarantees in the random design setting with fairly mild conditions. This pushes our motivation to study some theoretical properties of aggregation estimators. In the following, we mention aggregation methods that perform well, either in the fixed or random design settings. In particular, we present the exponentially weighted aggregate which is at the center of this thesis.

1.4.4 Exponentially weighted aggregation and its variations

In the following, we present the exponentially weighted aggregate in the continuous settings for linear regression. This family of aggregation estimators is widely used in the literature (Dalalyan and Tsybakov, 2008, 2012b; Rigollet and Tsybakov, 2012b; Chernousova et al., 2013; Dai et al., 2014; Dalalyan and Salmon, 2012; Golubev and Ostrovski, 2014).

With the notations we used earlier in this chapter, the exponentially weighted agregate estimate is of the form

$$\widehat{\boldsymbol{\beta}}_{EWA} = \int_{\Lambda \subset \mathbb{R}^p} \boldsymbol{u}\pi_\tau(\boldsymbol{u})\mathrm{d}\boldsymbol{u}, \tag{1.4.2}$$

with $\tau > 0$ a parameter named the temperature and where π is the normalized posterior,

$$\pi_\tau(\boldsymbol{\beta}) = \frac{\exp(-\frac{L(\boldsymbol{\beta})+\mathcal{P}(\boldsymbol{\beta})}{\tau})}{\int_E \exp(-\frac{L(\boldsymbol{u})+\mathcal{P}(\boldsymbol{u})}{\tau})\mathrm{d}\boldsymbol{u}}. \tag{1.4.3}$$

The only difference between the Bayesian estimator we introduced in Equations 1.3.3 and 1.3.4 and the exponentially weighted agregate is the parameter τ in Equations 1.4.2 and 1.4.3. This parameter is called the temperature with a reference to methods derived from the physics literature.

A takeaway from the literature is that exponentially weighted aggregates perform well (Dalalyan and Tsybakov, 2008; Cesa-Bianchi and Lugosi, 2006; Littlestone and Warmuth, 1994; Alquier and Lounici, 2011; Guedj and Alquier, 2013). One of the first optimal results has been proven in Catoni (1999) in the context of progressive mixture methods which rely on exponential weights. In Dalalyan and Tsybakov (2008), results are proven in the continuous case. Exponentially weighted aggregates offer optimal guarantees in the (nearly-)sparse setting (Dalalyan and Tsybakov, 2012b; Rigollet and Tsybakov, 2012b; Tsybakov, 2014).

In the fixed design settings, Theorem 1.4.2 is an example of guarantees that can be achieved.

Some variations of this aggregation methods have been developed and studied on the theoretical standpoint.

The exponential screening estimator has been introduced in Rigollet and Tsybakov (2011b). In the fixed design setting, the authors prove that the exponential screening estimator benefits from optimal performance universally, which means optimal results in different settings including model selection, convex and linear aggregation (see Tsybakov (2014) for further explanation on the concept of universal aggregate). Exponential screening is very closely related to the exponentially weighted aggregation. It consists in artificially creating two sets of random variables $(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)})$ from the output $\boldsymbol{y}$ and additional noise. One of the sets will be used to build a first maximum a posteriori estimator $\widehat{\boldsymbol{\beta}}_{(1)}$ of the parameter on the loss of interest. The prediction $\widehat{f}^{(1)}$ is then used to determine the empirical loss

$$r_n = \ell\Big(\widehat{f}^{(1)}\big(\boldsymbol{y}^{(1)}\big), \boldsymbol{y}^{(2)}\Big)$$

that is used to define the weights of the exponentially weighted aggregation.

In the random design setting, mirror averaging (MA) is the analogue of the exponentially weighted aggregation. What differs from the original exponentially weighted aggregation pro-

cedure is that the estimated posterior density that determines the weights of the aggregation is recursively computed and averaged. Using our notations, let define $\pi_{0,\tau} = 1$ and then recursively, for any $i \in [n]$,

$$\pi_{i,\tau} = \frac{\exp(-\frac{L_i(\boldsymbol{\beta})-\mathcal{P}_i(\boldsymbol{\beta})}{\tau})}{\int_E \exp(-\frac{L(\boldsymbol{u})-\mathcal{P}(\boldsymbol{u})}{\tau})\mathrm{d}\boldsymbol{u}},$$

where L_i and $\mathcal{P}_i$ are respectively the log-likelihood and the log-prior from the i first observations of the data set. Then the posterior that is finally used in the aggregation procedure is the average

$$\pi_\tau(\boldsymbol{\beta}) = \frac{1}{n+1}\sum_{i=0}^{n} \pi_{i,\tau}(\boldsymbol{\beta}).$$

This procedure has been inspired by the mirror descent algorithm in the field of optimization (Nemirovskii et al., 1983; Ben-Tal and Nemirovski, 1999). To the best of our knowledge, the mirror averaging procedure has been first introduced in Juditsky et al. (2005). In Juditsky et al. (2008), the mirror averaging procedure is proven to be optimal in expectation with respect to the model selection setting. In Dalalyan and Tsybakov (2012a), regression, density estimation and classification problems are studied within the sparse setting. The authors offer general results in the random design setting. The authors provide a PAC-bound in expectation in Dalalyan and Tsybakov (2012a)[Theorem 1] and sharp oracle inequalities in Dalalyan and Tsybakov (2012a)[Theorem 2]. Moreover, Dalalyan and Tsybakov (2012a)[Proposition 3] proposes bounds in the sparsity setting that are nearly optimal (up to a logarithmic factor) in the model selection, convex and linear settings[15].

Remark 1.4.2. *The method called Q-aggregation studied in Dai et al. (2012) in the fixed design setting and Lecué and Rigollet (2014) in the random design setting shows interesting theoretical properties as well. In particular, under conditions given in Lecué and Rigollet (2014), the Q-aggregation is proven to be an optimal aggregate in the random design setting in Lecué and Rigollet (2014). Results are given in probability and in expectation. However, we will not discuss them further as we focus on the exponentially weighted aggregate in the context of this thesis.*

The exponentially weighted aggregation and its variations (exponential screening and mirror averaging) benefit from (nearly)-optimal results either in fixed design or random design. Moreover, in the (nearly)-sparse setting, the literature has proved optimal results that legitimate the exponentially weighted aggregate as an interesting alternative to the Lasso. Indeed, obtaining optimal bound for the exponentially weighted aggregate needs less restrictive conditions on the

[15] A study on the limit of the mirror averaging procedure can be found in Lecué and Mendelson (2013).

Gram matrix than the Lasso.

Even though the literature provides optimal rates of convergence for these methods in the non-asymptotic high-dimensional frame, the computational complexity may represent the bottleneck of these methods. This is particularly important in the high-dimensional settings where averaging from Monte Carlo processes might be very costly. The next section briefly reviews this question.

1.4.5 Computational challenges and Langevin Monte Carlo

Aggregation methods require to generate a sample of random vectors with respect to a given (pseudo-)posterior distribution. This might be a difficult task, especially in the continuous high-dimensional settings.

For practical puproses, it is very useful to know the number of iterations K needed to achieve a targeted accuracy ϵ. Otherwise, how could we choose the number of iterations to achieve a desired accuracy?

Moreover, the required number of iterations should not grow too quickly with the dimension p of the problem, neither should it grow when the accuracy ϵ gets small.

Hence, a good sampling method would benefit from properties that guarantee the sufficiency of an explicit number of iterations K that remains *reasonably small*[16] for large value of p or for a small error tolerance ϵ.

There is no need to mention that not every sampling process benefits from such properties. The need of efficient sampling methods is an important constraint on the choice of the aggregation method, and more generally, of any statistical learning method requiring sample generation. As a consequence, among all aggregation methods, some may be more suitable to practical purposes.

The subject of this thesis is focus on the exponentially weighted aggregate estimator. This method is of great interest when the distribution of the pseudo-posterior is sampled from a discretized Langevin Monte Carlo process. Indeed, when the posterior can be written as in Equation 1.4.3 such that $L+\lambda\mathcal{P}$ is strongly convex, then any targeted accuracy can be achieved with an explicit number of iterations K (Dalalyan, 2016). There are numerous analogies between the computational questions in the *maximum a posterior* penalized estimation where the convex penalties guarantee fast computation from gradient descent and in the aggregation averaging

[16]Here, the terms *reasonably small*, *large dimension* and *small error tolerance* are not formally defined. It depends on the context and on how much resources one individual is willing to allocate to generate the sample as well. And, clearly this notion will dramatically change over time with improvements in computational technologies and statistical algorithms.

where convex penalties can guarantee the Langevin Monte Carlo algorithm to perform well. Insightful discussion can be found in Dalalyan (2016) and Dalalyan (2017) about the analogy and relationship between sampling and optimization properties.

Let consider the function[17] f, then in Dalalyan (2016), the Langevin diffusion $\boldsymbol{L}_t$ is defined for any $t \geq 0$, by

$$\mathrm{d}\boldsymbol{\vartheta}_t = -\nabla f(\boldsymbol{\vartheta}_t)\mathrm{d}t + \sqrt{2}p\mathbf{b}_t,$$

where the function $\mathbf{b}$ is a vector of Brownian motion of dimension p. The Langevin Monte Carlo relies on a discretization of the Langevin diffusion. In our work, we will consider the Euler discretization where the $k+1$-th iteration is recursively defined by

$$\boldsymbol{\upsilon}_{k+1}^h = \boldsymbol{\upsilon}_k^h - h\nabla f(\boldsymbol{\upsilon}_k^h) + \sqrt{2h}\boldsymbol{\xi}_k,$$

for a given step $h > 0$ and where $\boldsymbol{\xi}_k$ is a p-dimensional standard Gaussian vector. Other discretizations, such as the Ozaki discretization, are considered in the literature (Dalalyan, 2016).

The quality (or accuracy) of the sampling method is relative to an arbitrarily chosen distance between the distribution from which the sampling is generated and the targeted distribution. In the literature, common metrics such as the Kullback-Leibler (KL), the Chi-Square (χ^2) divergences (c.f. Definition 1.4.2), or the total variation norm, have been widely used.

Definition 1.4.2 (Kullback-Leibler and χ^2 divergences)**.** *Let ν and μ be two probability measures over a set Ω, then if ν is absolutely continuous with respect to μ, the Kullback-Leibler divergence is defined by*

$$KL(\nu\|\mu) = \int_\Omega \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\nu.$$

The χ^2 divergence is defined by

$$\chi^2(\nu\|\mu) = \int_\Omega \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu} - 1\right)^2 \mathrm{d}\mu.$$

Some results on the convergence of the Langevin Monte Carlo have been proven in these different metrics in Dalalyan (2016). More recently, the Wasserstein distance of order 2 has been more commonly used.

Definition 1.4.3 (Wasserstein distance)**.** *The Wasserstein distance of order $l \in \mathbb{N}^\star$ between*

[17] The condition of the function f will be discussed in Chapter 3.

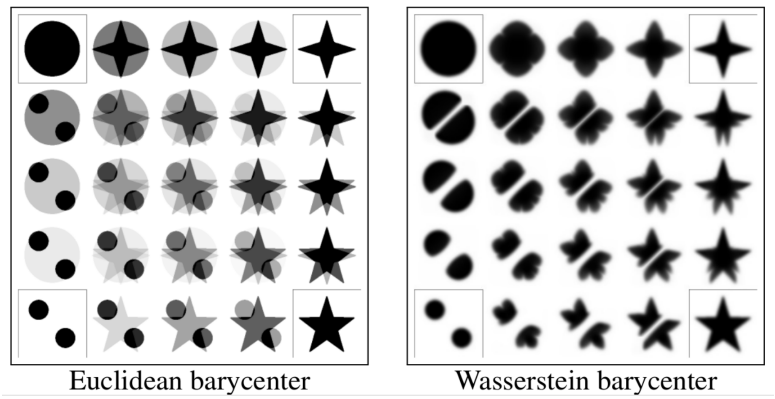


Figure 1-7: Shape interpolation in 2-dimensions using the Euclidean barycenter (left) and the Wasserstein barycenter (right). The Wasserstein barycenter shape interpolation procedure is discussed in Solomon et al. (2015). Credit: Solomon et al. (2015)

two measures of probability ν *and* η, $W_l(\nu, \eta)$ *is defined by*

$$W_l(\nu,\eta) = \inf_{\psi \in \Psi(\nu,\eta)} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{u} - \boldsymbol{v}\|_l^l \mathrm{d}\psi(\boldsymbol{u},\boldsymbol{v}) \right\}^{1/l}$$

where $\Psi(\nu, \eta)$ *is the set of probability measures on* $\mathbb{R}^p \times \mathbb{R}^p$ *with marginals* ν *and* η.

In (Durmus and Moulines, 2016; Dalalyan, 2017) the question of proving the accuracy of a sampling Langevin Monte Carlo with various discretization schemes are proven in the sense of the Wasserstein distance when the log-density is strongly convex. Using the Wasserstein distance is a very promising approach. In Dalalyan (2017), it is proven that if f is m-strongly convex and has a continuous M-Lipschitz gradient, then the distribution of the samples generated by the Langevin Monte Carlo method with a Euler discretization is such that the accuracy ϵ will be achieved (in the sense of the Wasserstein distance of order 2) for a number of iterations K proportional to $p\epsilon^{-2}\log(p/\epsilon)$.

Remark 1.4.3. *As a personal note, I find it very tempting to believe that the Wasserstein approach will offer great technical tools to generalize statistical and simulation theories. However, computing the Wasserstein distance is a very challenging task. Even methods to approximate the Wasserstein distance is an ongoing subject. In Solomon et al. (2015) an efficient algorithm is introduced to do so. However, no general minimization methods of the Wasserstein distance has been found yet. As such, the Wasserstein is very interesting on the statistical theory point of view but will require further understanding to improve sampling results.*

New results on the Langevin Monte Carlo methods brought the missing part to make the

exponentially weighted aggregate estimators (or mirror averaging in the random design settings) an interesting estimator from both a computational and a statistical standpoint, with fairly mild conditions.

For the various reasons mentioned in this section, the exponentially weighted aggregate is of great interest and motivates our work in this thesis.

1.5 Roadmap

Now that we have established the importance of statistical theory and more specifically the notions we need from high-dimensional statistical learning, we will present what will be studied in Chapters 2, 3 and 4. As mentioned earlier, the exponentially weighted aggregate is an efficient estimator to recover signal in a nearly-sparse context (Dalalyan and Tsybakov, 2012a,b). Different priors have been proposed to do so. It is amazingly surprising to note that among all the priors we can read in the litterature, the Laplace prior was never successfully used. Indeed, aggregation with Laplace prior is to the aggregation what the Lasso is to the maximum a posteriori paradigm. Indeed, the Lasso is arguably one of the most studied estimators in the context of penalized regression methods for nearly-sparse high-dimensional problem. Lasso has proven to be very efficient to recover sparse signal (even though restrictive conditions are necessary) and the statistical community has proven strong theoretical properties.

The exponentially weighted aggregate with Laplace prior can be defined by the procedure of the standard EWA estimator described in Equation 1.4.2 with a pseudo-posterior density defined by

$$\widehat{\pi}_n(\boldsymbol{\beta}) \propto \exp(-V_n(\boldsymbol{\beta})/\tau), \qquad \text{where} \quad V_n(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \left\|\boldsymbol{\beta}\right\|_1. \tag{1.5.1}$$

There is a chasm between interest and results shown for the exponentially weighted aggregate with Laplace prior and the huge literature on the Lasso. Our main motivation is to cross that chasm by developing results on the exponentially weighted aggregate with Laplace prior in terms of theoretical guarantees and computational efficiency. Obviously, our aim is strongly supported by already existing results on general theory about the exponentially weighted aggregation. We already mentioned earlier in this chapter some of the most remarkable results. However, the use of existing oracle inequalities for the exponentially weighted aggregate (Dalalyan and Tsybakov (2012b)[Theorem 1]) is not very optimistic when directly applied to the Laplace prior. They are far from being sharp in comparison to what can be achieved with the maximum a

posteriori analogue, the Lasso. Moreover, a result in Castillo et al. (2015)[Theorem 7] proves that the Bayesian Lasso cannot recover sparse signal well enough. However, an inspirational observation can push the need to study this estimator. If we look briefly at the exponentially weighted aggregate with Laplace prior, we remark that

$$\lim_{\tau \to 0} \widehat{\boldsymbol{\beta}}_{\tau}^{EWA} = \widehat{\boldsymbol{\beta}}^{Lasso}.$$

Therefore, it is legitimate to question the quality of the exponentially weighted aggregation with Laplace prior to recover sparse signal and to compare it with the Lasso performance. This was the seminal question of this thesis. Chapter 2 will give some insights to this question. We provide oracle inequalities in the regression settings under compatibility factor assumption for the prediction loss. These guarantees are very similar to the one of the Lasso additioned with a price to pay $p\tau$. We provide explicit value of the temperature parameter τ for which sharp oracle inequalities are guaranteed. From this point of view, we develop a generalization of the Lasso results to a broader family of estimators and we match similar performance than the one in Dalalyan et al. (2017) for the Lasso. Relying on the study in Bobkov and Madiman (2011), we also prove pseudo-posterior concentration results. The derived results can be easily generalized to other convex penalty norm such as the total variation as in Harchaoui and Lévy-Leduc (2010). Hence it gives insight on the relationships between maximum a posteriori penalized estimators and their aggregation analogues. To illustrate this, we establish results in the matrix regression settings in relation with the nuclear norm penalty. To do so, we introduce the compatibility factor adapted to the matrix case. Using the exponentially weighted aggregate with nuclear norm prior, we extend the results of the oracle inequality in Koltchinskii et al. (2011a) and we provide pseudo posterior concentration. These results legitimate the exponentially weighted aggregation with sparse-inducing priors as a good alternative to the maximum a posteriori analogues for learning purposes in the nearly-sparse high-dimensional settings. Furthermore, as the literature on Lasso oracle inequalities in the random design settings is quite arid, such results could help to find mirror averaging oracle inequality in the random design. However, even though this family of estimators is performing well on the theoretical standpoint, the approximation of this estimator is more challenging. Indeed, averaging with respect to a distribution is a difficult and resource consuming task. The computation question is addressed in Chapter 3. A direct application of Langevin Monte Carlo from (Dalalyan, 2016, 2017; Durmus and Moulines, 2016) would not necessarily guarantee that any accuracy can be achieved in a finite number of iterations. Indeed, these results require the log-posterior to be strongly

convex and smooth, and yet with the Laplace prior the differentiability does not hold and the strong-convexity cannot be assumed for any data design. In Chapter 3, we partially solve this issue. We study the behaviour of an Euler-discretization Langevin Monte Carlo sampling. We focus on the Wasserstein distance accuracy with respect to the targeted exponentially weighted aggregate with Laplace prior. This work relies on the shoulder of the monograph Ledoux (2005). Our goal is to adjust the algorithm proposed in Dalalyan (2016) in order to circumvent the non differentiability of the pseudo posterior. We provide an explicit number of iterations K that is comparable to the one in Dalalyan (2016) in regards to the error tolerance ϵ and the dimension p. However, this result is not completely satisfying as it relies on conditions on the Gram matrix that are not realistic in the high-dimensional settings. Indeed, in Chapter 3, we assume that the smallest eigenvalue of the Gram matrix is positive. Despite this limitation, this work provides a solution to guarantee a good approximation of the targeted density in a slightly more general context than existing results in the literature. The last chapter can be seen as a related study on the standard Lasso in the context of transductive and semi-supervised learning. Even though, this work treats a different problem, it could complete (and respectively be completed by) the other results of this thesis. In Chapter 4, we show that unlabeled data should be used in the estimator to infer the variance-covariance matrix. We introduce two adaptations of the Lasso estimators that substantially improve the prediction performance in respectively the transductive and the partially labeled settings. Under compatibility factor conditions, this last chapter proves sharp oracle inequalities in the random design settings.

Chapter 2

On the Exponentially Weighted Aggregate with the Laplace Prior

A joint work with Arnak Dalalyan and Quentin Paris.

Contents

Abstract

In this paper, we study the statistical behaviour of the Exponentially Weighted Aggregate (EWA) in the problem of high-dimensional regression with fixed design. Under the assumption that the underlying regression vector is sparse, it is reasonable to use the Laplace distribution as a prior. The resulting estimator and, specifically, a particular instance of it referred to as the Bayesian lasso, was already used in the statistical literature because of its computational convenience, even though no thorough mathematical analysis of its statistical properties was carried out. The present work fills this gap by establishing sharp oracle inequalities for the EWA with the Laplace prior. These inequalities show that if the temperature parameter is small, the EWA with the Laplace prior satisfies the same type of oracle inequality as the lasso estimator does, as long as the quality of estimation is measured by the prediction loss. Extensions of the proposed methodology to the problem of prediction with low-rank matrices are considered.

2.1 Introduction

We investigate statistical properties of the Exponentially Weighted Aggregate (EWA) in the context of high-dimensional linear regression with fixed design and under the sparsity scenario. This corresponds to considering data that consist of n random observations $y_1, \dots, y_n \in \mathbb{R}$ and p fixed covariates $\boldsymbol{x}^1, \dots, \boldsymbol{x}^p \in \mathbb{R}^n$. We further assume that there is a vector $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that the residuals $\xi_i = y_i - \beta_1^\star \boldsymbol{x}_i^1 - \ldots - \beta_p^\star \boldsymbol{x}_i^p$ are independent, zero mean random variables. In vector notation, this reads as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi}, \tag{2.1.1}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ is the response vector, $\mathbf{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{\xi}$ is the noise vector. For simplicity, in all mathematical results, the noise vector is assumed to be distributed according to the Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. We are mainly interested in obtaining mathematical results that cover the high-dimensional setting. This means that our goal is to establish risk bounds that can be small even if the ambient dimension p is large compared to the sample size. In order to attain this goal, we will consider the, by now, usual sparsity scenario. In other words, the established risk bounds are small if the underlying large vector $\boldsymbol{\beta}^*$ is well approximated by a sparse vector. Note that this setting can be extended to the matrix case, sometimes termed trace-regression (Rohde and Tsybakov, 2011; Koltchinskii et al., 2011a). Indeed, if the rows $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of the design matrix $\mathbf{X}$ are replaced by $m_1 \times m_2$ matrices $\mathbf{X}_1, \dots, \mathbf{X}_n$, then the regression vector $\boldsymbol{\beta}^\star$ is replaced by a $m_1 \times m_2$ matrix $\mathbf{B}^\star$ and the model of trace regression is

$$y_i = \mathrm{tr}(\mathbf{X}_i^\top \mathbf{B}^\star) + \xi_i, \qquad i = 1, \dots, n. \tag{2.1.2}$$

Our focus here is on the statistical properties related to the prediction risk. The important questions of variable selection and estimation in various norms are beyond the scope of the present work.

In the aforementioned vector- and trace-regression models, the most thoroughly studied statistical procedures of estimation and prediction rely on the principle of penalised least squares[1]. In the vector-regression model, assuming that the quadratic loss is used, this corresponds to analysing the properties of the estimator

$$\widehat{\boldsymbol{\beta}}^{\mathrm{PLS}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\,\mathrm{Pen}(\boldsymbol{\beta}) \right\}, \tag{2.1.3}$$

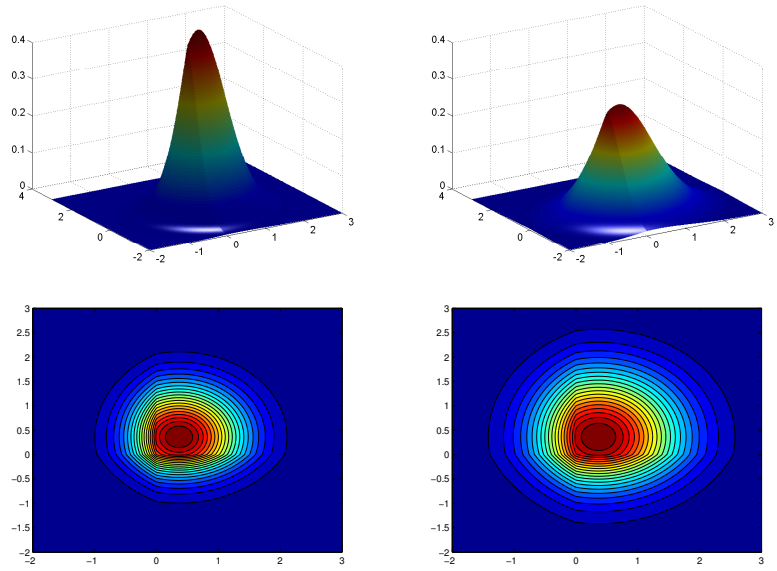
where $\lambda > 0$ is a tuning parameter and $\mathrm{Pen} : \mathbb{R}^p \to \mathbb{R}$ is a sparsity promoting penalty function. The literature on this topic is so rich that it would be impossible to cite here all the relevant papers. We refer the interested reader to the books (Bühlmann and van de Geer, 2011; Koltchinskii, 2011; Giraud, 2015; van de Geer, 2016) and the references therein. Among the sparsity promoting penalties, one can mention the ℓ_0 penalty (which for various choices of λ leads to the BIC (Schwarz, 1978), the AIC (Akaike, 1974) or to Mallows's Cp (Mallows, 1973)), the ℓ_1 penalty or the lasso (Tibshirani, 1996b), the ℓ_q (with $0 < q < 1$) or the bridge penalty (Frank and Friedman, 1993; Fu, 1998), the SCAD (Fan and Li, 2001), the minimax concave penalties (Zhang, 2010), the entropy (Koltchinskii, 2009), the SLOPE (Bogdan et al., 2015; Su and Candes, 2016), etc.

The aggregation by exponential weights is an alternative approach to the problems of estimation and prediction that, roughly speaking, replaces the minimisation by the averaging. Assuming that every vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is a candidate for estimating the true vector $\boldsymbol{\beta}^\star$, aggregation (cf., for instance, the survey (Tsybakov, 2014)) consists in computing a weighted average of the candidates. Naturally, the weights are to be chosen in a data-driven way. In the case of the exponentially weighted aggregate (EWA), the weight $\widehat{\pi}_n(\boldsymbol{\beta})$ of each candidate vector $\boldsymbol{\beta}$ has the exponential form

$$\widehat{\pi}_n(\boldsymbol{\beta}) \propto \exp\big(-V_n(\boldsymbol{\beta})/\tau\big), \qquad \text{where} \quad V_n(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\,\mathrm{Pen}(\boldsymbol{\beta}) \tag{2.1.4}$$

is the potential used above for defining the penalised least squares estimator and $\tau > 0$ is an additional tuning parameter referred to as the temperature. Using this notation, the EWA is

[1] Or, more generally, on the penalised empirical risk minimisation



$\tau = 0.5$ $\tau = 0.8$

Figure 2-1: **Top:** the plots of the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior for the temperature $\tau = 0.5$ (left) and $\tau = 0.8$ (right). One can observe that decreasing the value of τ strengthens the peakedness of the density. **Bottom:** the level curves of the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior for the temperature $\tau = 0.5$ (left) and $\tau = 0.8$ (right). One clearly observes the non-differentiability of the density along the axes β_1 and β_2 (caused by the non-differentiability of the ℓ_1-norm).

defined by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\beta}\, \widehat{\pi}_n(\boldsymbol{\beta})\, \mathrm{d}\boldsymbol{\beta}. \tag{2.1.5}$$

Exponential weights have been used for a long time in statistical learning theory (cf., for instance, Vovk (1990)). Their use in statistics was initiated by Yuhong Yang in (Yang, 2000a,b,c, 2001a) and by Olivier Catoni in a series of preprints, later on included in (Catoni, 2004, 2007). Precise risk bounds for the EWA in the model of regression with fixed design have been established in (Leung and Barron, 2006; Dalalyan and Tsybakov, 2007, 2008, 2012a; Dalalyan and Salmon, 2012; Dai et al., 2012; Golubev and Ostrovski, 2014; Chernousova et al., 2013). In the model of regression with random design, the counterpart of the EWA, often referred to as mirror averaging, has been thoroughly studied in (Juditsky et al., 2005, 2008; Audibert, 2009; Chesneau and Lecué, 2009; Gaïffas and Lecué, 2007; Dalalyan and Tsybakov, 2012a; Lecué and Mendelson, 2013). Note that when the temperature τ equals σ^2/n, the EWA coincides with the Bayesian posterior mean in the regression model with Gaussian noise provided that the prior is defined by $\pi_0(\boldsymbol{\beta}) \propto \exp(-\lambda \,\mathrm{Pen}(\boldsymbol{\beta})/\tau)$. Thanks to this analogy, we will call $\widehat{\pi}_n$ pseudo-posterior density. Let us mention here that, considering the path $\tau \mapsto \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ for $\tau \in (0, \sigma^2/n]$, we get a continuous interpolation between the penalised least squares and the Bayesian posterior mean. Along with these studies, several authors have demonstrated the ability of the EWA to optimally

estimate a sparse signal. To this end, various types of priors have been used. For instance, (Leung and Barron, 2006; Rigollet and Tsybakov, 2011b; Alquier and Lounici, 2011; Arias-Castro and Lounici, 2014) have employed discrete priors over the set of least-squares estimators with varying supports whereas (Dalalyan and Tsybakov, 2008, 2012b) have used Student-type heavy-tailed priors. In the context of structured sparsity, the EWA has been successfully used in (Alquier and Biau, 2013; Guedj and Alquier, 2013; Dalalyan et al., 2014a). Given the close relationship between the EWA and the Bayes estimator, it is worth mentioning here that the problem of sparse estimation has also received much attention in the literature on Bayesian Statistics (Wipf et al., 2003; Park and Casella, 2008; Hans, 2009). Posterior concentration properties for these methods have been investigated in (Castillo and van der Vaart, 2012; Castillo et al., 2015; van der Pas et al., 2016; Gao et al., 2015).

Despite these efforts, some natural questions remain open. One of them, described in details below, is at the origin of this work. Let us consider the prediction error of a candidate vector $\boldsymbol{\beta}$ with respect to the quadratic loss

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) = \frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{\beta} - \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star)^2. \tag{2.1.6}$$

On the one hand, theoretical studies of the lasso (Candes and Tao, 2007; Bickel et al., 2009; Belloni et al., 2014; Dalalyan et al., 2014b; Bellec et al., 2016b,a), established[2] sharp upper bounds for the prediction risk of the PLS estimator (2.1.3) for the ℓ_1-penalty $\mathrm{Pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. Therefore, one could expect the EWA with the Laplace prior $\pi_0(\boldsymbol{\beta}) \propto \exp(-\lambda\|\boldsymbol{\beta}\|_1/\tau)$ to have a high prediction performance. On the other hand, to the best of our knowledge, there is no result in the literature establishing accurate risk bounds for the EWA with Laplace prior. Indeed, a straightforward application of the PAC-Bayesian type risk bounds (McAllester, 1998) for the EWA (such as, for instance, Theorem 1 in (Dalalyan and Tsybakov, 2012b)) to the Laplace prior leads to strongly sub-optimal remainder terms. This raises the following questions:

Q1. Is the EWA with the Laplace prior suitable for prediction under the sparsity scenario?

Q2. If it is, what is the range of temperature τ providing good prediction accuracy?

Q3. How do the statistical properties of the EWA with the Laplace prior compare with those of the lasso?

Related questions are considered in (Castillo et al., 2015). Indeed, for $\boldsymbol{\beta}^\star = \mathbf{0}_p$, $p = n$ and $\mathbf{X}^\top\mathbf{X}/n = \mathbf{I}_n$, Theorem 7 from (Castillo et al., 2015) establishes the following property. For

[2]Provided that the Gram matrix $\mathbf{X}^\top\mathbf{X}/n$ satisfies suitable assumptions (restricted isometry, restricted eigenvalues, compatibility, etc.).

all the reasonable choices[3] of the tuning parameter λ, if the temperature τ in the EWA with the Laplace prior is chosen as $\tau = \sigma^2/n$, then the resulting posterior puts asymptotically no mass on the ball centred at $\boldsymbol{\beta}^\star$ and of radius $\mathrm{Const}(\log n/n)^{1/2}$, the latter corresponding to the optimal rate of convergence in this model. This negative result, stated in terms of the posterior contraction rate, can be easily adapted in order to show that, under the previous conditions, the Bayesian posterior mean is sub-optimal.

The present paper completes the picture by establishing some positive results. In particular, it turns out that if the temperature parameter of the EWA with the Laplace prior is of the order $s\sigma^2/(pn)$, where s is the sparsity of $\boldsymbol{\beta}^\star$, then the EWA with the Laplace prior does attain the optimal rate of convergence. Furthermore, it satisfies the same type of sharp sparsity inequality as the lasso does. Interestingly, the proof of this result is based on arguments which differ from those used in the aggregation literature. Indeed, the two previously used techniques for getting oracle inequalities for the EWA and related procedures rely either on the PAC-Bayesian inequality or on the Stein unbiased risk estimate. Instead, the key idea of our proof is to take advantage of the following relations:

$$\int_{\mathbb{R}^p} \nabla\big(\beta_j^\alpha e^{-V_n(\boldsymbol{\beta})/\tau}\big)\,\mathrm{d}\boldsymbol{\beta} = 0, \qquad j = 1,\dots,p, \quad \alpha = 0, 1. \tag{2.1.7}$$

Hence, most of our arguments are independent of the noise distribution and can be extended to other settings (as opposed to the results relying on the Stein formula). Elaborating on this, we prove that the pseudo-posterior $\widehat{\pi}_n$ puts an overwhelming weight on the set of vectors $\boldsymbol{\beta}$ satisfying a sharp oracle inequality with rate-optimal remainder term. In the case of the Gaussian noise, we also obtain the explicit form of the Stein unbiased estimator of the risk of $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$, which can be used for choosing the tuning parameter. Finally, we extend these results to the model of trace regression when the underlying true matrix $\mathbf{B}^\star$ has low rank.

The rest of the paper is organised as follows. The notation used throughout the paper is introduced in the next section. 2.3 analyses the prediction loss of the EWA with the Laplace prior, and 2.4 gathers results characterising the concentration of the pseudo-posterior $\widehat{\pi}_n$. Extensions of these results to the case where the unknown parameter is a (nearly) low-rank matrix are considered in 2.5. A brief summary of the obtained results along with some conclusions is given in 2.6. Finally, the proofs are postponed to 2.7.

[3] By "reasonable" we understand here the choice $\lambda = \mathrm{Const}\,\sigma(\frac{\log p}{n})^{1/2}$, for which the lasso is provably rate optimal under the sparsity scenario, provided that the design satisfies a version of the restricted eigenvalue condition.

2.2 Notation

This paragraph collects notation used throughout the paper. For every integer $k \geq 1$, we write $\mathbf{1}_k$ (resp. $\mathbf{0}_k$) for the vector of $\mathbb{R}^k$ having all coordinates equal to one (resp. zero). We set $[k] = \{1, \ldots, k\}$. For every $q \in [0, \infty]$, we denote by $\|\boldsymbol{u}\|_q$ the usual ℓ_q-norm of $\boldsymbol{u} \in \mathbb{R}^k$, that is $\|\boldsymbol{u}\|_q = (\sum_{j\in[k]} |u_j|^q)^{1/q}$ when $0 < q < \infty$, $\|\boldsymbol{u}\|_0 = \mathrm{Card}(\{j : u_j \neq 0\})$ and $\|\boldsymbol{u}\|_\infty = \max_{j\in[k]} |u_j|$. For every integer $k \geq 1$ and any $T \subset [k]$, we denote by T^c and $|T|$ the complementary set $[p] \setminus T$ and the cardinality of T, respectively. For $\boldsymbol{u} \in \mathbb{R}^k$ and $T \subset [k]$, we denote $\boldsymbol{u}_T \in \mathbb{R}^{|T|}$ the vector obtained from $\boldsymbol{u}$ by removing all the coordinates belonging to the set T^c.

In Sections 2.3 and 2.4, we recall that $\mathbf{X} \in \mathbb{R}^{n\times p}$ refers to the deterministic design matrix with columns $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p \in \mathbb{R}^n$ and rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$. Finally, our analysis will involve the compatibility factor of the design matrix defined, for any $J \subset [p]$ and $c > 0$, by

$$\kappa_{J,c} = \inf_{\boldsymbol{u}\in\mathbb{R}^p : \|\boldsymbol{u}_{J^c}\|_1 < c\|\boldsymbol{u}_J\|_1} \frac{c^2|J|\|\mathbf{X}\boldsymbol{u}\|_2^2}{n(c\|\boldsymbol{u}_J\|_1 - \|\boldsymbol{u}_{J^c}\|_1)^2}. \tag{2.2.1}$$

Note that the compatibility factor, often used for the analysis of the lasso, is slightly larger[4] than the restricted eigenvalue (Bickel et al., 2009). For a better understanding of these (and related) quantities we refer the reader to (Bickel et al., 2009, Sections 3 and 4) and (van de Geer and Bühlmann, 2009).

Risk bounds established in the present work for the EWA contain a new term, as compared to the analogous risk bounds for the lasso. This term reflects the peakedness of the pseudo-posterior density $\widehat{\pi}_n$ and is defined by

$$H(\tau) = p\tau - \int G(\boldsymbol{u})\widehat{\pi}_n(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + G(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}), \tag{2.2.2}$$

where $G(\boldsymbol{u}) = {}^1\!/_n\|\mathbf{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1$. When the temperature τ is low, close to zero, the pseudo-posterior $\widehat{\pi}_n$ is close to a Dirac measure centred at the lasso, which implies that $H(\tau)$ is close to zero. Furthermore, since the above function G is convex, we have the following bound

$$H(\tau) \leq p\tau. \tag{2.2.3}$$

In 2.3 and 2.4 we will occasionally use the following matrix notation. For all integers $p \geq 1$, $\mathbf{I}_p$ refers to the identity matrix in $\mathbb{R}^{p\times p}$. For any integers $p \geq 1$ and $q \geq 1$, any matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$ and any subset T of $[q]$, we denote by $\mathbf{A}_T$ the matrix obtained from $\mathbf{A}$ by removing all the

[4] Since this factor appears in the denominator of the risk bound, the larger is the better.

columns belonging to T^c. Finally the transpose and the Moore-Penrose pseudoinverse of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^\top$ and $\mathbf{A}^\dagger$, respectively.

2.3 Risk bound for the EWA with the Laplace prior

This section is devoted to discussing statistical properties of the EWA with the Laplace prior. Recall that it is defined by (2.1.5) as the average with respect to the pseudo-posterior density

$$\widehat{\pi}_n(\boldsymbol{\beta}) \propto \exp(-V_n(\boldsymbol{\beta})/\tau), \qquad \text{where} \quad V_n(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{2.3.1}$$

The emphasis is put on non-asymptotic guarantees in terms of the prediction loss. It is important to mention here that the Laplace prior, $\pi_0(\boldsymbol{\beta}) \propto \exp(-\lambda\|\boldsymbol{\beta}\|_1/\tau)$, makes use of the same scale for all the coordinates of the vector $\boldsymbol{\beta}$. This presumes that the covariates (columns of the matrix $\mathbf{X}$) are already rescaled so that their Euclidean norms are almost equal. An alternative approach (see, for instance, Bunea et al. (2007b); Bickel et al. (2009))—that we will not follow here—would consist in replacing the ℓ_1-norm of $\boldsymbol{\beta}$ by the weighted ℓ_1-norm $\sum_{j\in[p]} \|\boldsymbol{x}^j\||\beta_j|$. The next result provides the main risk bound for the EWA.

Theorem 2.3.1. *Assume that data are generated by model (2.1.1) with $\boldsymbol{\xi}$ drawn from the Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I}_n)$ and that the covariates are rescaled so that $\max_{j\in[p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \le 1$. Suppose, in addition, that $\lambda \ge 2\sigma({}^2\!/_n \log(p/\delta))^{1/2}$, for some $\delta \in (0,1)$. Then, with probability at least $1-\delta$,*

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}, \boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p \\ J\subset[p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2|J|}{4\kappa_{J,3}} \right\} + 2p\tau, \tag{2.3.2}$$

where ℓ_n is defined in (2.1.6) and $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ is defined in (2.1.5) and (2.3.1).

For the lasso estimator, risk bounds of this nature have been developed in (Koltchinskii et al., 2011a; Sun and Zhang, 2012a; Dalalyan et al., 2014b; Bellec et al., 2016a). The risk bound in (2.3.2) extends the risk bounds available for the lasso (cf. Theorem 2 in (Dalalyan et al., 2014b)) to the EWA with the Laplace prior. Indeed, letting the temperature τ go to zero, the last term in the right-hand side of (2.3.2) disappears and we retrieve the risk bound for the lasso. An attractive feature of risk bound (2.3.2) is that the factor in front of the term $\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star)$ is equal to one; this is often referred to as a sharp or exact oracle inequality. Furthermore, the other three terms in the right-hand side of (2.3.2) are neat and have a simple interpretation. The second

term, $4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1$, accounts for the approximate sparsity; when $\mathbf{X}\boldsymbol{\beta}^\star$ is well approximated by $\mathbf{X}\bar{\boldsymbol{\beta}}$ with a s-sparse vector $\bar{\boldsymbol{\beta}}$, then choosing $J = \{j : \bar{\beta}_j \neq 0\}$ annihilates this term. The third term of the risk bound corresponds to the optimal rate, up to a logarithmic factor, of estimation of a vector $\boldsymbol{\beta}^\star$ concentrated on the known set J. Indeed, if $|J| = s$ and the compatibility factor is bounded away from zero, this term is of order $s/n \log(p)$. Finally, the last term in the above risk bound, $2p\tau$, reflects the influence of the temperature parameter τ. In particular, it shows that if $\tau = \sigma^2/(pn)$ then this term is negligible with respect to the other remainder terms.

The inequality stated in 2.3.1 is a simplified version of the following one (proved in 2.7): for any $\gamma > 1$, in the event $\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \leq n\lambda/\gamma$, it holds

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}, \boldsymbol{\beta}^\star) \leq \inf_{\substack{\bar{\boldsymbol{\beta}} \in \mathbb{R}^p \\ J \subset [p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{J,(\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau), \tag{2.3.3}$$

where $H(\tau)$ is defined in (2.2.2). On the one hand, one can use this more general result for getting an oracle inequality under more general assumptions on the noise distribution such as those considered, for instance, in (Bunea et al., 2007b; Belloni et al., 2014). On the other hand, one can infer from (2.3.3) that the term $H(\tau)$ highlights the difference, in terms of statistical complexity, between the lasso and the EWA with the Laplace prior. It is therefore important to get a precise evaluation of $H(\tau)$ as a function of τ, p and n, and to understand how tight the inequality $H(\tau) \leq p\tau$ is. To answer this question, we restrict our attention to orthonormal designs and show the tightness of the aforementioned inequality. To this end, let us introduce the scaled complementary error function $\Psi_v(t) = e^{t^2/2v} \frac{1}{\sqrt{2\pi v}} \int_t^\infty e^{-u^2/2v} \, \mathrm{d}u$.

Proposition 2.3.1. *Let $\widehat{\boldsymbol{\Sigma}}_n = {}^1\!/_n \mathbf{X}^\top \mathbf{X}$ be the Gram matrix and $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}} = {}^1\!/_n \widehat{\boldsymbol{\Sigma}}_n^\dagger \mathbf{X}^\top \boldsymbol{y}$ be the least-squares estimator. Then, we have*

$$H(\tau) = \|\widehat{\boldsymbol{\Sigma}}_n^{1/2} \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\|_2^2 + \lambda \|\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\|_1 - (\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}})^\top \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}. \tag{2.3.4}$$

Furthermore, when the design is orthonormal, that is $\widehat{\boldsymbol{\Sigma}}_n = \mathbf{I}_p$, then the EWA with the Laplace prior is a thresholding estimator, $\widehat{\beta}_j^{\mathrm{EWA}} = \mathrm{sign}(\widehat{\beta}_j^{\mathrm{LS}})\big(|\widehat{\beta}_j^{\mathrm{LS}}| - \lambda w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|)\big)$, where

$$w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|) = \frac{\Psi_\tau(\lambda - |\widehat{\beta}_j^{\mathrm{LS}}|) - \Psi_\tau(\lambda + |\widehat{\beta}_j^{\mathrm{LS}}|)}{\Psi_\tau(\lambda - |\widehat{\beta}_j^{\mathrm{LS}}|) + \Psi_\tau(\lambda + |\widehat{\beta}_j^{\mathrm{LS}}|)}, \tag{2.3.5}$$

and

$$H(\tau) = \sum_{j=1}^p \lambda\big(|\widehat{\beta}_j^{\mathrm{LS}}| - \lambda w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|)\big)\big(1 - w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|)\big). \tag{2.3.6}$$

The last expression of $H(\tau)$ provided by the proposition may be used for a numerical evaluation. First, let us note that if we set $\bar{\beta}_j = \widehat{\beta}_j^{\mathrm{LS}}/\sqrt{\tau}$ and $\bar{\lambda} = \lambda/\sqrt{\tau}$, the function $H(\tau)/\tau$ is independent of τ. Indeed, we have $H(\tau)/\tau = \sum_j h(\bar{\lambda}, |\bar{\beta}_j|)$ where

$$h(\bar{\lambda}, z) = \bar{\lambda}\big(z - \bar{\lambda} w(1, \bar{\lambda}, z)\big)\big(1 - w(1, \bar{\lambda}, z)\big), \qquad \forall z > 0.$$

In Fig. 2-2 below, we plot the curves of the functions $z \mapsto h(\bar{\lambda}, z)$ for different values of the parameter $\bar{\lambda}$.

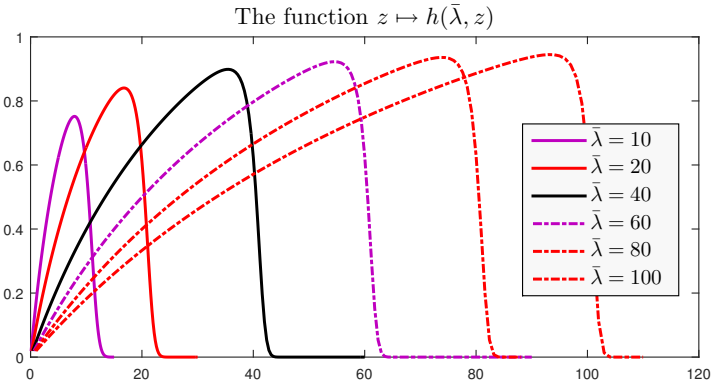


Figure 2-2: For different values $\bar{\lambda} \in \{10, 20, 40, 60, 80, 100\}$, we plot the function $z \mapsto h(\bar{\lambda}, z)$.

These curves clearly show that the bound $H(\tau) \leq p\tau$, a consequence of $h(\bar{\lambda}, z) \leq 1$, is tight. Another interesting observation is that the function $H(\tau)$ is always nonnegative. This basically implies that the value of τ minimising the right-hand side of (2.3.3) is $\tau = 0$. In other terms, the lowest risk bound is obtained for the lasso. This legitimately raises the following question: is there any advantage of using the EWA with the Laplace prior as compared to the lasso? Our firm conviction is that there is an advantage, and will try to explain our viewpoint in the rest of this section.

The point is that the lasso estimator is a nonsmooth function of the data. One of the consequences of this is that the Stein unbiased risk estimate (SURE) for the lasso is a discontinuous function of data. Indeed, as proved in (Tibshirani and Taylor, 2012), The SURE for the lasso (see also the earlier work (Donoho and Johnstone, 1995; Zou et al., 2007)) is given by

$$\widehat{R}^{\mathrm{lasso}}(\lambda) = \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}(\lambda)\|_2^2 - \sigma^2 + \frac{2\sigma^2}{n}\mathrm{rank}(\mathbf{X}_{\mathcal{A}(\lambda)}),$$

where $\mathcal{A}(\lambda) = \{j \in [p] : \widehat{\beta}_j^{\mathrm{lasso}}(\lambda) \neq 0\}$ is the active set for the lasso estimator with the tuning parameter λ. In theory, this quantity $\widehat{R}^{\mathrm{lasso}}(\lambda)$ can be used for choosing the tuning parameter λ of the lasso. However, in practice, this solution is rarely employed, since $\mathcal{A}(\lambda)$ has a very

unstable behaviour as a function of λ and $\boldsymbol{y}$. As a consequence, not only one can get very different "optimal" values of λ for two very close vectors $\boldsymbol{y}$ and $\boldsymbol{y}'$, but is also likely to obtain very different "optimal" values of λ for the same vector $\boldsymbol{y}$ if using two different optimisation algorithms for computing an approximate solution to the lasso problem.

Using Stein's lemma, in the case where $\boldsymbol{\xi}$ is drawn from the Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ distribution, one checks that

$$\widehat{R}^{\mathrm{EWA}}(\lambda, \tau) = \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}_{\lambda,\tau}\|_2^2 - \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2 \tau} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}_{\lambda,\tau})\|_2^2 \, \widehat{\pi}_{n,\lambda,\tau}(\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\beta} \tag{2.3.7}$$

is an unbiased estimator of the risk $\mathbb{E}[\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}, \boldsymbol{\beta}^\star)]$. Furthermore, the function $(\lambda, \tau) \mapsto \widehat{R}^{\mathrm{EWA}}(\lambda, \tau)$ is clearly continuous on $(0, \infty) \times (0, \infty)$. One can also check that the unbiased risk estimate $\widehat{R}^{\mathrm{EWA}}(\lambda, \tau)$ depends continuously on the data vector $\boldsymbol{y}$. Therefore, this quantity us arguably more robust to the variation in data and more regular as a function of the tuning parameters as compared to $\widehat{R}^{\mathrm{lasso}}$. This implies that minimising $\widehat{R}^{\mathrm{EWA}}(\lambda, \tau)$ with respect to λ or τ might be a good strategy for choosing these parameters adaptively.

Of course, this requires to be able to numerically compute the right-hand side of (2.3.7) or, equivalently, the mean and the covariance matrix of the pseudo-posterior distribution $\widehat{\pi}_n$. For smooth and strongly log-concave densities, the cost of such computations has been recently assessed in (Dalalyan, 2016; Durmus and Moulines, 2016). The adaptation of the approaches developed therein to the pseudo-posterior $\widehat{\pi}_n$, which is neither smooth nor strongly log-concave (but can be approximated by such a function), is an ongoing work.

2.4 Pseudo-Posterior concentration

Since the EWA estimator has a Bayesian flavour, it is appealing to look at the concentration properties of the pseudo-posterior distribution $\widehat{\pi}_n$. This is particularly important in the light of the results in Castillo et al. (2015) establishing that, for the temperature $\tau = \sigma^2/n$, the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior puts asymptotically no mass on the set of vectors $\boldsymbol{\beta}$ having a small prediction error. Furthermore, this result is proven for the orthonormal design matrix $\mathbf{X}$, which, intuitively, is a rather favourable situation for the Laplace prior.

The first property that we establish here and that characterises the concentration of the pseudo-posterior around its average is the following upper bound on the variance of the prediction $\mathbf{X}\boldsymbol{\beta}$ when $\boldsymbol{\beta}$ is drawn from $\widehat{\pi}_n$. (Recall that the matrix $\mathbf{X}$ has n rows, so the normalisation by multiplicative factor $1/n$ is natural.)

Proposition 2.4.1. *If $\widehat{\pi}_n(\boldsymbol{u}) \propto \exp\left(-V_n(\boldsymbol{u})/\tau\right)$ is the pseudo-posterior with the Laplace prior defined by (2.3.1), then, for every $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$, we have*

$$\int_{\mathbb{R}^p} V_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \leq p\tau + V_n(\bar{\boldsymbol{\beta}}) - \frac{1}{2n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}. \tag{2.4.1}$$

Furthermore, choosing $\bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{u}\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$, we get

$$\frac{1}{n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}})\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \leq p\tau. \tag{2.4.2}$$

The proof of this result is rather simple and plays an important role in the proof of the oracle inequality stated in 2.3.1. For these reasons, we opted for presenting this proof in this section, instead of postponing it to 2.7.

Proof. The convexity of the function $\bar{\boldsymbol{\beta}} \mapsto \|\bar{\boldsymbol{\beta}}\|_1$ readily implies that the function $\bar{\boldsymbol{\beta}} \mapsto W_n(\bar{\boldsymbol{\beta}}) = V_n(\bar{\boldsymbol{\beta}}) - {}^1\!/_{2n}\|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2$ is a convex function, for every fixed $\boldsymbol{u} \in \mathbb{R}^p$. Furthermore, we have $W_n(\boldsymbol{u}) = V_n(\boldsymbol{u})$ and $\nabla W_n(\boldsymbol{u}) = \nabla V_n(\boldsymbol{u})$ at any point $\boldsymbol{u}$ of differentiability of V_n. Therefore,

$$V_n\big(\bar{\boldsymbol{\beta}}\big) \geq V_n(\boldsymbol{u}) + \big(\bar{\boldsymbol{\beta}} - \boldsymbol{u}\big)^\top \nabla V_n(\boldsymbol{u}) + \frac{1}{2n}\big\|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\big\|_2^2, \tag{2.4.3}$$

for all $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$ and for almost all $\boldsymbol{u} \in \mathbb{R}^p$ (those for which V_n is continuously differentiable at $\boldsymbol{u}$). Using the fundamental theorem of calculus, we remark that

$$\int_{\mathbb{R}^p} \nabla V_n(\boldsymbol{u})\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = -\tau \int_{\mathbb{R}^p} [\nabla \widehat{\pi}_n(\boldsymbol{u})]\, \mathrm{d}\boldsymbol{u} = \mathbf{0}_p \tag{2.4.4}$$

and that

$$\int_{\mathbb{R}^p} \boldsymbol{u}^\top \nabla V_n(\boldsymbol{u})\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} - p\tau = \int_{\mathbb{R}^p} \sum_{j=1}^{p} \left(\beta_j \frac{\partial V_n}{\partial \beta_j}(\boldsymbol{u}) - \tau\right) \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \tag{2.4.5}$$

$$= -\tau \int_{\mathbb{R}^p} \sum_{j=1}^{p} \frac{\partial[u_j \widehat{\pi}_n(\boldsymbol{u})]}{\partial u_j}\, \mathrm{d}\boldsymbol{u} = 0. \tag{2.4.6}$$

Integrating inequality (2.4.3) on $\mathbb{R}^p$ with respect to the density $\widehat{\pi}_n$ and using relations (2.4.4) and (2.4.6), we arrive at

$$V_n\big(\bar{\boldsymbol{\beta}}\big) \geq \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} - p\tau + \frac{1}{2n} \int_{\mathbb{R}^p} \big\|\mathbf{X}\big(\boldsymbol{u} - \bar{\boldsymbol{\beta}}\big)\big\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}. \tag{2.4.7}$$

This completes the proof of the first claim of the proposition.

To prove the second claim, we replace $\bar{\boldsymbol{\beta}}$ by $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ in (2.4.7). After rearranging the terms, this yields

$$\frac{1}{2n}\int_{\mathbb{R}^p} \big\|\mathbf{X}\big(\boldsymbol{u}-\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big)\big\|_2^2\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \le p\tau + V_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) - \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}. \tag{2.4.8}$$

Using once again the fact that $\boldsymbol{u}\mapsto W_n(\boldsymbol{u}) = V_n(\boldsymbol{u}) - {}^1\!/_{2n}\|\mathbf{X}(\boldsymbol{u}-\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}})\|_2^2$ is a convex function, we obtain $V_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) = W_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) \le \int W_n(\boldsymbol{u})\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$, which is equivalent to

$$V_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) - \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \le -\frac{1}{2n}\int_{\mathbb{R}^p} \big\|\mathbf{X}\big(\boldsymbol{u}-\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big)\big\|_2^2\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}. \tag{2.4.9}$$

This inequality, combined with (2.4.8), completes the proof of (2.4.2) and of the proposition. □

Remark 2.4.1. *A careful inspection of the proof reveals that the claims of the proposition are independent of the precise form of the ℓ_1-penalty. Therefore, the proposition still holds if we replace the ℓ_1-norm by any convex penalty.*

The second claim of the proposition establishes that the dispersion of the distribution $\widehat{\pi}_n$ around its average value $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ is of the order $(p\tau)^{1/2}$. Interestingly, we show below that the same order of magnitude appears when we determine a region of concentration for the pseudo-posterior $\widehat{\pi}_n$. A key argument in the proof of the latter claim is the following result.

Proposition 2.4.2 (Bobkov and Madiman (2011), Theorem 1.1)**.** *Let $\widehat{\pi}_n(\boldsymbol{u}) \propto \exp(-V_n(\boldsymbol{u})/\tau)$ be a log-concave probability density*[5] *and let $\boldsymbol{\beta}$ be a random vector drawn from $\widehat{\pi}_n$. Then, for any $t > 0$, the inequality*

$$V_n(\boldsymbol{\beta}) \le \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \tau\sqrt{p}\,t \tag{2.4.10}$$

holds with probability at least $1 - 2e^{-t/16}$.

Using this proposition, we establish the following result (the proof of which is postponed to 2.7) characterising the concentration of $\widehat{\pi}_n$.

Theorem 2.4.1 (Posterior concentration bound)**.** *Assume that data are generated by model (2.1.1) with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ and rescaled covariates, i.e., $\max_{j\in[p]} {}^1\!/_n\|\boldsymbol{x}^j\|_2^2 \le 1$. Let the quality of an estimator be measured by the squared prediction loss (2.1.6). Assume that the tuning parameter λ satisfies $\lambda \ge 2\sigma\big({}^2\!/_n \log({}^p\!/_\delta)\big)^{1/2}$, for some $\delta \in (0,1)$. Then, with probability at least*

[5] This means that V_n is a convex function.

$1-\delta$, *the pseudo-posterior* $\widehat{\pi}_n$ *with the Laplace prior defined by* (2.3.1) *satisfies*

$$\widehat{\pi}_n\left(\boldsymbol{\beta} : \ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p \\ J\subset[p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2|J|}{2\kappa_{J,3}} \right\} + 8p\tau \right) \ge 1 - 2e^{-\sqrt{p}/16}. \quad (2.4.11)$$

The latter theorem, in conjunction with 2.3.1, tells us that if we generate a random vector $\boldsymbol{\beta}$ distributed according to the density $\widehat{\pi}_n$, then with high probability it will have a prediction loss almost as small as the one of the EWA, the average with respect to $\widehat{\pi}_n$. This remark might be attractive from the computational point of view, since, at least for some distributions, drawing a random sample is easier than computing the expectation. Note also that by increasing the factor in front of the term $p\tau$ it is possible to make the $\widehat{\pi}_n$-probability of the event considered in 2.4.1 even closer to one.

2.5 Sparsity oracle inequality in the matrix case

In this section, we extend the results of the previous sections to the problem of matrix regression with a low-rankness inducing prior. We first need to introduce additional notations used throughout this section.

2.5.1 Specific notation

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, the scalar product is defined by

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B}). \quad (2.5.1)$$

The nuclear norm of a $p \times q$ matrix $\mathbf{A}$ is $\|\mathbf{A}\|_1 = \sum_{k=1}^r s_{\mathbf{A},k}$, where $s_{\mathbf{A},k}$ is the k-th largest singular value of $\mathbf{A}$ and $r = \mathrm{rank}(\mathbf{A})$. The operator norm is $\|\mathbf{A}\| = \sup_{\boldsymbol{x}\in\mathbb{R}^q} \|\mathbf{A}\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2 = s_{\mathbf{A},1}$. We denote by $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{n\times m_1\times m_2}$ the three-dimensional tensor playing the role of the design matrix. Besides, let $\|\mathbf{A}\|^2_{L_2(\mathcal{X})} = \langle \mathbf{A}, \mathbf{A}\rangle_{L_2(\mathcal{X})}$ be the prediction loss defined via the "scalar product" $\langle \mathbf{A}, \mathbf{C}\rangle_{L_2(\mathcal{X})} = \frac{1}{n}\sum_{i=1}^n (\langle \mathbf{X}_i, \mathbf{A}\rangle\langle \mathbf{X}_i, \mathbf{C}\rangle)$. We will use the notation $\boldsymbol{u}^\top\mathcal{X} = \sum_{i\in[n]} u_i\mathbf{X}_i \in \mathcal{M}_{m_1,m_2}$ for the product of the tensor $\mathcal{X}$ with the vector $\boldsymbol{u} \in \mathbb{R}^n$.

We now need to define the matrix compatibility factor. Its definition is more involved than in the vector case because of the fact that the left and right singular spaces differ from one matrix to another. Let $\bar{\mathbf{B}}$ be any $m_1 \times m_2$ matrix of rank $r = \mathrm{rank}(\bar{\mathbf{B}})$ having the singular value decomposition $\bar{\mathbf{B}} = \mathbf{V}_1\boldsymbol{\Sigma}\mathbf{V}_2^\top$. Here, $\boldsymbol{\Sigma}$ is a $r \times r$ diagonal matrix with positive diagonal entries, $\boldsymbol{\Sigma}_{11} \ge \ldots \ge \boldsymbol{\Sigma}_{rr} > 0$, and $\mathbf{V}_j$ is a $m_j \times r$ matrix with orthonormal columns for $j = 1, 2$. For

any $J \subset [r]$ and $j = 1, 2$, we define $\mathbf{V}_{j,J}$ as the $m_j \times |J|$ matrix obtained from $\mathbf{V}_j$ by removing the columns with indices lying outside of J. This allows us to introduce the linear operators $\mathcal{P}_{\bar{\mathbf{B}},J^c}$ and $\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}$ from $\mathcal{M}_{m_1,m_2}$ to $\mathcal{M}_{m_1,m_2}$

$$\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = (\mathbf{I}_{m_1} - \mathbf{V}_{1,J}\mathbf{V}_{1,J}^\top)\mathbf{U}(\mathbf{I}_{m_2} - \mathbf{V}_{2,J}\mathbf{V}_{2,J}^\top), \quad \mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = \mathbf{U} - \mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U}).$$

We define, for every $\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$, $J \subset [\mathrm{rank}(\bar{\mathbf{B}})]$ and $c > 0$, the compatibility factor

$$\kappa_{\bar{\mathbf{B}},J,c} = \inf_{\substack{\mathbf{U}\in\mathcal{M}_{m_1,m_2} \\ \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1 < c\|\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1}} \frac{c^2|J|\,\|\mathbf{U}\|^2_{L_2(\mathcal{X})}}{\big(c\|\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1 - \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1\big)^2}. \tag{2.5.2}$$

When $J = [\mathrm{rank}(\bar{\mathbf{B}})]$, we use the notation $\kappa_{\bar{\mathbf{B}},c}$ instead of $\kappa_{\bar{\mathbf{B}},J,c}$. Note that the set $\mathcal{C}(\bar{\mathbf{B}}, J, c) = \{\mathbf{U} \in \mathcal{M}_{m_1,m_2} : \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1 < c\|\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1\}$ defines the cone of dimensionality reduction. It consists of matrices $\mathbf{U}$ that can be written as a sum of two matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ such that $\mathbf{U}_1$ is of small rank and dominates the possibly full-rank matrix $\mathbf{U}_2$, in the sense that $\|\mathbf{U}_2\|_1 \le c\|\mathbf{U}_1\|_1$. Indeed, it suffices to set $\mathbf{U}_1 = \mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U})$ and to remark that $\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = \mathbf{V}_{1,J}\mathbf{V}_{1,J}^\top\mathbf{U} + (\mathbf{I}_{m_1} - \mathbf{V}_{1,J}\mathbf{V}_{1,J}^\top)\mathbf{U}\mathbf{V}_{2,J}\mathbf{V}_{2,J}^\top$ is of rank not exceeding $2|J|$.
Similarly to (2.2.2), we also define the function

$$H(\tau) = m_1 m_2 \tau - \int_{\mathcal{M}_{m_1,m_2}} G(\mathbf{U})\,\widehat{\pi}_n(\mathbf{U})\,\mathrm{d}\mathbf{U} + G\big(\widehat{\mathbf{B}}\big), \tag{2.5.3}$$

where $G(\mathbf{U}) = \|\mathbf{U}\|^2_{L_2(\mathcal{X})} + \lambda\|\mathbf{U}\|_1$. The convexity property of the function G entails that $H(\tau) \le m_1 m_2 \tau$ for every $\tau > 0$.

2.5.2 Nuclear-norm prior and the exponential weights

The observed outcomes are n real random variables $y_1, \dots, y_n \in \mathbb{R}$. Contrary to Sections 2.3 and 2.4 where the design points are $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$, this section studies the situation in which we consider n design matrices $\mathbf{X}_i \in \mathbb{R}^{m_1\times m_2}$ for $i \in [n]$. We further assume that there is a regression matrix $\mathbf{B}^\star \in \mathcal{M}_{m_1,m_2}$ such that

$$y_i = \mathrm{tr}(\mathbf{X}_i^\top \mathbf{B}^\star) + \xi_i, \qquad i \in [n], \tag{2.5.4}$$

where the residuals ξ_i are independent and identically distributed according to a centred Gaussian distribution with variance σ^2. This model is referred to as trace-regression; see, for instance, Rohde and Tsybakov (2011). In this model, the nuclear norm is akin to the ℓ_1 norm in the

vector case. Therefore, to some extent, the equivalent of the lasso estimator $\widehat{\mathbf{B}}_\lambda^{\mathrm{NNP-LS}}$ with a positive smoothing parameter λ, is defined by

$$\widehat{\mathbf{B}}_\lambda^{\mathrm{NNP-LS}} \in \arg\min_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \left\{ \frac{1}{2n}\sum_{i\in[n]} (\boldsymbol{y}_i - \langle \mathbf{X}_i, \mathbf{B}\rangle)^2 + \lambda\|\mathbf{B}\|_1 \right\}. \tag{2.5.5}$$

This is the nuclear-norm penalized least-squares estimator. Similarly to the vector case, the above defined estimator $\widehat{\mathbf{B}}_\lambda^{\mathrm{NNP-LS}}$ is the maximum a posteriori estimator corresponding to the nuclear-norm prior

$$\pi_0(\mathbf{B}) \propto \exp\left\{ -\frac{\lambda\sigma^2\|\mathbf{B}\|_1}{n} \right\}. \tag{2.5.6}$$

This section investigates the prediction performance of the procedure obtained by replacing the optimisation step by averaging. In the matrix case, we define the potential function V_n and the pseudo-posterior, respectively, by

$$V_n(\mathbf{B}) = \frac{1}{2n}\sum_{i\in[n]} (\boldsymbol{y}_i - \langle \mathbf{X}_i, \mathbf{B}\rangle)^2 + \lambda\|\mathbf{B}\|_1, \qquad \text{and} \qquad \widehat{\pi}_n(\mathbf{B}) \propto \exp\left\{-1/\tau V_n(\mathbf{B})\right\}. \tag{2.5.7}$$

Using these concepts, we define the EWA with the nuclear-norm prior by

$$\widehat{\mathbf{B}}^{\mathrm{EWA}} = \int_{\mathcal{M}_{m_1,m_2}} \mathbf{B}\, \widehat{\pi}_n(\mathbf{B})\, \mathrm{d}\mathbf{B}. \tag{2.5.8}$$

We aim at studying the performance of this estimator in terms of the prediction loss

$$\ell_n\big(\widehat{\mathbf{B}}, \mathbf{B}^\star\big) = \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_{L^2(\mathcal{X})}^2 = \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \widehat{\mathbf{B}} - \mathbf{B}^\star\rangle^2. \tag{2.5.9}$$

2.5.3 Oracle Inequality

The problem of assessing the quality of the nuclear-norm penalised estimators has received a great deal of attention; see, for instance, (Srebro and Shraibman, 2005; Candès and Tao, 2010; Candes and Plan, 2011; Bunea et al., 2011; Gaiffas and Lecué, 2011; Negahban and Wainwright, 2011, 2012; Klopp, 2014). Such an interest in these methods is mainly motivated by the variety of applications in computer vision and image analysis (Shen and Wu, 2012; Harchaoui et al., 2012), recommendation systems (Zhou et al., 2008; Lim and Teh, 2007), and many other areas. Bayesian approaches to the problem of low-rank matrix estimation and prediction has been recently analysed by Alquier and Biau (2013); Mai and Alquier (2015); Cottet and Alquier (2016).

Making the parallel with the sparse vector estimation and prediction problem, we can note that the counterpart of the vector sparsity $s = \|\boldsymbol{\beta}^\star\|_0$ in the matrix case is the product $(m_1 + m_2)\operatorname{rank}(\mathbf{B}^\star)$, representing the number of potentially nonzero terms in the singular values decomposition of $\mathbf{B}^\star$. Similarly, the counterpart of the ambient dimension p is the overall number of entries in $\mathbf{B}^\star$ that is $m_1 m_2$. In view of these analogies, the next theorem is a natural extension of 2.3.1 to the model of trace-regression. To state it, we need the following notation:

$$v_{\mathcal{X}} = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right\|^{1/2} \bigvee \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{X}_i \right\|^{1/2}. \tag{2.5.10}$$

Theorem 2.5.1. *Assume that data are generated by model (2.5.4) with $\boldsymbol{\xi}$ drawn from the Gaussian distribution $\mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. Suppose, in addition, that $\lambda \geq 2\sigma v_{\mathcal{X}} \{2/n \log((m_1 + m_2)/\delta)\}^{1/2}$, for some $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, the matrix $\widehat{\mathbf{B}}^{\mathrm{EWA}}$ defined in (2.5.8) satisfies*

$$\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^\star) \leq \inf_{\substack{\bar{\mathbf{B}} \in \mathcal{M}_{m_1, m_2} \\ J \subset [\operatorname{rank}(\bar{\mathbf{B}})]}} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}}, J^c}(\bar{\mathbf{B}})\|_1 + \frac{9\lambda^2 |J|}{4\kappa_{\bar{\mathbf{B}}, J, 3}} \right\} + 2m_1 m_2 \tau. \tag{2.5.11}$$

This result can be seen as an extension of (Koltchinskii et al., 2011a, Theorem 2) to the exponentially weighted aggregate with a prior proportional to the scaled nuclear norm. Indeed, if we upper bound the infimum over all matrices $\mathbf{B}$ by the infimum over matrices such that $\operatorname{rank}(\mathbf{B}) \leq r$ for some given integer r, we easily see that (2.5.11) yields

$$\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^\star) \leq \inf_{\substack{\bar{\mathbf{B}} \in \mathcal{M}_{m_1, m_2} \\ \operatorname{rank}(\bar{\mathbf{B}}) \leq r}} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + \frac{9\lambda^2 r}{4\kappa_{\bar{\mathbf{B}}, 3}} \right\} + 2m_1 m_2 \tau. \tag{2.5.12}$$

An advantage of inequality (2.5.11) is that it offers a continuous interpolation between the so called "slow" and "fast" rates. "Slow" rates refer typically to risk bounds that are proportional to λ, whereas "fast" rates are proportional to λ^2. For procedures based on ℓ_1-norm or nuclear-norm penalty, "slow" rates are known to hold without any assumption on the design, while "fast" rates require a kind of compatibility assumption. In (2.5.11), taking $J = \varnothing$, the term with λ^2 disappears and we get the "slow" rate proportional to $\lambda \|\bar{\mathbf{B}}\|_1$. The other extreme case corresponding to $J = [\operatorname{rank}(\bar{\mathbf{B}})]$ leads to the "fast" rate proportional to $\lambda^2 \operatorname{rank}(\bar{\mathbf{B}})$, provided that the compatibility factor is bounded away from zero. The risk bound in (2.5.11) bridges these two extreme situations by providing the rate $\min_{q \in [r]} \{\lambda(s_{q+1, \bar{\mathbf{B}}} + \ldots + s_{r, \bar{\mathbf{B}}}) + \lambda^2 q\}$, where $r = \operatorname{rank}(\bar{\mathbf{B}})$ and $s_{\ell, \bar{\mathbf{B}}}$ is the ℓ-th largest singular value of $\bar{\mathbf{B}}$. Thus, our risk bound quantifies

the quality of prediction in the situations where the true matrix (or the best prediction matrix) is nearly low-rank, but not necessarily exactly low-rank.

As well as in the vector case, the inequality stated in 2.5.1 is a simplified version of the following one: for any $\gamma > 1$, in the event $\|\boldsymbol{\xi}^\top \mathcal{X}\| \leq n\lambda/\gamma$, it holds

$$\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^\star) \leq \inf_{\substack{\mathbf{B}\in\mathcal{M}_{m_1,m_2} \\ \mathcal{P}\in\bar{\mathscr{P}}}} \left\{ \ell_n(\mathbf{B}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}})\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{\bar{\mathbf{B}},J,(\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau), \quad (2.5.13)$$

where H is defined by (2.5.3). This inequality as well as 2.5.1 is proved in 2.7.

2.5.4 Pseudo-posterior concentration

In what follows, we state the result on the pseudo-posterior concentration in the matrix case. Akin to the vector case, one of the main building blocks is (Bobkov and Madiman, 2011, Theorem 1.1), see 2.4.2 above. Since the potential V_n in (2.5.7) is convex, the proposition applies and implies that, for every $t > 0$,

$$\widehat{\pi}_n\Big(\mathbf{B} : \ V_n(\mathbf{B}) \leq \int_{\mathcal{M}_{m_1,m_2}} V_n(\mathbf{U})\, \widehat{\pi}_n(\mathbf{U})\, \mathrm{d}\mathbf{U} + \tau\sqrt{m_1 m_2 t}\Big) \geq 1 - 2e^{-t/16}. \quad (2.5.14)$$

After some nontrivial algebra, this allows us to show that a risk bound similar to (2.5.1) holds not only for the pseudo-posterior-mean $\widehat{\mathbf{B}}^{\mathrm{EWA}}$, but also for any matrix $\mathbf{B}$ randomly sampled from $\widehat{\pi}_n$.

Theorem 2.5.2. *Let data be generated by model (2.5.4) with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$ and let the quality of an estimator be measured by the squared prediction loss (2.5.9). Assume that the tuning parameter λ satisfies $\lambda \geq 2\sigma v_{\mathcal{X}} \{2/n \log((m_1+m_2)/\delta)\}^{1/2}$, for some $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the pseudo-posterior $\widehat{\pi}_n$ with the nuclear-norm prior defined by (2.5.7) is such that the probability*

$$\widehat{\pi}_n\left(\mathbf{B} : \ell_n(\mathbf{B}, \mathbf{B}^\star) \leq \inf_{\substack{\bar{\mathbf{B}}\in\mathcal{M}_{m_1,m_2} \\ J\subset[\mathrm{rank}(\bar{\mathbf{B}})]}} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}})\|_1 + \frac{9\lambda^2|J|}{4\kappa_{\bar{\mathbf{B}},J,3}} \right\} + 8m_1 m_2 \tau \right) \quad (2.5.15)$$

is larger than $1 - 2e^{-\sqrt{m_1 m_2}/16}$.

We postpone the proof of Theorem 2.5.2 to Section 2.7. One can deduce from 2.5.2 that if the temperature parameter τ is sufficiently small, for instance, $\tau \leq \lambda^2/(m_1 m_2)$, then a random matrix sampled from the pseudo-posterior $\widehat{\pi}_n$ satisfies nearly the same oracle inequality as the nuclear-norm penalized least-squares estimator. Indeed, the term $8m_1 m_2 \tau$, which is the only

difference between the two upper bounds, is in this case negligible with respect to the term involving λ^2.

2.6 Conclusions

We have considered the model of regression with fixed design and established risk bounds for the exponentially weighted aggregate with the Laplace prior. This class of estimators encompasses important particular cases such as the lasso and the Bayesian lasso. The risk bounds established in the present work exhibit a range of values for the temperature parameter for which the EWA with the Laplace prior has a risk bound of the same order as the lasso. This offers a valuable complement to the negative results by Castillo et al. (2015), which show that the Bayesian lasso is not rate-optimal in the sparsity scenario. Note that the Bayesian lasso corresponds to the EWA with the Laplace prior for the temperature parameter $\tau = \sigma^2/n$, where σ^2 is the variance of the noise. Our results imply that in order to get rate-optimality in the sparsity scenario, it is sufficient to choose τ smaller than $\sigma^2/(np)$.

We have extended the result outlined in the previous paragraph in two directions. First, we have shown that one can replace the pseudo-posterior mean by any random sample from the pseudo-posterior distribution. This eventually increases the risk by a negligible additional term, but might be useful from a computational point of view. Second, we have established risk bounds of the same flavour in the case of trace-regression, when the unknown parameter is a nearly low-rank large matrix. This result extends those of (Koltchinskii et al., 2011a) and unifies the risks bounds leading to the "slow" and "fast" rates. Furthermore, our result offers an interpolation between these two extreme cases, see the discussion following 2.5.1.

With some additional work, all the results established in the present work can be extended to the model of regression with random design. Furthermore, the case of a partially labelled sample can be handled by coupling the methodology of the present work with that of Chapter 4. An interesting line of future research is to apply our approach to other priors constructed from convex penalties such as the mixed ℓ_1/ℓ_2-norm used in the group-lasso (Yuan and Lin, 2006), or the weighted ℓ_1-norm of ordered entries used in the slope (Bogdan et al., 2015). Another highly relevant and challenging topic for future work will be to investigate the computational complexity of various methods for approximating the pseudo-posterior mean or for drawing a sample from the pseudo-posterior density.

2.7 Proofs

2.7.1 Proof of the oracle inequality of Theorem 2.3.1

To ease notation, throughout this section we write $\widehat{\boldsymbol{\beta}}$ instead of $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$. Furthermore, for a function $h : \mathbb{R}^p \to \mathbb{R}$, we often write $\int h \, \widehat{\pi}_n$ instead of $\int_{\mathbb{R}^p} h(\boldsymbol{u}) \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}$. We split the proof into three steps. The first step, carried out in 2.7.1, consists in deriving an initial upper bound on the prediction loss from the fundamental inequality stated in (2.4.1). The second step, performed in 2.7.2, shares many common features with the analogous developments for the lasso and provides a proof of (2.3.3). Finally, the third step is a standard bound of the probability of the event $\mathcal{E}_\gamma = \{\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \leq n\lambda/\gamma\}$ based on the union bound and properties of the Gaussian distribution.

Lemma 2.7.1. *For any $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$, we have*

$$\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \leq \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) + 2H(\tau) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2.$$

Proof. On the one hand, inequality (2.4.1) can be rewritten as

$$V_n(\widehat{\boldsymbol{\beta}}) \leq V_n(\bar{\boldsymbol{\beta}}) + \underbrace{V_n(\widehat{\boldsymbol{\beta}}) - \int_{\mathbb{R}^p} V_n(\boldsymbol{u}) \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} + p\tau - \frac{1}{2n}\int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}}_{:=A}. \quad (2.7.1)$$

On the other hand, one can check that

$$V_n(\widehat{\boldsymbol{\beta}}) - \int_{\mathbb{R}^p} V_n(\boldsymbol{u}) \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = \frac{1}{2n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \int_{\mathbb{R}^p} \Big(\frac{1}{2n}\|\mathbf{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1\Big) \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}, \quad (2.7.2)$$

$$\int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = \|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 + \int_{\mathbb{R}^p} \|\mathbf{X}\boldsymbol{u}\|_2^2 \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} - \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2. \quad (2.7.3)$$

These inequalities, combined with the definition of H, given in (2.2.2), yield

$$A = \frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \int_{\mathbb{R}^p} \Big(\frac{1}{n}\|\mathbf{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1\Big) \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} + p\tau - \frac{1}{2n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 \quad (2.7.4)$$

$$= H(\tau) - \frac{1}{2n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2. \quad (2.7.5)$$

Finally, using the definitions of the prediction loss ℓ_n and the potential V_n, we get

$$\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) - \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) = 2\big(V_n(\widehat{\boldsymbol{\beta}}) - V_n(\bar{\boldsymbol{\beta}})\big) + \frac{2}{n}\boldsymbol{\xi}^\top \mathbf{X}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1). \quad (2.7.6)$$

In view of the duality inequality, the term $\boldsymbol{\xi}^\top \mathbf{X}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$ is upper bounded in absolute value by $\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1$. Inserting this inequality and (2.7.1) in (2.7.6) and using relation (2.7.5), we get the claim of the lemma. □

According to 2.7.1, in the event $\mathcal{E}_\gamma = \{\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \leq n\lambda/\gamma\}$, we have

$$\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \leq \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + \frac{2\lambda}{\gamma}(\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1) + 2H(\tau) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2. \quad (2.7.7)$$

Lemma 2.7.2. *For every $J \subset [p]$, we have*

$$\frac{2\lambda}{\gamma}(\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 \leq 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{J,(\gamma+1)/(\gamma-1)}}.$$

This lemma is essentially a copy of Proposition 2 in Chapter 4. We provide here its proof for the sake of self-containedness.

Proof. Let us fix a $J \subset \{1, \ldots, p\}$ and set $\boldsymbol{u} = \widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$. We have

$$\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{u}_J\|_1 + \|\boldsymbol{u}_{J^c}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}_J\|_1 + \gamma\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}_J\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}_{J^c}\|_1. \quad (2.7.8)$$

Using inequalities $\|\bar{\boldsymbol{\beta}}_J\|_1 - \|\widehat{\boldsymbol{\beta}}_J\|_1 \leq \|\boldsymbol{u}_J\|_1$ and $\|\widehat{\boldsymbol{\beta}}_{J^c}\|_1 \geq \|\boldsymbol{u}_{J^c}\|_1 - \|\bar{\boldsymbol{\beta}}_{J^c}\|_1$, we deduce from equation (2.7.8) that

$$\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1 \leq (\gamma+1)\|\boldsymbol{u}_J\|_1 - (\gamma-1)\|\boldsymbol{u}_{J^c}\|_1 + 2\gamma\|\bar{\boldsymbol{\beta}}_{J^c}\|_1. \quad (2.7.9)$$

Now, by definition of the compatibility factor $\kappa_{J,c}$ given by equation (2.2.1), we obtain

$$\|\boldsymbol{u}_J\|_1 - \frac{\gamma-1}{\gamma+1}\|\boldsymbol{u}_{J^c}\|_1 \leq \left(\frac{|J|\|\mathbf{X}\boldsymbol{u}\|_2^2}{n\kappa_{J,(\gamma+1)/(\gamma-1)}}\right)^{1/2}. \quad (2.7.10)$$

Hence, inequalities (2.7.9) end (2.7.10) imply that

$$\frac{2\lambda}{\gamma}(\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 \leq 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + 2ab - a^2, \quad (2.7.11)$$

where we have used the notation $a^2 = \|\mathbf{X}\boldsymbol{u}\|_2^2/n$ and $b^2 = \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{J,(\gamma+1)/(\gamma-1)}}$. Finally, noticing that

$$2ab - a^2 \leq b^2 = \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{J,(\gamma+1)/(\gamma-1)}},$$

we get the claim of the lemma. □

Combining the claims of the previous lemmas and taking the minimum with respect to J and $\bar{\boldsymbol{\beta}}$, we obtain that the inequality

$$\ell_n\big(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star\big) \leq \inf_{\substack{\bar{\boldsymbol{\beta}} \in \mathbb{R}^p \\ J \subset [p]}} \left\{ \ell_n\big(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star\big) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{J,(\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau) \tag{2.7.12}$$

holds in the event $\mathcal{E}_\gamma$. The third and the last step of the proof consists in assessing the probability of this event.

Lemma 2.7.3. *If $\mathbf{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$ is a $n \times p$ deterministic matrix with columns $\boldsymbol{x}^j$ satisfying $\|\boldsymbol{x}^j\|_2^2 \leq n$ and if $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, then, for all $\varepsilon > 0$,*

$$\mathbb{P}\big(\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty > n\varepsilon\big) \leq p \, \exp\big(-n\varepsilon^2/(2\sigma^2)\big). \tag{2.7.13}$$

Proof. By the union bound, we get

$$\mathbb{P}\big(\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty > n\varepsilon\big) = \mathbb{P}\left(\max_{j \in [p]} |\boldsymbol{\xi}^\top \boldsymbol{x}^j| > n\varepsilon\right) \leq \sum_{i=1}^p \mathbb{P}\big(|\boldsymbol{\xi}^\top \boldsymbol{x}^j| > n\varepsilon\big). \tag{2.7.14}$$

Then, noticing that for each $j \in [p]$ the random variable $\boldsymbol{\xi}^\top \boldsymbol{x}^j$ is distributed according to $\mathcal{N}(0, \sigma^2 \|\boldsymbol{x}^j\|_2^2)$, we deduce that

$$\mathbb{P}\left(\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty > n\varepsilon\right) \leq 2 \sum_{j=1}^p \int_{n\varepsilon/(\sigma\|\boldsymbol{x}^j\|_2)}^{+\infty} \phi(u) \, \mathrm{d}u,$$

where ϕ stands for the probability density function of the standard Gaussian distribution. Finally, by using the inequality $\int_x^{+\infty} \phi(u) \, \mathrm{d}u \leq 1/2 \exp(-x^2/2)$ that holds for every $x > 0$, we obtain the result. □

A proof of 2.3.1 can be deduced from the three previous lemmas as follows. Choosing $\gamma = 2$ and $\varepsilon = \lambda/2 \geq \sigma\sqrt{(2/n)\log(p/\delta)}$ in 2.7.3, we get that the event $\mathcal{E}_\gamma$ has a probability at least $1 - \delta$. Furthermore, on this event, we have already established inequality (2.7.12). Finally, upper bounding $H(\tau)$ by $p\tau$ leads to the claim of the theorem.

2.7.2 Proof of the concentration property of Theorem 2.4.1

Let us introduce the set $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : V_n(\boldsymbol{\beta}) \leq \int V_n \, \widehat{\pi}_n + p\tau\}$. Applying 2.4.2 with $t = \sqrt{p}$, we get $\widehat{\pi}_n(\mathcal{B}) \geq 1 - 2e^{-\sqrt{p}/16}$. To prove 2.4.1, it is sufficient to check that in the event $\mathcal{E}_\gamma$ (in

particular, with $\gamma = 2$), every vector $\boldsymbol{\beta}$ from $\mathcal{B}$ satisfies the inequality

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}} \in \mathbb{R}^p \\ J \subset [p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{2\kappa_{J,3}} \right\} + 8p\tau. \tag{2.7.15}$$

In the rest of this proof, $\boldsymbol{\beta}$ is always a vector from $\mathcal{B}$. In view of (2.4.1), it satisfies

$$V_n(\boldsymbol{\beta}) \le 2p\tau + V_n(\bar{\boldsymbol{\beta}}) - \frac{1}{2n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}. \tag{2.7.16}$$

Note that (2.7.16) holds for every $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$. Therefore, it also holds for $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and yields

$$\frac{1}{n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \boldsymbol{\beta})\|_2^2 \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \le 4p\tau. \tag{2.7.17}$$

In addition, we have

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) - \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) = 2\big(V_n(\boldsymbol{\beta}) - V_n(\bar{\boldsymbol{\beta}})\big) + \frac{2}{n} \boldsymbol{\xi}^\top \mathbf{X}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}\|_1). \tag{2.7.18}$$

Combining (2.7.16), (2.7.18) and the duality inequality, we get that in $\mathcal{E}_\gamma$,

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) - \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \le 4p\tau - \frac{1}{n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \tag{2.7.19}$$

$$+ \frac{2\lambda}{\gamma} \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_1 + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}\|_1). \tag{2.7.20}$$

We use now the inequality $\|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \ge 1/2 \|\mathbf{X}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\|_2^2 - \|\mathbf{X}(\boldsymbol{u} - \boldsymbol{\beta})\|_2^2$, in conjunction with (2.7.17), to deduce from (2.7.20) that

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) - \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \le 8p\tau + \frac{2\lambda}{\gamma} \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_1 + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}\|_1) - \frac{1}{2n} \|\mathbf{X}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\|_2^2. \tag{2.7.21}$$

We can apply now 2.7.2 with $\boldsymbol{\beta}$ instead of $\widehat{\boldsymbol{\beta}}$ and $\mathbf{X}/\sqrt{2}$ instead of $\mathbf{X}$ in order to get the claim of 2.4.1.

2.7.3 Proof of Proposition 2.3.1

For the sake of simplicity, we abbreviate $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\beta}}^0 = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ throughout the proof. In particular, notation $\widehat{\beta}_j$ (resp. $\widehat{\beta}_j^0$) will refer to the j-th entry of $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ (resp. $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$). First, observe

that one can write the posterior density as $\widehat{\pi}(\boldsymbol{u}) \propto \exp(-\bar{V}_n(\boldsymbol{u})/\tau)$ with

$$\bar{V}_n(\boldsymbol{u}) = V_n(\boldsymbol{u}) - \frac{1}{2n}\|\boldsymbol{y}\|^2 + \frac{1}{2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\widehat{\boldsymbol{\beta}}^0\|_2^2 \tag{2.7.22}$$

$$= \frac{1}{2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^0)\|_2^2 + \lambda\|\boldsymbol{u}\|_1. \tag{2.7.23}$$

On the one hand, the integration by parts formula yields

$$\int_{\mathbb{R}^p} [\boldsymbol{u}^\top \nabla \bar{V}_n(\boldsymbol{u})]\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = -\tau \int_{\mathbb{R}^p} \boldsymbol{u}^\top \nabla \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = p\tau.$$

On the other hand, the expression of $\bar{V}_n(\boldsymbol{u})$ written in (2.7.23) leads directly to

$$\int_{\mathbb{R}^p} [\boldsymbol{u}^\top \nabla \bar{V}_n(\boldsymbol{u})]\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = \int_{\mathbb{R}^p} G(\boldsymbol{u})\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{\beta}}^0, \tag{2.7.24}$$

where we recall that $G(\boldsymbol{u}) = \|\mathbf{X}\boldsymbol{u}\|_2^2/n + \lambda\|\boldsymbol{u}\|_1 = \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1$. This yields

$$\int_{\mathbb{R}^p} G(\boldsymbol{u})\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = p\tau + \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{\beta}}^0,$$

and, hence,

$$H(\tau) = p\tau - \frac{1}{n}\int_{\mathbb{R}^p} G(\boldsymbol{u})\widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} + \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \tag{2.7.25}$$

$$= \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{\beta}}^0, \tag{2.7.26}$$

which proves the first claim of Proposition 2.3.1. Let us now consider the case where $\widehat{\boldsymbol{\Sigma}}_n = \mathbf{I}_p$. Then, recalling the definition of $\bar{V}_n(\boldsymbol{u})$ in (2.7.23), a straightforward calculation reveals that

$$\bar{V}_n(\boldsymbol{u}) = \frac{\lambda^2 p}{2} + \sum_{j=1}^{p} \left[\frac{1}{2}\Big(u_j - \widehat{\beta}_j^0 + \lambda\,\mathrm{sign}(u_j)\Big)^2 + \lambda\widehat{\beta}_j^0\,\mathrm{sign}(u_j)\right]. \tag{2.7.27}$$

Hence, we deduce that $\widehat{\pi}(\boldsymbol{u}) = \prod_{j=1}^p \widehat{\pi}_j(u_j)$ where

$$\widehat{\pi}_j(t) \propto \exp\Big(-\frac{1}{2\tau}(t - \widehat{\beta}_j^0 + \lambda\,\mathrm{sign}(t))^2 - \frac{\lambda}{\tau}\widehat{\beta}_j^0\,\mathrm{sign}(t)\Big). \tag{2.7.28}$$

Next, let $\varphi(t) = \int_t^{+\infty} \phi(x)\mathrm{d}x$ where ϕ denotes the density function of the standard normal distribution. For a fixed $j \in [p]$, we consider the abbreviations $a = \lambda/\sqrt{\tau}$ and $b = \widehat{\beta}_j^0/\sqrt{\tau}$. Then, the change of variable $u = t/\sqrt{\tau}$ in the first integral below, together with the observation

that $\operatorname{sign}(t) = \operatorname{sign}(t/\sqrt{\tau})$ for all real t, leads to

$$\widehat{\beta}_j = \int t\,\widehat{\pi}_j(t)\,\mathrm{d}t = \sqrt{\tau}\,\frac{\int u \exp\{-\frac{1}{2}(u-b+a\operatorname{sign}(u))^2 - ab\operatorname{sign}(u)\}\,\mathrm{d}u}{\int \exp\{-\frac{1}{2}(u-b+a\operatorname{sign}(u))^2 - ab\operatorname{sign}(u)\}\,\mathrm{d}u} \tag{2.7.29}$$

$$= \sqrt{\tau}\,\frac{(a+b)e^{ab}\varphi(a+b) - (a-b)e^{-ab}\varphi(a-b)}{e^{ab}\varphi(a+b) + e^{-ab}\varphi(a-b)} \tag{2.7.30}$$

$$= \sqrt{\tau}\,\operatorname{sign}(b)\frac{(a+|b|)e^{a|b|}\varphi(a+|b|) - (a-|b|)e^{-a|b|}\varphi(a-|b|)}{e^{a|b|}\varphi(a+|b|) + e^{-a|b|}\varphi(a-|b|)} \tag{2.7.31}$$

$$= \widehat{\beta}^0_j + \lambda\operatorname{sign}(\widehat{\beta}^0_j)\frac{e^{a|b|}\varphi(a+|b|) - e^{-a|b|}\varphi(a-|b|)}{e^{a|b|}\varphi(a+|b|) + e^{-a|b|}\varphi(a-|b|)} \tag{2.7.32}$$

$$= \widehat{\beta}^0_j + \lambda\operatorname{sign}(\widehat{\beta}^0_j)\frac{\Psi(a+|b|) - \Psi(a-|b|)}{\Psi(a+|b|) + \Psi(a-|b|)}, \tag{2.7.33}$$

where $\Psi(t) = e^{t^2/2}\varphi(t)$. In other terms, noticing that $\Psi_\tau(t) = \Psi(t/\sqrt{\tau})$, we have obtained

$$\widehat{\beta}_j = \operatorname{sign}(\widehat{\beta}^0_j)\left(|\widehat{\beta}^0_j| - \lambda w(\tau,\lambda,|\widehat{\beta}^0_j|)\right), \tag{2.7.34}$$

where we have denoted $w(\tau,\lambda,t) = (\Psi_\tau(\lambda-t) - \Psi_\tau(\lambda+t))/(\Psi_\tau(\lambda-t) + \Psi_\tau(\lambda+t))$. Finally, injecting (2.7.34) in (2.7.26) leads easily to the desired expression for H.

2.7.4 Proofs for Stein's unbiased risk estimate (2.3.7)

In what follows, we denote $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ for brevity. The dependance on $\boldsymbol{y}$ will sometimes be made explicit in the proof for clarity. Below, it is understood that all gradients are taken with respect to variable $\boldsymbol{y}$ and that ∂_i refers to the i-th element of the gradient. In addition, for every function $h : \mathbb{R}^n \to \mathbb{R}^n$, we use the notation $\nabla \cdot h$ for the divergence operator $\sum_i \partial_i h_i$. Finally, function f will refer to the non-normalized pseudo-posterior, $f(\boldsymbol{\beta},\boldsymbol{y}) = \exp(-V_n(\boldsymbol{\beta},\boldsymbol{y})/\tau)$, and g to its integral (the normalizing constant), *i.e.*

$$g(\boldsymbol{y}) = \int_{\mathbb{R}^p} f(\boldsymbol{\beta},\boldsymbol{y})\,\mathrm{d}\boldsymbol{\beta}. \tag{2.7.35}$$

According to Stein's formula, an unbiased estimate of the risk of $\widehat{\boldsymbol{\beta}}$—under Gaussian noise—is given by

$$\widehat{R}^{\mathrm{EWA}}(\lambda,\tau) = \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2 - \frac{\sigma^2}{n} + \frac{2\sigma^2}{n}\nabla\cdot(\mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{2.7.36}$$

Therefore, to prove (2.3.7), we need only to show that

$$\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) \;=\; \frac{\mathbf{Cov}_{\widehat{\pi}}(\boldsymbol{\beta})}{n\tau}\mathbf{X}^\top \in \mathbb{R}^{p\times n}. \tag{2.7.37}$$

Indeed, this will imply that

$$\nabla \cdot (\mathbf{X}\widehat{\boldsymbol{\beta}}) = \sum_i \boldsymbol{x}_i^\top \partial_i \widehat{\boldsymbol{\beta}}(\boldsymbol{y}) = \frac{1}{n\tau} \sum_i \boldsymbol{x}_i^\top \mathbf{Cov}_{\widehat{\pi}}(\boldsymbol{\beta}) \boldsymbol{x}_i = \frac{1}{n\tau} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\mathrm{d}\boldsymbol{\beta}),$$

which, combined with (2.7.36), leads to (2.3.7). To do so, we proceed in two steps. First, we prove that

$$\partial_i \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}) \quad = \quad \frac{\boldsymbol{x}_i}{n\tau} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{y})) \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}). \tag{2.7.38}$$

Secondly, we show that

$$\partial_i \widehat{\beta}(\boldsymbol{y}) \quad = \quad \frac{\mathbf{Cov}_{\widehat{\pi}}(\boldsymbol{\beta})}{n\tau} \boldsymbol{x}_i^\top. \tag{2.7.39}$$

Given the notations introduced above, we have

$$\widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}) = \frac{f(\boldsymbol{\beta}, \boldsymbol{y})}{g(\boldsymbol{y})}. \tag{2.7.40}$$

Then, notice that

$$\partial_i f(\boldsymbol{\beta}, \boldsymbol{y}) = -\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{n\tau} \right) f(\boldsymbol{\beta}, \boldsymbol{y}). \tag{2.7.41}$$

Hence, combining (2.7.40) and (2.7.41) yields,

$$\frac{\partial_i f(\boldsymbol{\beta}, \boldsymbol{y})}{g(\boldsymbol{y})} \quad = \quad -\frac{1}{n\tau} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \, \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}). \tag{2.7.42}$$

Moreover, using one more time (2.7.41), we get

$$\begin{aligned} \frac{\partial_i g(\boldsymbol{y})}{g(\boldsymbol{y})} &= \frac{\int \partial_i f(\boldsymbol{\beta}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta}}{g(\boldsymbol{y})} \\ &= -\frac{1}{n\tau} \frac{\int (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) f(\boldsymbol{\beta}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta}}{g(\boldsymbol{y})} \\ &= \frac{\boldsymbol{x}_i^\top}{n\tau} \frac{\int \boldsymbol{\beta} f(\boldsymbol{\beta}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta}}{g(\boldsymbol{y})} - \frac{y_i}{n\tau} \\ &= \frac{1}{n\tau} (\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - y_i), \end{aligned} \tag{2.7.43}$$

where (2.7.43) follows from the definition of $\widehat{\boldsymbol{\beta}}$, f and g. With these remarks in mind, observe that

$$\partial_i \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}) = \frac{\partial_i f(\boldsymbol{\beta}, \boldsymbol{y}) g(\boldsymbol{y}) - f(\boldsymbol{\beta}, \boldsymbol{y}) \partial_i g(\boldsymbol{y})}{g(\boldsymbol{y})^2} \tag{2.7.44}$$

$$= \frac{\partial_i f(\boldsymbol{\beta}, \boldsymbol{y})}{g(\boldsymbol{y})} - \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}) \frac{\partial_i g(\boldsymbol{y})}{g(\boldsymbol{y})} \tag{2.7.45}$$

$$= \frac{\boldsymbol{x}_i^\top}{n\tau} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{y}))\, \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}), \tag{2.7.46}$$

where the last line follows easily by combining (2.7.42) and (2.7.43). We have therefore proved (2.7.38) and now proceed to showing (2.7.39). To that aim, we write

$$\partial_i \widehat{\boldsymbol{\beta}}(\boldsymbol{y}) = \partial_i \int_{\mathbb{R}^p} \boldsymbol{\beta}\, \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\beta} = \int_{\mathbb{R}^p} \boldsymbol{\beta}\, \partial_i \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\beta}. \tag{2.7.47}$$

Using (2.7.38) and then transposing the product $\boldsymbol{x}_i^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{y})) \in \mathbb{R}$ we have,

$$\begin{aligned}
\partial_i \widehat{\boldsymbol{\beta}}(\boldsymbol{y}) &= \frac{1}{n\tau} \int_{\mathbb{R}^p} \boldsymbol{\beta} (\boldsymbol{x}_i^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{y})) \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y})) \mathrm{d}\boldsymbol{\beta} \\
&= \frac{1}{n\tau} \int_{\mathbb{R}^p} (\boldsymbol{\beta}\boldsymbol{\beta}^\top - \boldsymbol{\beta} \widehat{\boldsymbol{\beta}}(\boldsymbol{y})^\top) \boldsymbol{x}_i\, \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta} \\
&= \frac{1}{n\tau} \left(\int_{\mathbb{R}^p} \boldsymbol{\beta}\boldsymbol{\beta}^\top \widehat{\pi}(\boldsymbol{\beta}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{y}) \widehat{\boldsymbol{\beta}}(\boldsymbol{y})^\top \right) \boldsymbol{x}_i,
\end{aligned}$$

which is equivalent to (2.7.39) and concludes the proof of (2.3.7).

2.7.5 Proof of the results in the matrix case

To ease notation, throughout this section we write $\widehat{\mathbf{B}}$ instead of $\widehat{\mathbf{B}}^{\mathrm{EWA}}$. Furthermore, for a function $h : \mathcal{M}_{m_1,m_2} \to \mathbb{R}$, we often use the notation $\int h\, \widehat{\pi}_n$ or $\int_{\mathcal{M}} h(\mathbf{U})\, \widehat{\pi}_n(\mathrm{d}\mathbf{U})$ instead of $\int_{\mathcal{M}_{m_1,m_2}} h(\mathbf{U})\, \widehat{\pi}_n(\mathbf{U})\, \mathrm{d}\mathbf{U}$. In this section, we prove 2.5.1 and 2.5.2. To do so, we state and prove 2.7.1 as well as 2.7.2 that will be used throughout the proofs.

2.7.1 is an extension of the fundamental theorem of calculus in the case of locally-Lipschitz functions. It will be very useful to work with any (pseudo-)posterior of the form $\widehat{\pi}_n$ corresponding to convex penalties.

Let us first recall that a function $f : \mathbb{R} \to \mathbb{R}$ is called locally-Lipschitz-continuous, or locally-Lipschitz, if it is Lipschitz-continuous on any bounded interval. Clearly, any locally-Lipschitz function is absolutely continuous (in the sense of Definition 7.17 in Rudin (1987)) and, therefore, is almost everywhere (with respect to the Lebesgue measure) differentiable.

Proposition 2.7.1. *For any locally-Lipschitz function f such that $\lim_{|x|\to\infty} f(x) = 0$ and $f' \in L^1(\mathbb{R})$, we have*

$$\int_{\mathbb{R}} f'(x)\mathrm{d}x = 0.$$

Proof. The result of (Rudin, 1987, Theorem 7.20) implies that for any $a > 0$,

$$\int_{-a}^{a} f'(x)\,\mathrm{d}x = f(a) - f(-a).$$

Since, by assumption, the derivative f' is absolutely integrable over $\mathbb{R}$, we have

$$\int_{\mathbb{R}} f'(x)\,\mathrm{d}x = \lim_{a\to+\infty} \int_{-a}^{a} f'(x)\,\mathrm{d}x = \lim_{a\to+\infty} \big(f(a) - f(-a)\big) = 0.$$

This completes the proof. □

Corollary 2.7.1. *Let $\mathbf{0}_{m_1,m_2}$ be the null element of $\mathcal{M}_{m_1,m_2}$. Then*

$$\int_{\mathcal{M}} \nabla\widehat{\pi}_n(\mathbf{U})\,\mathrm{d}\mathbf{U} = \mathbf{0}_{m_1,m_2}.$$

Proof. We want to prove that $\int_{\mathcal{M}} [\partial_{\mathbf{U}_{\boldsymbol{sd}}} \widehat{\pi}_n(\mathbf{U})]\,\mathrm{d}\mathbf{U} = 0$ for any $d := (k,l) \in [m_1] \times [m_2]$. To this end, we will simply prove that $\int_{\mathbb{R}} [\partial_{\mathbf{U}_{\boldsymbol{sd}}} \widehat{\pi}_n(\mathbf{U})]\,\mathrm{d}\mathbf{U}_{\boldsymbol{sd}} = 0$, where the integration is done with respect to the $\boldsymbol{sd}$-th entry of $\mathbf{U}$ when all the other entries are fixed.

The function $\mathbf{U} \mapsto V_n(\mathbf{U})$ is locally-Lipschitz as the sum of a continuously differentiable function (the quadratic term) and a Lipschitz term (the nuclear norm). We note in passing that any norm in a finite-dimensional space is Lipschitz continuous thanks to the triangle inequality and the equivalence of norms. In addition, one easily checks that $\|\mathbf{U}\|_1^2 \geq \|\mathbf{U}\|^2 = \|\mathbf{U}^\top\mathbf{U}\| \geq \max_l (\mathbf{U}^\top\mathbf{U})_{l,l} \geq \mathbf{U}_{\boldsymbol{sd}}^2$. This implies that if $\mathbf{U}_{\boldsymbol{sd}}$ tends to infinity while all the other entries of $\mathbf{U}$ remain fixed, the nuclear norm $\|\mathbf{U}\|_1$ tends to infinity[6].

As a consequence, the function $\mathbf{U}_{\boldsymbol{sd}} \mapsto \pi_n(\mathbf{U}) \propto \exp\{-V_n(\mathbf{U})/\tau\}$ is locally-Lipschitz and tends to zero when $|\mathbf{U}_{\boldsymbol{sd}}| \to \infty$. This implies that we can apply 2.7.1 and the claim of the corollary follows. □

Corollary 2.7.2. *With the notation introduced in 2.5, we have*

$$\int_{\mathcal{M}} \langle \mathbf{U}, \nabla V_n(\mathbf{U})\rangle\, \widehat{\pi}_n(\mathbf{U})\,\mathrm{d}\mathbf{U} = \tau m_1 m_2. \tag{2.7.48}$$

[6]This assertion can be also established for any other norm using the equivalence of norms in $\mathcal{M}_{m_1,m_2}$.

Proof of Corrolary 2.7.2. We first remark that (2.7.48) can be equivalently written as

$$\sum_{d\in[m_1]\times[m_2]} \int_{\mathcal{M}} \mathbf{U}_{sd}\left[\partial_{\mathbf{U}_{sd}} V_n(\mathbf{U})\right] \widehat{\pi}_n(\mathbf{U})\, \mathrm{d}\mathbf{U} = \tau m_1 m_2. \tag{2.7.49}$$

To establish this identity, it suffices to prove that each integral of the left-hand side is equal to τ. We have already checked in the proof of 2.7.1 that the mapping $\mathbf{U}_{sd} \mapsto \pi_n(\mathbf{U})$ is locally-Lipschitz and tends to zero when $\mathbf{U}_{sd}$ tends to infinity. Furthermore, the latter convergence is exponential so that $\mathbf{U}_{sd}\pi_n(\mathbf{U})$ tends to zero as well, when $\mathbf{U}_{sd}$ tends to infinity. In view of 2.7.1, this yields

$$\int_{\mathcal{M}} \frac{\partial[\mathbf{U}_{sd}\, \widehat{\pi}_n(\mathbf{U})]}{\partial \mathbf{U}_{sd}}\, \mathrm{d}\mathbf{U} = 0. \tag{2.7.50}$$

Moreover, we remark that

$$\frac{\partial[\mathbf{U}_{sd}\, \widehat{\pi}_n(\mathbf{U})]}{\partial \mathbf{U}_{sd}} = \mathbf{U}_{sd} \frac{\partial \widehat{\pi}_n(\mathbf{U})}{\partial \mathbf{U}_{sd}} + \widehat{\pi}_n(\mathbf{U}) = -\mathbf{U}_{sd} \frac{\partial V_n(\mathbf{U})}{\tau \partial \mathbf{U}_{sd}}\, \widehat{\pi}_n(\mathbf{U}) + \widehat{\pi}_n(\mathbf{U}). \tag{2.7.51}$$

Therefore, multiplying by τ and integrating over $\mathcal{M}_{m_1,m_2}$, we get

$$\int_{\mathcal{M}} \mathbf{U}_{sd} \frac{\partial V_n(\mathbf{U})}{\partial \mathbf{U}_{sd}}\, \widehat{\pi}_n(\mathbf{U})\, \mathrm{d}\mathbf{U} = \tau \int_{\mathcal{M}} \widehat{\pi}_n(\mathbf{U}) \mathrm{d}\mathbf{U} = \tau.$$

This completes the proof. □

The next Proposition is the matrix analogue of 2.4.1.

Proposition 2.7.2. *Let $\widehat{\pi}_n(\mathbf{U}) \propto \exp\left(-V_n(\mathbf{U})/\tau\right)$ be the pseudo-posterior defined by (2.5.7). Then, for every $\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$, we have*

$$\int_{\mathcal{M}} V_n(\mathbf{U})\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) \le V_n(\bar{\mathbf{B}}) - \frac{1}{2} \int_{\mathcal{M}} \|\bar{\mathbf{B}} - \mathbf{U}\|^2_{L_2(\mathcal{X})} \widehat{\pi}_n(\mathrm{d}\mathbf{U}) + m_1 m_2 \tau. \tag{2.7.52}$$

Furthermore,

$$\int_{\mathcal{M}} \|\mathbf{U} - \widehat{\mathbf{B}}^{\mathrm{EWA}}\|^2_{L_2(\mathcal{X})}\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) \le m_1 m_2 \tau. \tag{2.7.53}$$

Proof. The convexity of $\mathbf{U} \mapsto \|\mathbf{U}\|_1$ and the strong convexity of the function $\boldsymbol{\theta} \mapsto \|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2$ applied in $\boldsymbol{\theta} = \sum_{i\in[n]} \langle \mathbf{X}_i, \mathbf{U}\rangle$ imply that for any $\mathbf{U}, \bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$,

$$V_n\big(\bar{\mathbf{B}}\big) \ge V_n(\mathbf{U}) + \langle \bar{\mathbf{B}} - \mathbf{U}, \nabla V_n(\mathbf{U})\rangle + \frac{1}{2}\|\bar{\mathbf{B}} - \mathbf{U}\|^2_{L_2(\mathcal{X})}. \tag{2.7.54}$$

In order to prove 2.7.2 we rely on Corrolaries 2.7.1 and 2.7.2 from 2.7.1:

$$\int_{\mathcal{M}} \nabla V_n(\mathbf{U})\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) = 0 \qquad \text{and} \qquad \int_{\mathcal{M}} \langle \mathbf{U}, \nabla V_n(\mathbf{U}) \rangle\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) = m_1 m_2 \tau. \tag{2.7.55}$$

We integrate inequality (2.7.54) over $\mathcal{M}_{m_1,m_2}$ with respect to the density $\widehat{\pi}_n$ and use equalities (2.7.55). This yields

$$V_n\big(\bar{\mathbf{B}}\big) \geq \int_{\mathcal{M}} V_n(\mathbf{U})\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) - m_1 m_2 \tau + \frac{1}{2} \int_{\mathcal{M}} \|\bar{\mathbf{B}} - \mathbf{U}\|^2_{L_2(\mathcal{X})}\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}), \tag{2.7.56}$$

which concludes the proof of the first assertion of 2.7.2. The second assertion follows from the first one by choosing $\bar{\mathbf{B}} = \widehat{\mathbf{B}}$. □

Lemma 2.7.4. *In the event $\mathcal{E}_\gamma = \{\|\boldsymbol{\xi}^\top \mathcal{X}\| \leq n\lambda/\gamma\}$, for any $\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$, we have*

$$\ell_n(\widehat{\mathbf{B}}, \mathbf{B}^\star) \leq \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + \frac{2\lambda}{\gamma}\big(\gamma\|\bar{\mathbf{B}}\|_1 - \gamma\|\widehat{\mathbf{B}}\|_1 + \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|_1\big) - \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})} + 2H(\tau). \tag{2.7.57}$$

Proof. On the one hand, using the definitions of the prediction loss ℓ_n and the empirical loss L_n, as well as the Von Neumann inequality, we get

$$\ell_n(\widehat{\mathbf{B}}, \mathbf{B}^\star) - \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) = 2(V_n(\widehat{\mathbf{B}}) - V_n(\bar{\mathbf{B}})) + \frac{2}{n}\sum_{i\in[n]} \xi_i \langle \mathbf{X}_i, \widehat{\mathbf{B}} - \bar{\mathbf{B}} \rangle + 2\lambda(\|\bar{\mathbf{B}}\|_1 - \|\widehat{\mathbf{B}}\|_1) \tag{2.7.58}$$

$$\leq 2(V_n(\widehat{\mathbf{B}}) - V_n(\bar{\mathbf{B}})) + \frac{2}{n}\|\boldsymbol{\xi}^\top \mathcal{X}\|\|\widehat{\mathbf{B}} - \bar{\mathbf{B}}\|_1 + 2\lambda(\|\bar{\mathbf{B}}\|_1 - \|\widehat{\mathbf{B}}\|_1) \tag{2.7.59}$$

$$\overset{(\text{in } \mathcal{E}_\gamma)}{\leq} 2(V_n(\widehat{\mathbf{B}}) - V_n(\bar{\mathbf{B}})) + 2\lambda(\|\bar{\mathbf{B}}\|_1 - \|\widehat{\mathbf{B}}\|_1) + \frac{2\lambda}{\gamma}\|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|_1. \tag{2.7.60}$$

Notice that inequality (2.7.52) can be rewritten as

$$V_n(\widehat{\mathbf{B}}) \leq V_n(\bar{\mathbf{B}}) + \underbrace{V_n(\widehat{\mathbf{B}}) - \int_{\mathcal{M}} V_n \widehat{\pi}_n + m_1 m_2 \tau - \frac{1}{2}\int_{\mathcal{M}} \|\bar{\mathbf{B}} - \mathbf{U}\|^2_{L_2(\mathcal{X})}\, \widehat{\pi}_n(\mathrm{d}\mathbf{U})}_{:=A}. \tag{2.7.61}$$

One can check that

$$V_n(\widehat{\mathbf{B}}) - \int_{\mathcal{M}} V_n \, \widehat{\pi}_n = \frac{1}{2}\|\widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})} + \lambda\|\widehat{\mathbf{B}}\|_1 - \int_{\mathcal{M}} \Big(\frac{1}{2}\|\mathbf{U}\|^2_{L_2(\mathcal{X})} + \lambda\|\mathbf{U}\|_1\Big)\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}), \tag{2.7.62}$$

$$\int \|\mathbf{U} - \bar{\mathbf{B}}\|^2_{L_2(\mathcal{X})}\, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) = \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})} + \int \|\mathbf{U}\|^2_{L_2(\mathcal{X})}\ \widehat{\pi}_n(\mathrm{d}\mathbf{U}) - \|\widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})}. \tag{2.7.63}$$

These inequalities, combined with the definition of H, given in (2.5.3), yield

$$A = H(\tau) - \frac{1}{2}\|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})}. \tag{2.7.64}$$

Inserting this inequality in (2.7.61) and using relation (2.7.60), we get the claim of the lemma. □

The next step is to establish the counterpart of 2.7.2 in the matrix setting.

Lemma 2.7.5. *For every $J \in [\mathrm{rank}(\bar{\mathbf{B}})]$, we have*

$$\frac{2\lambda}{\gamma}\big(\gamma\|\bar{\mathbf{B}}\|_1 - \gamma\|\widehat{\mathbf{B}}\|_1 + \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|_1\big) - \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})} \le 4\lambda\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}})\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{\bar{\mathbf{B}},J,(\gamma+1)/(\gamma-1)}}. \tag{2.7.65}$$

Proof. To ease notation, let us write $\bar{\mathbf{B}}_J$ and $\bar{\mathbf{B}}_{J^c}$ instead of $\mathcal{P}_{\bar{\mathbf{B}},J}(\bar{\mathbf{B}}) = \mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}})$ and $\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}})$, respectively. Clearly, $\bar{\mathbf{B}} = \bar{\mathbf{B}}_J + \bar{\mathbf{B}}_{J^c}$. Recall that $r = \mathrm{rank}(\bar{\mathbf{B}})$ and $\bar{\mathbf{B}} = \mathbf{V}_1\mathbf{\Sigma}\mathbf{V}_2^\top$ is the singular value decomposition of $\bar{\mathbf{B}}$. Note that the matrices $\mathbf{\Pi}_{1,J^c} = \mathbf{I}_{m_1} - \mathbf{V}_{1,J}\mathbf{V}_{1,J}^\top$ and $\mathbf{\Pi}_{2,J^c} = \mathbf{I}_{m_2} - \mathbf{V}_{2,J}\mathbf{V}_{2,J}^\top$ are orthogonal projectors and, for every matrix $\mathbf{U} \in \mathcal{M}$, we have $\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = \mathbf{\Pi}_{1,J^c}\mathbf{U}\mathbf{\Pi}_{2,J^c}$.

Let $\mathbf{W}$ be a $m_1 \times m_2$ matrix such that $\|\mathbf{W}\| = 1$ and $\langle \mathcal{P}_{\bar{\mathbf{B}},J^c}(\widehat{\mathbf{B}}), \mathbf{W}\rangle = \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\widehat{\mathbf{B}})\|_1$. We set $\mathbf{D} = \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top + \mathbf{\Pi}_{1,J^c}\mathbf{W}\mathbf{\Pi}_{2,J^c}$. It is clear that

$$\|\bar{\mathbf{B}}\|_1 \le \|\bar{\mathbf{B}}_J\|_1 + \|\bar{\mathbf{B}}_{J^c}\|_1 = \langle \bar{\mathbf{B}}_J, \mathbf{D}\rangle + \|\bar{\mathbf{B}}_{J^c}\|_1 \tag{2.7.66}$$

and, in view of the von Neumann inequality, $\|\widehat{\mathbf{B}}\|_1 \ge \langle \widehat{\mathbf{B}}, \mathbf{D}\rangle$. This implies that

$$\|\bar{\mathbf{B}}\|_1 - \|\widehat{\mathbf{B}}\|_1 \le \|\bar{\mathbf{B}}_{J^c}\|_1 + \langle \bar{\mathbf{B}}_J - \widehat{\mathbf{B}}, \mathbf{D}\rangle. \tag{2.7.67}$$

As shown in (Koltchinskii et al., 2011a), $\langle \bar{\mathbf{B}}_J - \widehat{\mathbf{B}}, \mathbf{D}\rangle \le \|\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}})\|_1 - \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\widehat{\mathbf{B}})\|_1$. For

the sake of self-containedness, we reproduce their proof here. We have

$$\langle \bar{\mathbf{B}}_J - \widehat{\mathbf{B}}, \mathbf{D} \rangle = \langle \bar{\mathbf{B}}_J - \widehat{\mathbf{B}}, \mathbf{V}_{1,J}\mathbf{V}_{2,J} + \mathbf{\Pi}_{1,J^c}\mathbf{W}\mathbf{\Pi}_{2,J^c} \rangle \tag{2.7.68}$$

$$= \langle \bar{\mathbf{B}}_J - \widehat{\mathbf{B}}, \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top \rangle + \langle \mathbf{\Pi}_{1,J^c}(\bar{\mathbf{B}}_J - \widehat{\mathbf{B}})\mathbf{\Pi}_{2,J^c}, \mathbf{W} \rangle \tag{2.7.69}$$

$$= \langle \bar{\mathbf{B}} - \widehat{\mathbf{B}}, \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top \rangle - \langle \mathcal{P}_{\bar{\mathbf{B}},J^c}(\widehat{\mathbf{B}}), \mathbf{W} \rangle \tag{2.7.70}$$

$$= \langle \bar{\mathbf{B}} - \widehat{\mathbf{B}}, \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top \rangle - \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\widehat{\mathbf{B}})\|_1. \tag{2.7.71}$$

In addition, using the triangle inequality, we get $\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\widehat{\mathbf{B}})\|_1 \geq \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}})\|_1 - \|\bar{\mathbf{B}}_{J^c}\|_1$. Thus, we get

$$\langle \bar{\mathbf{B}}_J - \widehat{\mathbf{B}}, \mathbf{D} \rangle \leq \langle \bar{\mathbf{B}} - \widehat{\mathbf{B}}, \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top \rangle - \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}})\|_1 + \|\bar{\mathbf{B}}_{J^c}\|_1. \tag{2.7.72}$$

Finally, one easily checks that $\langle \bar{\mathbf{B}} - \widehat{\mathbf{B}}, \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top \rangle = \langle \mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}}), \mathbf{V}_{1,J}\mathbf{V}_{2,J}^\top \rangle \leq \|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}})\|_1$. Combining this inequality with (2.7.67) and (2.7.72), we get

$$\|\bar{\mathbf{B}}\|_1 - \|\widehat{\mathbf{B}}\|_1 \leq 2\|\bar{\mathbf{B}}_{J^c}\|_1 + \|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}})\|_1 - \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}} - \widehat{\mathbf{B}})\|_1. \tag{2.7.73}$$

If we set $\mathbf{M} = \bar{\mathbf{B}} - \widehat{\mathbf{B}}$, then we have already shown that

$$\frac{2\lambda}{\gamma}\big\{\gamma\|\bar{\mathbf{B}}\|_1 - \gamma\|\widehat{\mathbf{B}}\|_1 + \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|_1\big\} - \|\bar{\mathbf{B}} - \widehat{\mathbf{B}}\|^2_{L_2(\mathcal{X})} \tag{2.7.74}$$

$$\leq 4\lambda\|\bar{\mathbf{B}}_{J^c}\|_1 + \frac{2\lambda}{\gamma}\big(\gamma\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 - \gamma\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 + \|\mathbf{M}\|_1\big) - \|\mathbf{M}\|^2_{L_2(\mathcal{X})}. \tag{2.7.75}$$

We remark that

$$\gamma\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 - \gamma\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 + \|\mathbf{M}\|_1 \leq (\gamma+1)\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 - (\gamma-1)\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 \tag{2.7.76}$$

Now, by definition of the compatibility factor $\kappa_{\bar{\mathbf{B}},J,c}$ given by equation (2.5.2), we obtain

$$\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 - \frac{\gamma-1}{\gamma+1}\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 \leq \left(\frac{|J|\,\|\mathbf{M}\|^2_{L_2(\mathcal{X})}}{n\kappa_{\bar{\mathbf{B}},J,(\gamma+1)/(\gamma-1)}}\right)^{1/2}. \tag{2.7.77}$$

Hence, inequalities (2.7.76) end (2.7.77) imply that

$$\frac{2\lambda}{\gamma}\big(\gamma\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 - \gamma\|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{M})\|_1 + \|\mathbf{M}\|_1\big) - \|\mathbf{M}\|^2_{L_2(\mathcal{X})} \leq 2ab - a^2, \tag{2.7.78}$$

where we have used the notation $a^2 = \|\mathbf{M}\|^2_{L_2(\mathcal{X})}$ and $b^2 = \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{\bar{\mathbf{B}},J,(\gamma+1)/(\gamma-1)}}$. Finally, noticing

that $2ab - a^2 \leq b^2$ we get the claim of the lemma. □

Combining the claims of the previous lemmas and taking the minimum with respect to J and $\bar{\mathbf{B}}$, we obtain that the inequality

$$\ell_n\big(\widehat{\mathbf{B}}, \mathbf{B}^\star\big) \leq \inf_{\substack{\bar{\mathbf{B}} \in \mathcal{M} \\ J \subset [\mathrm{rank}(\bar{\mathbf{B}})]}} \left\{ \ell_n\big(\bar{\mathbf{B}}, \mathbf{B}^\star\big) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}}, J^c}(\bar{\mathbf{B}})\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{\bar{\mathbf{B}}, J, (\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau) \quad (2.7.79)$$

holds in the event $\mathcal{E}_\gamma$. At this point, we remark that we have proved the more general result of Inequality (2.5.13).

The third and last step of the proof consists in assessing the probability of the event $\mathcal{E}_\gamma$. We rely on Theorem 4.1.1 from (Tropp, 2015) that provides a comprehensive account on matrix concentration inequalities.

Lemma 2.7.6. *Let $\mathcal{X}$ be a fixed design tensor and $v_\mathcal{X}$ be defined by (2.5.10). If $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, then, for all $\varepsilon > 0$,*

$$\mathbb{P}\big(\|\boldsymbol{\xi}^\top \mathcal{X}\| > n\varepsilon\big) \leq (m_1 + m_2) \exp\big(-n\varepsilon^2/(2\sigma^2 v_\mathcal{X}^2)\big). \quad (2.7.80)$$

Proof. It is clear that ξ_i/σ are standard gaussian random variables. Therefore, we can apply (Tropp, 2015, Theorem 4.1.1) to the $m_1 \times m_2$ matrix

$$\mathbf{Z} = \sum_{i=1}^n \xi_i \mathbf{X}_i / \sigma. \quad (2.7.81)$$

One easily checks that

$$v(\mathbf{Z}) = \|\mathbb{E}(\mathbf{Z}\mathbf{Z}^\top)\| \vee \|\mathbb{E}(\mathbf{Z}^\top \mathbf{Z})\| \quad (2.7.82)$$

$$= \left\| \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \right\| \bigvee \left\| \sum_{i=1}^n \mathbf{X}_i^\top \mathbf{X}_i \right\| = v_\mathcal{X}. \quad (2.7.83)$$

Therefore,

$$\mathbb{P}\big(\|\boldsymbol{\xi}^\top \mathcal{X}/\sigma\| > n\varepsilon/\sigma\big) \leq (m_1 + m_2) \exp\big(-n\varepsilon^2/(2\sigma^2 v_\mathcal{X})\big), \quad (2.7.84)$$

from which we deduce the claim of 2.7.6. □

A proof of 2.5.1 can be deduced from the three previous lemmas as follows. Choosing $\gamma = 2$ and $\varepsilon = \lambda/\gamma \geq \sigma v_\mathcal{X} \sqrt{2/n \log((m_1 + m_2)/\delta)}$ in 2.7.6, we get that the event $\mathcal{E}_\gamma$ has a probability at least $1 - \delta$. Furthermore, on this event, we have already established inequality (2.7.79), which coincides with the claim of 2.5.1.

We conclude this section by proving 2.5.2 which is the analogue of 2.4.1. Let us introduce the (random) set $\mathcal{B} = \{\mathbf{B} \in \mathcal{M}_{m_1,m_2} : V_n(\mathbf{B}) \le \int V_n \, \widehat{\pi}_n + m_1 m_2 \tau\}$. Applying (2.5.14) with $t = \sqrt{m_1 m_2}$, we get $\widehat{\pi}_n(\mathcal{B}) \ge 1 - 2e^{-\sqrt{m_1 m_2}/16}$. To prove 2.5.2, it is sufficient to check that in the event $\mathcal{E}_\gamma$ (in particular, with $\gamma = 2$), every matrix $\mathbf{B}$ from $\mathcal{B}$ satisfies the inequality

$$\ell_n(\mathbf{B}, \mathbf{B}^\star) \le \inf_{\substack{\bar{\mathbf{B}} \in \mathcal{M} \\ J \in [\mathrm{rank}(\bar{\mathbf{B}})]}} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}}, J^c}(\bar{\mathbf{B}})\|_1 + \frac{9\lambda^2 |J|}{2\kappa_{\bar{\mathbf{B}}, J, 3}} \right\} + 8 m_1 m_2 \tau. \tag{2.7.85}$$

In the rest of this proof, $\mathbf{B}$ is always a matrix from $\mathcal{B}$. In view of (2.7.52), it satisfies

$$V_n(\mathbf{B}) \le 2 m_1 m_2 \tau + V_n(\bar{\mathbf{B}}) - \frac{1}{2} \int_{\mathcal{M}} \|\mathbf{U} - \bar{\mathbf{B}}\|^2_{L_2(\mathcal{X})} \, \widehat{\pi}_n(\mathrm{d}\mathbf{U}). \tag{2.7.86}$$

Note that (2.7.86) holds for every $\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$. Therefore, it also holds for $\bar{\mathbf{B}} = \mathbf{B}$ and yields

$$\int_{\mathcal{M}} \|\mathbf{U} - \mathbf{B}\|^2_{L_2(\mathcal{X})} \, \widehat{\pi}_n(\mathrm{d}\mathbf{U}) \le 4 m_1 m_2 \tau. \tag{2.7.87}$$

In addition, we have

$$\ell_n(\mathbf{B}, \mathbf{B}^\star) - \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) = 2\big(V_n(\mathbf{B}) - V_n(\bar{\mathbf{B}})\big) + \frac{2}{n} \sum_{i=1}^n \xi_i \langle \mathbf{X}_i, \mathbf{B} - \bar{\mathbf{B}} \rangle + 2\lambda(\|\bar{\mathbf{B}}\|_1 - \|\mathbf{B}\|_1). \tag{2.7.88}$$

Combining (2.7.86), (2.7.88) and the Von Neuman inequality, we get that in $\mathcal{E}_\gamma$

$$\ell_n(\mathbf{B}, \mathbf{B}^\star) - \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) \le 4 m_1 m_2 \tau + \frac{2\lambda}{\gamma}\big(\gamma \|\bar{\mathbf{B}}\|_1 - \gamma \|\mathbf{B}\|_1 + \|\mathbf{B} - \bar{\mathbf{B}}\|_1\big) \tag{2.7.89}$$

$$- \int_{\mathcal{M}} \|\mathbf{U} - \bar{\mathbf{B}}\|^2_{L_2(\mathcal{X})} \, \widehat{\pi}_n(\mathrm{d}\mathbf{U}). \tag{2.7.90}$$

We use now the inequality $\|\mathbf{U} - \bar{\mathbf{B}}\|^2_{L_2(\mathcal{X})} \ge 1/2 \|\mathbf{B} - \bar{\mathbf{B}}\|^2_{L_2(\mathcal{X})} - \|\mathbf{U} - \mathbf{B}\|^2_{L_2(\mathcal{X})}$, in conjunction with (2.7.87), to deduce from (2.7.90) that

$$\ell_n(\mathbf{B}, \mathbf{B}^\star) - \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) \le 8 m_1 m_2 \tau + \frac{2\lambda}{\gamma} \|\mathbf{B} - \bar{\mathbf{B}}\|_1 + 2\lambda(\|\bar{\mathbf{B}}\|_1 - \|\mathbf{B}\|_1) - \frac{1}{2} \|\mathbf{B} - \bar{\mathbf{B}}\|^2_{L_2(\mathcal{X})}. \tag{2.7.91}$$

We can apply now 2.7.5 with $\mathbf{B}$ instead of $\widehat{\mathbf{B}}$ and $\mathcal{X}/\sqrt{2}$ instead of $\mathcal{X}$ in order to get the claim of 2.5.2.

Chapter 3

Computation guarantees with Langevin Monte Carlo

Contents

Abstract

We study the behaviour of the Langevin Monte Carlo approximation method to estimate log-concave densities. In the existing literature, the density is assumed to be strongly log-concave and its gradient Lipschitz continuous. In our case we provide results in the spirit of Dalalyan (2016), Durmus and Moulines (2016) and Dalalyan (2017) without assuming that the gradient of the potential of the density is Lipschitz continuous. In particular, it will allow us to consider the exponentially weighted aggregate with Laplace prior estimate as in Chapter 2 (Dalalyan et al. (2016)). In this study, provided that the gram matrix is invertible, we provide a method that offers guarantees in the sense of the Wasserstein metrics. These results will provide an explicit upper bound on the quality of the sampling and therefore of the approximation of the estimate.

3.1 Introduction, context and notations

Let p be a positive integer and π be a log-concave distribution density in $\mathbb{R}^p$. Then the potential f associated with π is a convex function and takes value in $\mathbb{R}^p$ such that for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\pi(\boldsymbol{\beta}) \propto \exp\{-f(\boldsymbol{\beta})\}. \tag{3.1.1}$$

The potential f can be seen as the negative log-likelihood or the negative log-posterior. In this paper, we study approximation methods of the quantity

$$\widehat{\boldsymbol{\beta}}_\tau = \frac{\int_{\mathbb{R}^p} \boldsymbol{u} \exp\{-\frac{f(\boldsymbol{u})}{\tau}\} \mathrm{d}\boldsymbol{u}}{\int_{\mathbb{R}^p} \exp\{-\frac{f(\boldsymbol{u})}{\tau}\} \mathrm{d}\boldsymbol{u}}, \tag{3.1.2}$$

for any $\tau \geq 0$ such that $\int_{\mathbb{R}^p} \exp\{-\frac{f(\boldsymbol{u})}{\tau}\} \mathrm{d}\boldsymbol{u}$ is finite. Provided it exists, the limit of $\widehat{\boldsymbol{\beta}}_\tau$ when τ tends to 0 is the maximum likelihood estimate. In that case, closed solutions or approximations of the estimates of such forms have been thoroughly studied in the literature. There are well-known guarantees on the quality of the approximations that often come from optimization methods as described in Boyd and Vandenberghe (2004).

However, almost every time τ is not null, the approximation of $\widehat{\boldsymbol{\beta}}_\tau$ requires sampling methods for which convergence properties are less understood. The authors of Dalalyan (2016), Durmus and Moulines (2016) and Dalalyan (2017) answered this question when $\widehat{\boldsymbol{\beta}}_\tau$ is approximated by averaging a discretized Langevin Monte Carlo sampling algorithm with respect to π. The authors of these papers offer non-asymptotic guarantees of the quality of the approximation in the sense of Kullback-Leibler metric (Dalalyan et al. (2016)) and in the sense of the Wasserstein distance (Durmus and Moulines (2016) and Dalalyan (2017)). Such results offer practical insights of the number of required iterations in order to obtain a desired precision. In these

papers, the rate of convergence of the approximates are directly linked to the property of strong convexity of f and to the Lipschitz property of ∇f, the gradient of f.

These results offer useful guarantees in the context of pseudo-bayesian methods. For example, this question can be motivated by the pseudo-bayesian analogue to penalized regression. Let us consider instances of the potential f that can be written as

$$f(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + g(\boldsymbol{\beta}), \tag{3.1.3}$$

where, from a frequentist point of view, L is the fitting term while g is the penalization term. On a Bayesian basis, L is the negative log-likelihood and g is the logarithm of the prior distribution of the parameter $\boldsymbol{\beta}$. In that context, $\widehat{\boldsymbol{\beta}}_\tau$ is the pseudo-Bayesian estimate.

No matter the design of the data, the results of Dalalyan et al. (2016), Durmus and Moulines (2016) and Dalalyan (2017) applied very well to the pseudo-bayesian analogue of the Ridge regression. Indeed, in that case, ∇f is lipschitz-continuous, g is the ℓ_2-norm that is strongly convex and L is convex, which makes in turn f strongly convex.

However, one may want to slightly relax the assumptions and to keep benefitting from non-asymptotic guarantees. For example, one may wish to consider the case where f is strongly convex but ∇f is not Lipschitz-continuous. Another instance of practical motivation is when f is subdifferentiable and not differentiable for some elements in $\mathbb{R}^p$.

In the classical optimization settings, there have been developed nearly equivalent results when the convex function to optimize is subdifferentiable instead of differentiable. It would make sense that we could offer analogous results in the averaging settings. In a sense, we aim at crossing this chasm by offering non-asymptotic guarantees in order to approximate averaging estimate with a given accuracy.

This question has been motivated by the challenge of approximating the exponential weighted aggregate with Laplace prior as in Chapter 2 (Dalalyan et al. (2016)) where the function g is the ℓ_1 -norm. Therefore, g is not differentiable for any $\boldsymbol{\beta}$ that has at least one null element. However, g is differentiable almost everywhere.

Even though this situation does not belong to the scope of the aforementioned papers, one would expect the theoretical properties of the Langevin Monte Carlo sampling approximation to hold, up to a few adjustments. Indeed, f is strongly convex (provided conditions on L) and its gradient is almost everywhere defined and Lipschitz-continuous.

In this study, we assume L and g to be convex functions. Besides, we assume L to be differentiable and m-strongly convex while g is assumed to be subdifferentiable. Let μ be the measure

of probability associated with the density π.

The rationale behind this work consists in approximating the measure μ by a well chosen measure μ^s associated with a density measure π_s and with a potential f_s that is differentiable, strongly convex and which gradient is Lipschitz continuous. Using the results of Durmus and Moulines (2016), we obtain an upper bound of the Wasserstein distance between the measure μ^s and the one associated with the simulation by a discretized Langevin Monte Carlo process, as described in Durmus and Moulines (2016) and Dalalyan (2016). If μ^s and μ are such that the Wasserstein distance is small enough, by the triangle inequality we would obtain an upper bound of the distance between μ and the discretized sampling process of μ^s.

The rest of this section provides us with notations and frames the context of our study. In particular we define the Wasserstein distance and the discretized Langevins Monte Carlo process. Section 3.2 gathers and combines results found in the literature in order to prove the quality of the approximation in a finite number of iteration. In Section 3.2, we also mention some useful upper bound tips on the Wasserstein distance. The main results is Corollary 3.2.3 that interprets the impact of the caracteristic of the measure and of the smooth approximation. In Section 3.3, we focus on our initial motivation, namely the challenge of approximating the exponentially weighted aggregate with Laplace prior estimate. We propose a well chosen measure μ^γ which depends on a smoothing parameter $\gamma > 0$. Corollary 3.3.2 provides a practical result to help practitioner to approximate the exponentially weighted aggregate estimate by choosing the parameters and by defining the number of iterations required to achieve a given accuracy.

3.1.1 Notations

Let p be a positive integer, for any $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{u}\|_2$ is the Euclidean norm and $\|\boldsymbol{u}\|_1 = \sum_{i\in[p]} |\beta_j|$ the ℓ_1-norm and, more generally, $\|\boldsymbol{u}\|_k$ refers to the ℓ_k-norm for any positive integer k. We will denote $\mathbf{0_p}$ the null vector of dimension p. The matrix $\mathbf{I}_p$ is the identity matrix of dimension p. The gradient operator of a function v is denoted ∇ and the subgradient set ∂v. We can now set the definitions of strong convexity and gradient Lipschitzness.

Definition 3.1.1 (m-strong convexity)**.** *Let f be a function taking values in $\mathbb{R}^p$, let m be a positive integer, f is m-strongly convex if and only if*

$$f(\boldsymbol{\beta}) - f(\bar{\boldsymbol{\beta}}) + \boldsymbol{u}^\top(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \geq \frac{m}{2}\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2^2, \forall \boldsymbol{u} \in \partial f(\bar{\boldsymbol{\beta}}), \forall \boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \in \mathbb{R}^p. \tag{3.1.4}$$

Definition 3.1.2 (gradient M-Lipschitz)**.** *Let f be a function taking values in $\mathbb{R}^p$, let $M > 0$,*

f is continuously differentiable and has a M-Lipschitz gradient if and only if

$$\|\nabla f(\boldsymbol{\beta}) - \nabla f(\bar{\boldsymbol{\beta}})\|_2 \leq M\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2, \forall \boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \in \mathbb{R}^p. \tag{3.1.5}$$

Let $m > 0$ and $M > 0$, we note $\mathcal{F}_{M,m}$ the set of functions that are m-strongly convex and which gradient is M-Lipschitz. It is interesting to remark that, if a function f belongs to the family $\mathcal{F}_{M,m}$, it implies necessarily that $m \leq M$. This comes from the study of Definition 3.1.1 when $\bar{\boldsymbol{\beta}} \to \boldsymbol{\beta}$.

Let k be a positive integer, we also define $\mathcal{P}_k$ the set of probability measures in $\mathbb{R}^p$ with finite k-moment.

For any set Ω, let us define $\mathcal{B}(\Omega)$ as the Borel set of Ω. For a probability measure ν and Markov kernel Q, we denote νQ the probability measure $\{(\nu Q)A = \int_{\mathbb{R}^p} \nu(\boldsymbol{u})Q(\boldsymbol{u}, A)\mathrm{d}\boldsymbol{u} : A \in \mathcal{B}(\mathbb{R}^p)\}$.

Throughout this work, we use the Wasserstein distance of order 2 as the distance of reference between two measures. This choice is motivated by the strong results of Durmus and Moulines (2016) using Wasserstein distance, instead of Kullback-Leibler as in Dalalyan (2016).

Definition 3.1.3 (Wasserstein distance)**.** *The Wasserstein distance of order 2 between two measures of probability ν and η, $W_2^2(\nu, \eta)$ is defined by*

$$W_2(\nu, \eta) = \inf_{\psi \in \Psi(\nu,\eta)} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 \mathrm{d}\psi(\boldsymbol{u}, \boldsymbol{v}) \right\}^{1/2} \tag{3.1.6}$$

where $\Psi(\nu, \eta)$ is the set of probability measures on $\mathbb{R}^p \times \mathbb{R}^p$ with marginals ν and η.

For clarity purpose, we also remind the definition of the Kullback-Leibler divergence between two probability measures ν and μ.

Definition 3.1.4 (Kullback-Leibler divergence)**.** *Let ν and μ be two probability measures over a set Ω, then if ν is absolutely continuous with respect to μ, the Kullback-Leibler divergence is defined by*

$$KL(\nu\|\mu) = \int_\Omega \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\nu. \tag{3.1.7}$$

Finally, we define the norms

$$\|f\|_{L_2(\pi)} = \int_{\mathbb{R}^p} \big(f(\boldsymbol{u})\big)^2 \pi(\mathrm{d}\boldsymbol{u}), \tag{3.1.8}$$

and

$$\|f\|_\infty = \max_{\mathbb{R}^p} |f(\boldsymbol{u})|. \tag{3.1.9}$$

3.1.2 The Langevin Monte Carlo algorithm

Let f_s be m-strongly convex, continuously differentiable and with a M-Lipschitz continuous gradient. We consider the Langevin stochastic differential equation associated with π_s

$$\mathrm{d}\boldsymbol{\vartheta}_t = -\nabla f_s(\boldsymbol{\vartheta}_t)\mathrm{d}t + \sqrt{2}p\mathbf{b}_t, \tag{3.1.10}$$

where $\mathbf{b}_t$ is a p-dimensional Brownian motion. The process studied in Durmus and Moulines (2016)[1] and Dalalyan (2016) is the Markov chain process based on the Euler-Maruyama discretization

$$\boldsymbol{\upsilon}_{k+1}^{s,h} = \boldsymbol{\upsilon}_k^{s,h} - h\nabla f_s(\boldsymbol{\upsilon}_k^{s,h}) + \sqrt{2h}\boldsymbol{\xi}_k, \tag{3.1.11}$$

where $h > 0$ is the discretization stepsize and $\boldsymbol{\xi}_k$ is a p-dimensional standard Gaussian variable. It is worth remarking that $\boldsymbol{\upsilon}_0^{s,h}$ can be either set arbitrarily in $\mathbb{R}^p$ or be the result of a random distribution. In the case where $\boldsymbol{\upsilon}_0^{s,h}$ is generated by a probability measure ν, we note $\nu P_{h,s}^K$ the measure of the Markov chain defined in Equation 3.1.11.

In this study we focus our attention on the Euler-Maruyama discretization process as described in Equation 3.1.11. Other discretization schemes might be considered and could be subject to further studies.

Section 3.2.1 offers explicit and computable upper bound of the quantity $W_2^2(\nu P_{h,s}^K, \mu^s)$ while Section 3.2.2 guarantees an upper bound of $W_2^2(\mu, \mu^s)$.

Although we will derive results that hold for functions f in the general case as described in Equation 3.1.3, we will consider some specific cases too. For example, in Section 3.3 we will explicitly investigate the computational challenge of the EWA with Laplace prior as described in Chapter 2 (Dalalyan et al. (2016)). In that case,

$$f(\boldsymbol{\beta}) = \frac{1}{\tau}\Big(\frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda\|\boldsymbol{\beta}\|_1\Big), \tag{3.1.12}$$

where $\tau > 0$ is the temperature parameter.

[1] In the paper of Durmus and Moulines (2016), the sampling process considered is more general since stepsize h are not assumed constant; however, in our work, we specify this sampling procedure to the case where h is constant.

3.2 Guarantees for the Wasserstein distance of subdifferentiable potentials

For practical reasons, one may want to sample data from a computable measure $\widehat{\mu}$ that approximates well a given measure μ. In this study, the approximation accuracy is in the sense of the Wasserstein distance; The measure $\widehat{\mu}$ approximates the targeted measure μ with accuracy $\epsilon > 0$ if $W_2(\mu, \widehat{\mu}) < \epsilon$.

Let $m > 0$ and $M > 0$, in this section we show that it is possible to approximate accurately the sampling process of a measure μ as long as there exists a measure μ^s associated with a smooth function $f_s \in \mathcal{F}_{M,m}$ such that $W_2(\mu, \mu^s)$ is smaller than the desired accuracy level ϵ. If such μ^s exists, we propose to approximate the process by applying the Euler-Maruyama discretization of the Langevin Monte Carlo process to μ^s as described in Equation 3.1.11. We named this process $\widehat{\mu}_k^{s,h}$ for any integer k.

The rationale behind this procedure is straightforward, since $f_s \in \mathcal{F}_{M,m}$, the guarantees proved in Dalalyan (2016), Durmus and Moulines (2016) and Dalalyan (2017) offer explicit upper bounds of $W_2(\mu^s, \widehat{\mu}_K^{s,h})$ for a given and explicit number of iterations K. Moreover, if f_s can be chosen such that $W_2(\mu, \mu^s)$ is upper bounded by a quantity smaller than the accuracy level ϵ. Then, using the triangle inequality,

$$W_2(\mu, \widehat{\mu}_K^{s,h}) \leq W_2(\mu, \mu^s) + W_2(\mu^s, \widehat{\mu}_K^{s,h}), \tag{3.2.1}$$

one may choose an explicit number of iterations K such that the Wasserstein distance between μ and $\widehat{\mu}_K^{s,h}$ is smaller than ϵ.

This section states this rationale in Corollaries 3.2.2 and 3.2.3. They are guarantees on the quality of the approximate of the estimation by Langevin Monte Carlo $\widehat{\mu}_K^{s,h}$ when the log-density f is close, in the sense of the Wasserstein distance, to an artefact log-density $f_s \in \mathcal{F}_{M,m}$.

Section 3.2.1 translates the results of Durmus and Moulines (2016) and Dalalyan (2017) in order to offer an upper bound of $W_2(\mu^s, \widehat{\mu}_K^{s,h})$. It is not always simple or even possible to obtain an explicit quantity of $W_2(\mu, \mu^s)$. Thus, we provide in Section 3.2.2 some tools to control the quantity $W_2(\mu, \mu^s)$ with respect to the Kullback-Leibler divergence. These results are mainly due to the very thorough book of Bakry et al. (2014). We seize this opportunity to upper bound the Wasserstein distance with other distances such as $\|f - f_s\|^2_{L_2(\pi)}$ or $\|f - f_s\|_\infty$. This can be useful in practical situations where one meets difficulties to upper bound explicitly the Kullback-Leibler divergence.

3.2.1 Theoretical guarantees from Durmus and Moulines (2016) and Dalalyan (2017)

This section outlines results from Dalalyan (2017) which are closely related to Durmus and Moulines (2016). These results exhibit upper bounds that apply to the last term of the right hand side of Inequality 3.2.1 $W_2(\mu^s, \widehat{\mu}_K^{s,h})$. Indeed these results hold for any log-density $f_s \in \mathcal{F}_{M,m}$ when the sampling of the Langevin Monte Carlo is built on the Euler-Maruyama discretization of the markov chain as described in Equation 3.1.11.

Theorem 3.2.1 is a mere translation of Dalalyan (2017)[Theorem 1] into our notations. Theorem 3.2.1 is very close to Durmus and Moulines (2016)[Theorem 3]. Actually, in the followings, as we will consider constant stepsize of the discretization process, Theorem 3.2.1 is equivalent to Durmus and Moulines (2016)[Corollary 5] that is a consequence of Durmus and Moulines (2016)[Theorem 3].

Theorem 3.2.1 considers the case where the discretization stepsize h is smaller than $2/M$ with M being the gradient Lipschitzness coefficient as defined in Definition 3.1.2. The smoother is the gradient the rougher can be the discretization scheme. Even though one could possibly choose $h = 2/M$ to benefit from theoretical Theorem 3.2.1, it may be interesting to choose smaller stepsize. Indeed, the upper bound has two different regimes depending whether h is smaller than $2/(M+m)$ or not.

Theorem 3.2.1 (Wasserstein upper bound of $f_s \in \mathcal{F}_{M,m}$)**.** *Let $f_s \in \mathcal{F}_{M,m}$ and μ^s the measure of probability associated with f_s. Let $h < \frac{2}{M}$ and $K > 1$, for any probability measure $\nu \in \mathcal{P}_2$, we consider the probability measure $\widehat{\mu}_K^{s,h}$ defined by the probability distribution $\nu P_{s,h}^K$, where $P_{s,h}^K$ is the discretized process diffusion approximation described in Equation 3.1.11. Then, if $h \leq 2/(M+m)$,*

$$W_2(\widehat{\mu}_K^{s,h}, \mu^s) \leq (1-mh)^K W_2(\nu, \mu^s) + 1.82\frac{M}{m}(hp)^{1/2}. \tag{3.2.2}$$

Alternatively, if $h \geq 2/(M+m)$,

$$W_2(\widehat{\mu}_K^{s,h}, \mu^s) \leq (Mh-1)^K W_2(\nu, \mu^s) + 1.82\frac{Mh}{2-Mh}(hp)^{1/2}. \tag{3.2.3}$$

It is worth commenting some elements of this result. The right hand side of Inequalities 3.2.2 and 3.2.3 are both composed of two elements. The first one decreases with the number of iterations K and depends on the initial choice of the measure $\nu \in \mathcal{P}_2$. We will name this

quantity the decreasing part. Respectively, we will name the constant part the second term in the right hand side of these inequalities. Let us first consider the case when one has chosen a stepsize h smaller than $2/(M+m)$. First of all it is easy to show that

$$\lim_{K\to+\infty}(1-mh)^K = 0. \tag{3.2.4}$$

Indeed, since $0 < h \leq 2/(M+m) \leq 1/m$, it implies that $0 \geq (1-mh) < 1$. Moreover, it implies that the decreasing part of the upper bound is decreasing with h. On the other hand the constant part

$$1.82\frac{M}{m}(hp)^{1/2} \tag{3.2.5}$$

increases with M and p and decreases with m and h. Therefore one faces a tradeoff. A small value of h minimizes the constant part while a value of h closer to $2/(M+m)$ minimizes the decreasing part. In practical situation, for a given targeted accuracy, one may not have so much choice. It is clear that the constant part has to be controlled with a small value of h and the number of iterations K will be chosen to counterbalance the small value of h.

In the case where $h \geq 2/(M+m)$, the interpretation is different. Even though

$$\lim_{K\to+\infty}(Mh-1)^K = 0 \tag{3.2.6}$$

increases with h, contrary to the quantity defined in Equation 3.2.4. The constant part

$$1.82\frac{Mh}{2-Mh}(hp)^{1/2} \tag{3.2.7}$$

increases with h too. In the case where $h \geq 2/(M+m)$, we recommend to choose a small h close to $2/(M+m)$.

Theorem 3.2.1 is key to our study. However, as mentioned in Dalalyan (2017), it does not provide explicit guarantees since it depends on the term $W_2(\nu, \mu^s)$, while ν has not been specified, and on W_2, which may often be difficult to compute for any $\nu \in \mathcal{P}_2$.

In order to offer an explicit upper bound as in Corollary 3.2.1, we remark that Theorem 3.2.1 holds for any $\nu \in \mathcal{P}_2$. In particular, it is true for any Dirac measure. It makes the consideration of deterministic initialization of the Langevin Monte Carlo process possible. Let us consider an arbitrarily chosen vector $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$, let us set the measure ν to be a Dirac in $\boldsymbol{\beta}^{(0)}$, $\nu = \delta_{\boldsymbol{\beta}^{(0)}}$. Let $\bar{\boldsymbol{\beta}}_s$ be the average with respect to the density π_s. Then we can propose a rough but explicit

upper bound of $W_2^2(\delta_{\boldsymbol{\beta}^{(0)}}, \mu^s)$ using the exact same arguments as in Dalalyan (2017). Indeed,

$$W_2^2(\delta_{\boldsymbol{\beta}^{(0)}}, \mu^s) \leq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2^2 + \int_{\mathbb{R}^p} \|\bar{\boldsymbol{\beta}}_s - \boldsymbol{\beta}\|_2^2 \pi_s(\mathrm{d}\boldsymbol{\beta}). \tag{3.2.8}$$

Therefore,

$$W_2^2(\delta_{\boldsymbol{\beta}^{(0)}}, \mu^s) \leq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2^2 + \frac{p}{m}. \tag{3.2.9}$$

This leads to the conclusion that,

$$W_2(\delta_{\boldsymbol{\beta}^{(0)}}, \mu^s) \leq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + \left(\frac{p}{m}\right)^{1/2}. \tag{3.2.10}$$

In a practical context, the use of a deterministic initialization enables, through Inequality 3.2.10, an explicit upper bound of $W_2(\widehat{\mu}_K^{s,h}, \mu^s)$. Of course, it requires $\boldsymbol{\beta}^{(0)}$ to be chosen. If the use of a deterministic initialization is useful for the proof of the existence of an explicit upper bound of $W_2(\widehat{\mu}_K^{s,h}, \mu^s)$, the choice of the initialization is of paramount importance in a practical context. Indeed, a not suited choice of $\boldsymbol{\beta}^{(0)}$ deteriors strongly the approximation accuracy. The following corollary is a consequence of a deterministic initialization, it provides an explicit upper bound and exhibits the importance of the choice of $\boldsymbol{\beta}^{(0)}$. One more time, Corollary 3.2.1 is a mere translation of the results of Dalalyan (2017).

Corollary 3.2.1. *Let $f_s \in \mathcal{F}_{M,m}$ and μ^s the measure of probability associated with f_s. Let $h < \frac{2}{M}$ and $K > 1$, for any probability measure $\nu \in \mathcal{P}_2$, we consider the probability measure $\widehat{\mu}_K^{s,h}$ defined by the probability distribution $\nu P_{s,h}^K$, where $P_{s,h}^K$ is the discretized process diffusion approximation described in Equation 3.1.11. Then, if $h \leq 2/(M+m)$,*

$$W_2(\widehat{\mu}_K^{s,h}, \mu^s) \leq (1-mh)^K \left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + \left(\frac{p}{m}\right)^{1/2} \right) + 1.82 \frac{M}{m} (hp)^{1/2}.$$

Alternatively, if $h \geq 2/(M+m)$,

$$W_2(\widehat{\mu}_K^{s,h}, \mu^s) \leq (Mh-1)^K \left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + \left(\frac{p}{m}\right)^{1/2} \right) + 1.82 \frac{Mh}{2-Mh} (hp)^{1/2}.$$

Corollary 3.2.1 will be used in Section 3.2.3. In the meanwhile, we study results that enable to control the distance between the smooth measure and the non-smooth targeted measure.

3.2.2 Bounding the Wasserstein distance between approximation and the target distribution

In Section 3.2.1, we remind some explicit upper bound of the distance $W_2(\widehat{\mu}_K^{s,h}, \mu^s)$ for a given stepsize h and an explicit number of iterations K, when $f_s \in \mathcal{F}_{M,m}$. Now, we study some inequalities that are useful to bound the second term of Inequality 3.2.1, $W_2(\mu, \mu^s)$. Close forms of Wasserstein distance between two measures are not always known, this section provides the practitioners with some tools to explicitly upper bound $W_2(\mu, \mu^s)$. The main result of this section is Proposition 3.2.1. It is a direct application of some results from the optimal transport theory applied in the trivial case that is $\mathbb{R}^p$. It states that the m-strong convexity of the smoothed potential f_s guarantees an upper bound of the Wasserstein distance between μ^s and the measure of probability of interest μ. It is a method to upper bound the distance $W_2(\mu, \mu^s)$ by a quantity that is related to the Kullback-Leibler divergence $KL(\mu\|\mu^s)$. This result is implied by Bakry et al. (2014)[Corollary 9.2.2] which is a specific case of the more general result Bakry et al. (2014)[Theorem 9.3.1]. These results can be interpreted as generalizations of the Talagrand inequality theorem as described in Bakry et al. (2014)[Theorem 9.2.1]. The goal of this section is to provide methodologies that can help upper bounding the quantity $W_2(\mu, \mu^s)$ in practical situations. Therefore, we mention complementary results that could offer guarantees on the upper bound of the Kullback-Leibler divergence, and consequently of the Wasserstein distance. So are the goals of Proposition 3.2.2 and Remark 3.2.1.
In order to understand the result of Proposition 3.2.1, we first introduce the definition of the quadratic transportation cost inequality $\mathcal{T}(\rho)$ for any positive real ρ as in Otto and Villani (2000)[Definition 2] and to Bakry et al. (2014)[Definition 9.2.2].

Definition 3.2.1 (Quadratic transport cost inequality)**.** *Let η be a probability measure we say that the quadratic transportation cost inequality $\mathcal{T}(\rho)$ holds for η (i.e. $\eta \in \mathcal{T}(\rho)$) if for any measure $\nu \in \mathcal{P}_2$, absolutely continuous with respect to η ,*

$$W_2^2(\eta, \nu) \leq \frac{2KL(\nu\|\eta)}{\rho}. \tag{3.2.11}$$

This definition states that Inequality 3.2.11 must hold for any $\nu \in \mathcal{P}_2$. Since we consider target measure μ in $\mathcal{P}_2$, if one is able to choose a measure μ^s in $\mathcal{T}(\rho)$ for a given $\rho > 0$, then the inequality

$$W_2^2(\mu, \mu^s) \leq \frac{2KL(\mu\|\mu^s)}{\rho} \tag{3.2.12}$$

would hold.

Proposition 3.2.1 gives conditions that are sufficient to guarantee that a measure μ^s belongs to the set $\mathcal{T}(\rho)$ for an explicit value of ρ as long as there exists an associated potential f_s that is twice differentiable and strongly convex. The parameter ρ will be defined by the coefficient of strong-convexity of f_s.

Proposition 3.2.1. *Let $m > 0$, and f_s the potential of π_s be twice differentiable and m-strongly convex, then $\mu^s \in \mathcal{T}(m)$.*

It implies that

$$W_2^2(\mu, \mu^s) \leq \frac{2KL(\mu\|\mu^s)}{m}. \tag{3.2.13}$$

This result can be deduced from Bakry et al. (2014)[Corollary 9.3.2] that states, with our notations, that if $\mathrm{d}\mu^s = \exp(-f_s)\mathrm{d}\boldsymbol{\beta}$ is a probability measure of $\mathbb{R}^p$ where f_s is a smooth function such that $\nabla^2(f_s) \succcurlyeq \rho \mathbf{I}_p$ for some $\rho > 0$ then μ^s satisfies the quadratic transportation cost of Inequality 3.2.11 (i.e. $\mu^s \in \mathcal{T}(\rho)$). Since $\mu^s \in \mathcal{T}(\rho)$, for any $\nu \in \mathcal{P}_2$,

$$W_2^2(\nu, \mu^s) \leq \frac{2KL(\nu\|\mu^s)}{\rho}. \tag{3.2.14}$$

In particular, $\mu \in \mathcal{P}_2$ and $f_s \in \mathcal{F}_{M,m}$ implies that $\nabla^2(f_s) \succcurlyeq m\mathbf{I}_p$, therefore

$$W_2^2(\mu, \mu^s) \leq \frac{2KL(\mu\|\mu^s)}{m}, \tag{3.2.15}$$

which concludes the proof of Corollary 3.2.1. □

Using Proposition 3.2.1 directly offers a solution to upper bound the second term of Inequality 3.2.1 as long as there exists $m > 0$ such that f_s is twice differentiable and m-strongly convex. Indeed, if so,

$$W_2(\mu, \mu^s) \leq \left(\frac{2KL(\mu\|\mu^s)}{m}\right)^{1/2}. \tag{3.2.16}$$

The result of Dalalyan (2016)[Lemma 3] upper bounds the Kullback-Leibler metric between two measures with some assumptions that are reasonable in the context of Section 3.3. This is the reason why Proposition 3.2.1 is interesting for our purpose. The result Dalalyan (2016)[Lemma 3] is described in our notations in Proposition 3.2.2.

Proposition 3.2.2 (Dalalyan (2016), Lemma 3)**.** *Let f_s and f be some potentials respectively associated with the measures μ^s and μ such that $f(\boldsymbol{\beta}) \leq f_s(\boldsymbol{\beta})$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$. If $\exp(-f)$ and $\exp(-f_s)$ are both integrable, then*

$$KL(\mu\|\mu^s) \leq \frac{1}{2}\|f - f_s\|^2_{L_2(\pi)}, \tag{3.2.17}$$

where π is the density associated with μ and f.

Remark 3.2.1. *From Inequality 3.2.17, it is trivial to remark that, with similar assumptions than the ones necessary to Proposition 3.2.2,*

$$KL(\mu\|\mu^s) \leq \frac{1}{2}\|f - f_s\|_\infty^2 = \max_{\boldsymbol{\beta}\in\mathbb{R}^p}\{|f(\boldsymbol{\beta}) - f_s(\boldsymbol{\beta})|\}^2. \tag{3.2.18}$$

Indeed,

$$\|f - f_s\|_{L_2(\pi)}^2 = \int_{\mathbb{R}^p} (f(\boldsymbol{u}) - f_s(\boldsymbol{u}))^2 \mathrm{d}\pi(\boldsymbol{u}), \tag{3.2.19}$$

therefore, π being a density, by construction, it integrates to one. It implies that,

$$\int_{\mathbb{R}^p} (f(\boldsymbol{u}) - f_s(\boldsymbol{u}))^2 \mathrm{d}\pi(\boldsymbol{u}) \leq \max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\left(f(\boldsymbol{\beta}) - f_s(\boldsymbol{\beta})\right)^2\right\}. \tag{3.2.20}$$

Combining Inequalities 3.2.17 and 3.2.20, we obtain Inequality 3.2.18.

In view of Definition 3.2.1, Proposition 3.2.2 and Remark 3.2.1, some conditions in the choice of the smooth potential f_s are now known to guarantee an upper bound of the Wasserstein distance $W_2(\mu, \mu^s)$.

If μ^s is such that one knows an upper bound of $KL(\mu\|\mu^s)$ then it is enough to show that there exists $\rho > 0$ such that $\mu^s \in \mathcal{T}(\rho)$. In particular, Proposition 3.2.2 states that if μ^s is twice differentiable and m-strongly convex, then $\mu^s \in \mathcal{T}(m)$. In the case one does not know an upper bound of the divergence $KL(\mu\|\mu^s)$, then Proposition 3.2.2 and Remark 3.2.1 provide one with some methodologies to upper bound the Wasserstein distance $W_2(\mu, \mu^s)$. If $f(\boldsymbol{\beta}) \leq f_s(\boldsymbol{\beta})$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, then,

$$W_2(\mu, \mu^s) \leq \left(\frac{2KL(\mu\|\mu^s)}{m}\right)^{1/2} \leq \frac{\|f - f_s\|_{L_2(\pi)}}{m^{1/2}} \leq \frac{\|f - f_s\|_\infty}{m^{1/2}}. \tag{3.2.21}$$

Even though these inequalities are very rough, they may be useful if one aims at obtaining explicit guarantees. Moreover, it is worth noting that these inequalities provide guidelines to choose μ^s and f_s. Indeed, the Wasserstein distance $W_2(\mu, \mu^s)$ decreases with the strong-convexity coefficient m of the potential f_s and increases with the divergence between μ and μ^s (in the sense of Kullback Leibler) or similarly with the distance between f and f_s.

Section 3.2.1 gives explicit upper bound of the Wasserstein distance $W_2(\widehat{\mu}_K^{s,h}, \mu^s)$, while this section defines conditions for which $W_2(\mu, \mu^s)$ is controlled. In the next section, we combine these results to prove, under a given set of assumptions, the existence of an explicit upper bound with a finite number of iterations K of $W_2(\widehat{\mu}_K^{s,h}, \mu^s)$.

3.2.3 Conclusion on theoretical guarantees in finite sampling

It is now possible to consider the case of interest by using the triangle inequality applied to previous results,

$$W_2(\mu, \widehat{\mu}_K^{s,h}) \leq W_2(\mu, \mu^s) + W_2(\mu^s, \widehat{\mu}_K^{s,h}). \tag{3.2.22}$$

For the sake of completeness, the authors Clement and Desch (2008) offer a complete proof of the triangle inequality property of the Wasserstein distance.

This being taken into consideration, we deduce a corollary from Theorem 3.2.1 and Corollary 3.2.1.

Corollary 3.2.2. *Let f_s be a smooth potential that approximates f such that $f_s \in \mathcal{F}_{M,m}$, and let μ^s be the measure of probability associated with f_s. Let $0 < h < 2/M$ and $K > 1$, for any probability measure $\nu \in \mathcal{P}_2$, we consider the probability measure $\widehat{\mu}_K^{s,h}$ defined by the probability distribution $\nu P_{s,h}^T$, where $P_{s,h}^K$ is the discretized process diffusion approximation described in Equation 3.1.11.*

Then, if $h \leq 2/(M+m)$,

$$W_2(\widehat{\mu}_K^{s,h}, \mu) \leq (1-mh)^K W_2(\nu, \mu^s) + 1.82\frac{M}{m}(hp)^{1/2} + \left(\frac{2KL(\mu\|\mu^s)}{m}\right)^{1/2}. \tag{3.2.23}$$

Alternatively, if $h \geq 2/(M+m)$,

$$W_2(\widehat{\mu}_K^{s,h}, \mu) \leq (Mh-1)^K W_2(\nu, \mu^s) + 1.82\frac{Mh}{2-Mh}(hp)^{1/2} + \left(\frac{2KL(\mu\|\mu^s)}{m}\right)^{1/2}. \tag{3.2.24}$$

Furthermore, it is possible to substitute the terms $W_2(\nu, \mu^s)$ and $KL(\mu\|\mu^s)^{1/2}$ in Equations 3.2.23 and 3.2.24 by more explicit quantities. If the substitution of $W_2(\nu, \mu^s)$ does not require additional assumptions, the upper bound of $KL(\mu\|\mu^s)^{1/2}$ by $(\|f - f_s\|_{L_2(\pi)})/2$ or $(\|f - f_s\|_\infty)/2$ requires so. Even though the substitution with $(\|f - f_s\|_\infty)$ is the roughest we represent Corollary 3.2.3 with this substitution, because it is often simpler to get a close form of $(\|f - f_s\|_\infty)$ than of $(\|f - f_s\|_{L_2(\pi)})$. Of course one could use the same corollary with the term $(\|f - f_s\|_{L_2(\pi)})$ instead of $(\|f - f_s\|_\infty)$.

Corollary 3.2.3. *Let $f_s \in \mathcal{F}_{M,m}$ and μ^s the measure of probability associated with f_s. Let $h < \frac{2}{M}$ and $K > 1$, for any probability measure $\nu \in \mathcal{P}_2$, we consider the probability measure $\widehat{\mu}_K^{s,h}$ defined by the probability distribution $\nu P_{s,h}^T$, where $P_{s,h}^K$ is the discretized process diffusion approximation described in Equation 3.1.11. Moreover, if f_s is twice differentiable and $f_s(\boldsymbol{\beta}) \geq$*

$f(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, then if $h \leq 2/(M+m)$,

$$W_2(\widehat{\mu}_K^{s,h}, \mu) \leq (1-mh)^K \left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + \left(\frac{p}{m}\right)^{1/2} \right) + 1.82 \frac{M}{m} (hp)^{1/2} + \frac{\|f - f_s\|_\infty}{m^{1/2}}.$$

Alternatively, if $h \geq 2/(M+m)$,

$$\begin{aligned} W_2(\widehat{\mu}_K^{s,h}, \mu) \leq & (Mh-1)^K \left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + \left(\frac{p}{m}\right)^{1/2} \right) + 1.82 \frac{Mh}{2-Mh} (hp)^{1/2} \\ & + \frac{\|f - f_s\|_\infty}{m^{1/2}}. \end{aligned}$$

This corollary is the combination of known results from the literature such as the one mentioned in Dalalyan (2017). However it is new in the sense that it provides an explicit guarantee of the quality of the sampling approximation of the measure μ, even though f does not belong to the set $\mathcal{F}_{M,m}$. Clearly, some questions remain since $\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2$ is not necessarily known, neither is $KL(\mu\|\mu^s)$ or $\|f - f_s\|_\infty$. However, these quantities are easier to measure, or at least to approximate than the quantity $W_2(\nu, \mu^s)$.

This result describes some explicit conditions to guarantee the existence and the reach of a given accuracy ϵ. Of course, the accuracy strongly relies on the choice of f_s. In the article Brosse et al. (2017), a general approach of smoothing using the proximal operator is defined. In order to optimize the solution, it would be interesting, for a given family $\mathcal{F}_{M,m}$ to minimize the problem

$$\arg \min_{g \in \mathcal{F}_{M,m}} W_2(\mu, \nu_g), \tag{3.2.25}$$

where ν_g is the measure associated with g. If solving this problem was feasible for any (m, M), then it would be possible to choose optimally the best sample approximation. The exact solution if this problem is a very difficult challenge. However, there is a literature on efficient computational methods for the approximation of the Wasserstein distance as in Solomon et al. (2015) that could be of help in order to solve this problem.

In the next section, we will consider an example of application of Corollary 3.3.2, the exponential weighted aggregate with Laplace prior estimation.

3.3 The case of EWA with Laplace prior approximation

In this section, we consider the exponentially weighted aggregate with Laplace prior estimate $\widehat{\boldsymbol{\beta}}_{EWA}$ and the model described in Chapter 2 (Dalalyan et al. (2016)). This corresponds to

considering data that consist of n random observations $y_1, \ldots, y_n \in \mathbb{R}$ and p fixed covariates $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p \in \mathbb{R}^n$. We further assume that there is a vector $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that the residuals $\xi_i = y_i - \beta_1^\star \boldsymbol{x}_i^1 - \ldots - \beta_p^\star \boldsymbol{x}_i^p$ are independent, zero mean random variables. In vector notation, this reads as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi}, \tag{3.3.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the response vector, $\mathbf{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{\xi}$ is the noise vector.

In this section, we assume the Gram matrix to be invertible; in particular, we assume that its smallest eigen value $\widehat{\sigma}_{\min}$ is positive. It is equivalent to

$$\mathbf{X}^\top \mathbf{X}/n \succcurlyeq \widehat{\sigma}_{\min} \mathbf{I}_p. \tag{3.3.2}$$

We define the fitting function L by

$$L(\boldsymbol{\beta}) = \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n}, \tag{3.3.3}$$

and the penalization term g by

$$g(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1. \tag{3.3.4}$$

In that context, the potential of the pseudo-posterior function is

$$f(\boldsymbol{\beta}) = \frac{1}{\tau} L(\boldsymbol{\beta}) + \frac{\lambda}{\tau} g(\boldsymbol{\beta}). \tag{3.3.5}$$

The function f corresponds to the potential associated with the pseudo posterior of the EWA with Laplace prior as defined in Chapter 2. The exponentially weighted aggregate with Laplace prior estimate is defined, for a given temperature τ, by Equation 3.1.2. As shown in Chapter 2, the exponentially weighted aggregate estimate $\widehat{\boldsymbol{\beta}}_{EWA}$ enjoys a fast rate property in the sense of the prediction error:

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) = \|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2, \tag{3.3.6}$$

for a given temperature parameter τ and a specific value of the penalty term λ. It is therefore a question of interest to determine a method that guarantees a feasible and practical method to approximate with a desired accuracy the EWA estimate $\widehat{\boldsymbol{\beta}}_{EWA}$. However, to the best of our knowledge, there is no guarantee in the literature that a given method could approximate the estimate $\widehat{\boldsymbol{\beta}}_{EWA}$ in a polynomial finite computational time for a fixed accuracy ϵ.

In this section, we first provide a smooth approximation f_γ of f such that $f_\gamma \in \mathcal{F}_{M,m}$ and such that an explicit upper bound of $W_2(\mu, \mu^\gamma)$ is guaranteed. We then show in Theorem 3.3.1 that this approximation enables to apply Theorem 3.2.1 and therefore to show that, with a given approximation of f, the sampling method can achieve accurate approximation in finite time. Finally, we provide in Corollary 3.3.2 practical requirements on the number of iterations K as well as of the stepsize h so that a desired accuracy ϵ is achieved.

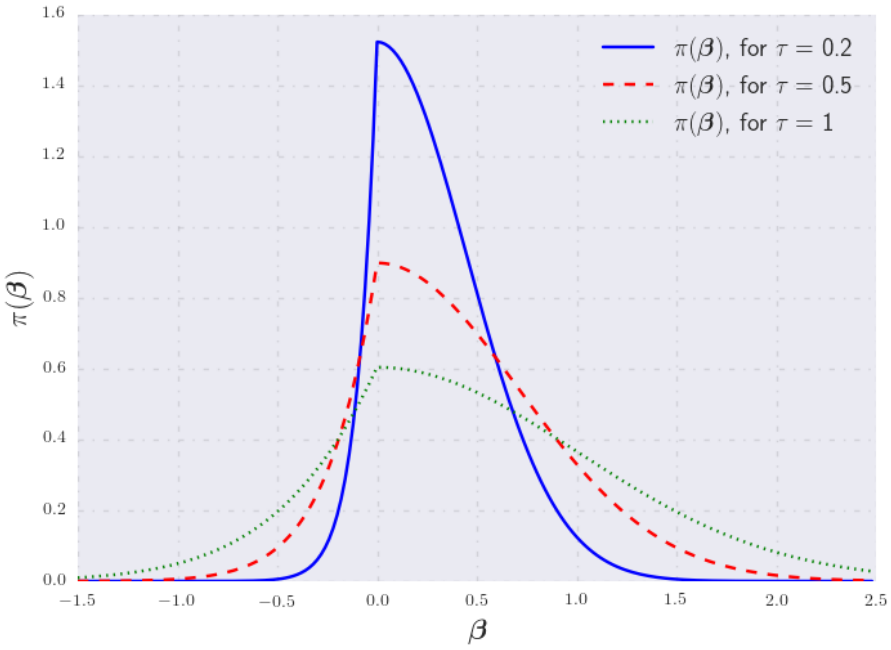


Figure 3-1: This figure illustrates the density function π associated with the log-density $f(\boldsymbol{\beta})$ as defined in Equation 3.3.5 for different values of the temperature parameter τ in the noise-free and unidimensional settings where $y = \boldsymbol{\beta}^\star = 1$ and $\lambda = 1$. The solid blue line represents π with $\tau = 1$ which is equivalent to the bayesian Lasso. The red dashed line is an illustration of the density π when $\tau = 0.5$. Finally, the dotted green line represents a smaller temperature $\tau = 0.2$. When τ is a very small quantity, the density π is converging, in the sense of distributions, into a Dirac concentrated at the lasso estimate.

The function f is $(\widehat{\sigma}_{\min}/\tau)$-strongly convex from Equation 3.3.2. However, $f \notin \mathcal{F}_{M,m}$, since g is not differentiable but only subdifferentiable. This is why direct applications of the results from Durmus and Moulines (2016) or from Dalalyan (2016) do not apply here. To tackle this challenge, we investigate the properties of g_γ, a smooth approximation of g:

$$g_\gamma(\boldsymbol{\beta}) = \sum_{j \in [p]} \sqrt{\gamma^2 + \beta_j^2}, \tag{3.3.7}$$

where $\gamma \geq 0$ is a parameter.

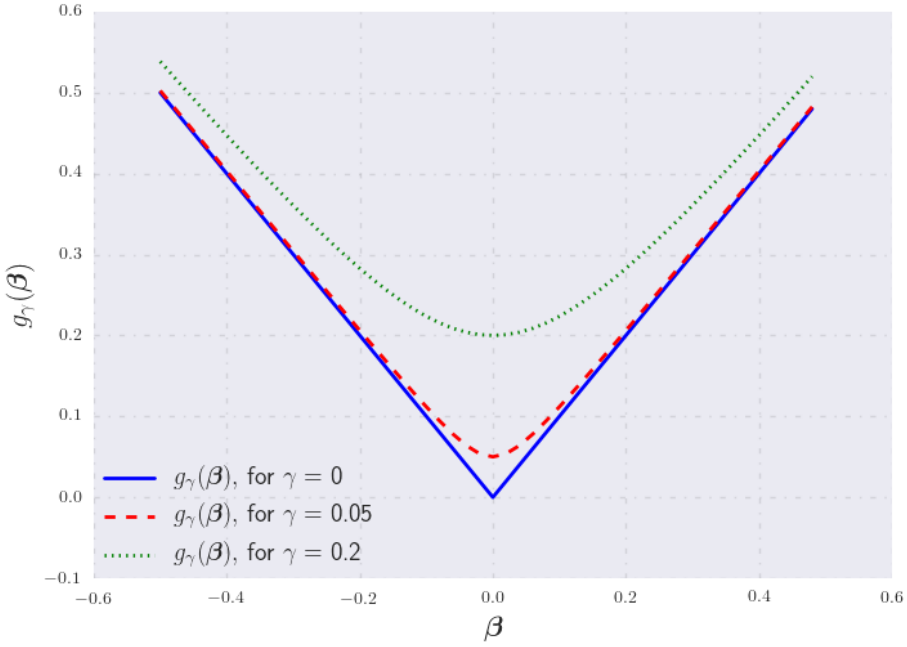


Figure 3-2: This figure represents the function g_γ as described in Equation 3.3.7 for different values of γ in the unidimensional settings ($\boldsymbol{\beta} \in \mathbb{R}$). The solid blue line is g_0, which is also the ℓ_1-norm. The red dashed line is $g_{0.05}$ and the green dotted line is $g_{0.2}$. The greater the value of γ, the smoother is the penalty function g_γ.

We can now define the smooth approximation f_γ of the potential f defined as

$$f_\gamma(\boldsymbol{\beta}) = \frac{1}{\tau} L(\boldsymbol{\beta}) + \frac{\lambda}{\tau} g_\gamma(\boldsymbol{\beta}). \tag{3.3.8}$$

The goal of this section is to investigate the quality of a Langevin Monte Carlo procedure applied to f_γ in order to sample the distribution characterized by the potential f.

In this section, we derivate and show some properties of f_γ that are useful to obtain explicit guarantees of the sampling process. In particular, we show that f_γ has all the necessary properties so that Corollary 3.2.2 can be applied.

We define $\widehat{\sigma}_{\min}$ (respectively $\widehat{\sigma}_{\max}$) as the smallest (resp. the largest) eigenvalue of the Gram matrix $\mathbf{X}^\top \mathbf{X}/n$. In the following, we will assume that the Gram matrix is invertible and that $\widehat{\sigma}_{\min} > 0$. These assumptions exclude some situations that are potentially important. In particular, the high-dimensional settings, where $p > n$, will not belong to the frame of our study. It would be of great interest to develop a method that could guarantee approximation accuracies in finite time with less restrictive assumptions. Indeed, the penalization methods

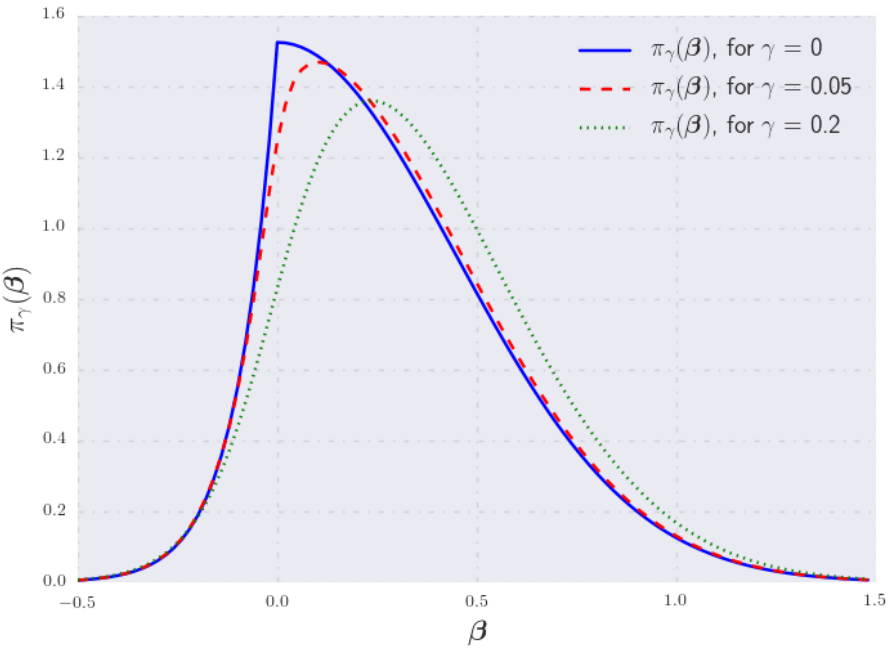


Figure 3-3: This plot represents the density function π_γ associated with the smooth log-density $f_\gamma(\boldsymbol{\beta})$ as defined in Equation 3.3.8 for a given temperature parameter $\tau = 0.2$. This plot considers the noise-free and unidimensional settings where $\boldsymbol{\beta}^\star = 1$, $y = \boldsymbol{\beta}^\star$ and $\lambda = 1$. The solid blue line is π_0, the non-smooth case, which is the standard EWA estimate. The red dash line is $\pi_{0.05}$ and the green dotted line is $\pi_{0.2}$.

such as the EWA with Laplace prior are efficient in the case of high-dimensional settings. However, this will not be the focus of this section, nor of this study.

Proposition 3.3.1 explicits constants M_γ, m_γ depending on the matrix $\mathbf{X}$ and on the parameter γ so that $f_\gamma \in \mathcal{F}_{M_\gamma, m_\gamma}$. Moreover, these constants are proved to be optimal in the sense that there is neither $m > m_\gamma$ such that f_γ is m-strongly convex, nor any $M < M_\gamma$ such that the gradient ∇f_γ is M-Lipschitz.

Proposition 3.3.1. *Let us consider a design matrix* $\mathbf{X}$ *such that the associated Gram matrix* $\mathbf{X}^\top\mathbf{X}/n$ *is invertible with its smallest eigen value* $\widehat{\sigma}_{\min}$ *positive and its greatest eigen value* $\widehat{\sigma}_{\max}$ *known. Let* γ *be a positive real number and* f_γ *be the smooth negative log-density defined in Equation 3.3.8 with respect to the smooth penalty* g_γ *defined in Equation 3.3.7. Then,* f_γ *belongs to the set* $\mathcal{F}_{M_\gamma, m_\gamma}$ *with*

$$M_\gamma = \frac{1}{\tau}\big(\widehat{\sigma}_{\max} + \frac{\lambda}{\gamma}\big), \tag{3.3.9}$$

and

$$m_\gamma = \frac{\widehat{\sigma}_{\min}}{\tau}. \tag{3.3.10}$$

Moreover, the constants (m_γ, M_γ) are optimal in the sense that there is no couple (m, M) with $m < m_\gamma$ or $M > M_\gamma$ such that $f_\gamma \in \mathcal{F}_{M,m}$.

The proof of Proposition 3.3.1 is postponed to Section 3.5. First of all, Proposition 3.3.1 mentions that the strong convexity property of f_γ is not affected by the parameter γ. Indeed, considering Equation 3.3.8, only g_γ varies with γ. As illustrated in Figure 3-2, g_γ is not strongly-convex but only strictly-convex as long as $\gamma > 0$. Therefore, no theoretical improvement is expected to come from an increase of the strong-convexity parameter $m = m_\gamma$. However, referring to Figure 3-2, g_γ is strongly-convex on any bounded set included in $\mathbb{R}^p$ as long as $\gamma > 0$. Consequently, some empirical benefits of the strong-convexity are expected with the increase of γ. On the other hand, γ has a significant impact on the Lipschitzness property of the gradient. Indeed, if $\gamma = 0$, f_γ is not differentiable since the ℓ_1-norm is not differentiable. By smoothing g_γ with the increase of γ, we observe theoretical improvement of the Lipschitzness of the gradient ∇f_γ. As a consequence, the increase of γ will theoretically improve the property of f_γ in the sense of the Lipschitzness of the gradient. It will therefore improve the accuracy of the approximation $\widehat{\mu}_K^{\gamma,h}$. However, if the term $W_2(\widehat{\mu}_K^{\gamma,h}, \mu^\gamma)$ will decrease with γ, the quality of the approximation μ^γ of the targeted measure μ will suffer from a rougher approximation of the ℓ_1-norm. This result is explicited by Proposition 3.3.2.

Proposition 3.3.2. *Let μ be the measure of probability associated with the potential f defined in Equation 3.3.5 and μ^γ the measure associated with f_γ defined in Equation 3.3.8, then*

$$W_2(\mu, \mu^\gamma) \leq \left(\frac{\widehat{\sigma}_{\min}}{\tau}\right)^{1/2} p\lambda\gamma. \tag{3.3.11}$$

The proof of Proposition 3.3.2 is postponed to Section 3.5. The proof relies on Equation 3.2.18 in Remark 3.2.1. This result provides a rough upper bound of $W_2(\mu, \mu^\gamma)$ and shows the impact of γ on the approximation quality. The next theorem combines Propositions 3.3.1 and 3.3.2 to explicit the rate of convergence of the approximation $\widehat{\mu}_K^{\gamma,h}$ with respect to μ. This result enables to state the result that motivates this work, namely Theorem 3.3.1.

Theorem 3.3.1. *Let μ be the probability measure of the pseudo-posterior of the EWA with Laplace prior, with temperature $\tau > 0$, as described by the potential f in Equation 3.3.5, and let μ^γ be the measure probability associated to f_γ as in Equation 3.3.8. Let $0 < h < \frac{2}{M}$ and $K > 1$, for any probability measure $\nu \in \mathcal{P}_2$, we consider the probability measure $\widehat{\mu}_{\gamma,h}^K$ defined by the probability distribution $\nu P_{\gamma,h}^T$, where $P_{\gamma,h}^K$ is the discretized process diffusion approximation of the measure μ^γ, as described in Equation 3.1.11.*

If $h \leq 2/(M_\gamma + m_\gamma)$,

$$W_2(\widehat{\mu}^K_{\gamma,h}, \mu) \leq (1 - m_\gamma h)^K W_2(\nu, \mu^\gamma) + 1.82 \frac{M_\gamma}{m_\gamma} (hp)^{1/2} + \frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}}. \tag{3.3.12}$$

Alternatively, if $h \geq 2/(M_\gamma + m_\gamma)$,

$$W_2(\widehat{\mu}^K_{\gamma,h}, \mu) \leq (M_\gamma h - 1)^K W_2(\nu, \mu^\gamma) + 1.82 \frac{M_\gamma h}{2 - M_\gamma h} (hp)^{1/2} + \frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}}. \tag{3.3.13}$$

Theorem 3.3.1 is the direct application of Theorem 3.2.1 in the context of the exponentially weighted aggregate with Laplace prior estimate. Of course, one may not know the quantity $W_2(\nu, \mu^\gamma)$. Corollary 3.3.1 offers an upper bound that turns out to be explicit.

Corollary 3.3.1. *Let* μ *be the probability measure of the pseudo-posterior of the EWA with Laplace prior with temperature* $\tau > 0$*, as described by the potential* f *in Equation 3.3.5, and let* μ^γ *be the measure probability associated to* f_γ *in Equation 3.3.8. Let* $0 < h < \frac{2}{M}$ *and* $K > 1$*, for any probability measure* $\nu \in \mathcal{P}_2$*, we consider the probability measure* $\widehat{\mu}^K_{\gamma,h}$ *defined by the probability distribution* $\nu P^T_{\gamma,h}$*, where* $P^K_{\gamma,h}$ *is the discretized process diffusion approximation of the measure* μ^γ *as described in Equation 3.1.11.*

If $h \leq 2/(M_\gamma + m_\gamma)$,

$$\begin{aligned} W_2(\widehat{\mu}^K_{\gamma,h}, \mu) \leq & (1 - m_\gamma h)^K \left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2 + \left(\frac{p}{m_\gamma}\right)^{1/2} \right) + 1.82 \frac{M_\gamma}{m_\gamma} (hp)^{1/2} \\ & + \frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}}. \end{aligned}$$

Alternatively, if $h \geq 2/(M_\gamma + m_\gamma)$,

$$\begin{aligned} W_2(\widehat{\mu}^K_{\gamma,h}, \mu) \leq & (M_\gamma h - 1)^K \left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2 + \left(\frac{p}{m_\gamma}\right)^{1/2} \right) + 1.82 \frac{M_\gamma h}{2 - M_\gamma h} (hp)^{1/2} \\ & + \frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}}. \end{aligned}$$

The proof of this corollary is straightforward in view of Corollary 3.2.2 applied to f_γ using the fact from Proposition 3.3.1 that $f_\gamma \in \mathcal{F}_{M_\gamma, m_\gamma}$. □

In Chapter 2 (Dalalyan et al. (2016)), sharp oracles inequalities of the prediction error and of the pseudo-posterior concentration are proven to hold when τ is of the order σ^2/np and λ of the order $\sigma\{\log(p)/n\}^{1/2}$. In the following, we study explicitly the rate of convergence of

$W_2(\widehat{\mu}^K_{\gamma,h}, \mu)$ when

$$\tau = \frac{\sigma^2}{np}, \tag{3.3.14}$$

and

$$\lambda = \sigma \left\{ \frac{\log(p)}{n} \right\}^{1/2}. \tag{3.3.15}$$

We assume $\widehat{\sigma}_{\min} > 0$, $\widehat{\sigma}_{\max}$ and σ^2 to be constant with respect to the number of observations n and the dimension p. Corollary 3.3.2 provides an explicit approximation process that achieves a given accuracy ϵ. Conditions on the stepsize h as well as on the number of iterations K are explicit. Similarly to previous results, two different rates of convergence are given conditionally to the stepsize h. In Corollary 3.3.2 we also recommend g_γ to be chosen with

$$\gamma = \frac{\epsilon \widehat{\sigma}^{1/2}_{\min}}{3p^{3/2}\log(p)^{1/2}}. \tag{3.3.16}$$

Corollary 3.3.2. *Let μ be the probability measure of the pseudo-posterior of the EWA with Laplace prior with temperature $\tau > 0$, as described by the potential f in Equation 3.3.5. Let γ be defined as in Equation 3.3.16 and let μ^γ be the measure probability associated to f_γ in Equation 3.3.8. Let $h \in \mathbb{R}^\star$ and $K \in \mathbb{N}^\star$.*

For any Dirac measure $\nu = \delta_{\boldsymbol{\beta}^{(0)}}$, we consider the probability measure $\widehat{\mu}^K_{\gamma,h}$ defined by the probability distribution $\nu P^K_{\gamma,h}$, where $P^K_{\gamma,h}$ is the discretized process diffusion approximation of the measure μ^γ as described in Equation 3.1.11.

If $h < 2/(M_\gamma + m_\gamma)$, then for any $\epsilon > 0$,

$$W_2(\widehat{\mu}^{s,h}_K, \mu^s) \leq \epsilon, \tag{3.3.17}$$

if

$$Kh \geq \sigma^2 \frac{\log\left(\frac{3}{\epsilon}\left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2 + \frac{\sigma}{\sqrt{n\widehat{\sigma}_{\min}}}\right)\right)}{np\widehat{\sigma}_{\min}}, \tag{3.3.18}$$

$$\frac{1}{h} > \frac{31}{\epsilon^2}\left(\frac{\widehat{\sigma}_{\max}\sqrt{p}}{\widehat{\sigma}_{\min}} + \frac{3\sigma p^2 \log(p)}{\epsilon\sqrt{n\widehat{\sigma}^3_{\min}}}\right)^2. \tag{3.3.19}$$

If we assume $\widehat{\sigma}_{\min}$, $\widehat{\sigma}_{\max}$ to be constant, $\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2$ to be of order $\sqrt{p}$ or smaller such that, $\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + (\sigma/\sqrt{n}) \leq \mathcal{O}(\sqrt{p})$ and h to be small enough, then, there is a number of iterations

$$K = \mathcal{O}\left(\frac{p^3 \log(p)^2}{(n\epsilon)^4}\log(\sqrt{p}/\epsilon)\right)$$

such that the target accuracy ϵ is achieved. This result should be compared to the paper Brosse et al. (2017). However, even though the authors defined a more general smoothing method, the smoothing parameter γ are not explicited. It is reasonable to consider γ to depend on parameters such as the dimension p. Therefore, it is not easy to compare the complexity of the general method described in Brosse et al. (2017) with the method we study. The result of Corollary 3.3.2 is to be compared with the results mentioned in Dalalyan (2017). Indeed, when $f \in \mathcal{F}_{M,m}$, a number of iterations $K = \mathcal{O}(p \log(p/\epsilon)\epsilon^{-2})$ is sufficient to achieve an accuracy of order ϵ. Thus, a higher number of iterations is needed when the Lispschitzness of the gradient does not hold.

Remark 3.3.1. *When the choice is given, a small stepsize such as $h < 2/(M_\gamma + m_\gamma)$ enables to reach accuracy ϵ in a finite number of iterations K as described in Corollary 3.3.2. However, one may wish to guarantee a convergence rate of the approximation in a situation where the discretization process has been generated with a given $h \geq 2/(M_\gamma + m_\gamma)$. In that case, if*

$$\frac{1}{h} < \frac{60p}{\epsilon^2} \vee \frac{M_\gamma}{2},$$

then for any $\epsilon > 0$, Inequality 3.3.17 holds if,

$$K \geq \frac{\log\left(\frac{3\left(\|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2 + \frac{\sigma}{\sqrt{n\widehat{\sigma}_{\min}}}\right)}{\epsilon}\right)}{2 - \left(np\widehat{\sigma}_{\max}/\sigma^2 + \frac{\log(p)^{3/2}p^{5/2}}{\sigma\epsilon\sqrt{\widehat{\sigma}_{\min}}}\right)h}.$$

Corollary 3.3.2 and Remark 3.3.1 are proven in Section 3.5.

3.4 Discussion and outlook

This study establishes guarantees in the spirit of Dalalyan (2017) for any log-density that is close, in the sense of the Wasserstein distance, to a strongly convex log-density with a Lipschitz gradient. A particularly suited application is the log-density of the exponentially weighted aggregate with Laplace prior. We have established explicit results in the context of this application with the assumption that the Gram matrix is invertible when the temperature parameter $\tau = \sigma^2/(np)$ and the penalty parameter $\lambda = \sigma[\log(p)/n]^{1/2}$, as recommended in Chapter 2. The sampling approximation is computable in polynomial time and explicit constants are given. Matching the optimization performance in the context of sampling is a very promising and exciting subject. In a practical situation, in order to improve the algorithm, the choice of an

optimal smooth log-density is important. To do so, the computation of the Wasserstein distance is necessary, some researches, such as Solomon et al. (2015), provide tools to start solving such questions.

In the context of regression in high-dimensional settings, guaranteeing a targeted accuracy when the design matrix is such that the Gram matrix is not invertible would be of great interest. The generalization of the aforementioned results to other discretization schemes could broaden the understanding of the sampling algorithms. In particular, considering the Ozaki scheme as defined in Ozaki (1992) could be useful to study the associated performance. The analysis of other log-densities inspired from other penalized regression problems such as the SLOPE, as in Bogdan et al. (2015) and Sepehri (2016), or even the nuclear-norm penalization in the matrix regression context, as in Koltchinskii et al. (2011b), would be of great interest.

Another promising field of study is the generalization of these results to other sampling methods. For example, guarantees on time-inhomogeneous Langevin-type processes, as defined in Andrieu et al. (2016), or on the Hamiltonian Monte Carlo, as explained in Neal (2011). Furthermore, such studies could provide useful insights to deepen the understanding of the existing similarities between optimization and sampling performances. On the spectrum of analogies with the optimization, the question of sampling on a convex subset is very important. The study of theoretical guarantees on samplings methods based on reflecting diffusion such as in Skorokhod (1961) and Tanaka (1979) would be very promising.

3.5 Proofs

In this section, we prove the claims of Propositions 3.3.1 and 3.3.2 as well as the results of Corollary 3.3.2 and Remark 3.3.1.

3.5.1 Proof of Proposition 3.3.1

Let $\mathbf{X}$ be the design matrix such that the associated Gram matrix $\mathbf{X}^\top\mathbf{X}/n$ is invertible. We define $\widehat{\sigma}_{\min}$ (respectively $\widehat{\sigma}_{\max}$) to be the smallest (resp. greatest) eigen value of the Gram matrix and we assume $\widehat{\sigma}_{\min} > 0$. Let γ be a positive real number. With respect to $\mathbf{X}$ and γ, the negative log-likelihood f_γ is defined in Equation 3.3.8. In order to prove Proposition 3.3.1, we will first study the strong-convexity property of f_γ with Lemmas 3.5.1 and 3.5.2 which respectively provide results on the strong-convexity of L and g_γ. In a second time, Lemmas 3.5.3 and 3.5.4 will ensure properties on the Lipschitz property of the gradient of f_γ with respect

to $\boldsymbol{\beta}$, ∇f_γ.

These various lemmas will use some derivative results up to the third order of the functions L and g_γ. We start the proof of Proposition 3.3.1 with the definition of these derivatives.

The function L is well known in the statistical literature and the following properties are well known (see Montgomery et al. (2015)[Section 3.2.1] for example):

$$\nabla L(\boldsymbol{\beta}) = \frac{\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top \boldsymbol{y}}{n}, \tag{3.5.1}$$

and

$$\nabla^2 L(\boldsymbol{\beta}) = \frac{\mathbf{X}^\top \mathbf{X}}{n}. \tag{3.5.2}$$

Furthermore, it is then trivial that

$$\nabla^3 L(\boldsymbol{\beta}) = 0. \tag{3.5.3}$$

The function g_γ is differentiable and for any $j \in [p]$,

$$\frac{\partial g_\gamma}{\partial \beta_j}(\boldsymbol{\beta}) = \frac{\beta_j}{\left(\gamma^2 + \beta_j^2\right)^{1/2}}. \tag{3.5.4}$$

On the second order derivative, the Hessian matrix of g_γ is a diagonal matrix such that diagonal coefficients are defined by

$$\frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) = \frac{1}{(\gamma^2 + \beta_j^2)^{1/2}} - \frac{\beta_j^2}{(\gamma^2 + \beta_j^2)^{3/2}}, \tag{3.5.5}$$

and every non-diagonal term is null. Finally, we can show that the third differentiable of g_γ is null everywhere but in the element $\frac{\partial^3 g_\gamma}{\partial \beta_j^3}$, for any $j \in [p]$,

$$\frac{\partial^3 g_\gamma}{\partial \beta_j^3}(\boldsymbol{\beta}) = \frac{3\beta_j}{(\gamma^2 + \beta_j^2)^{3/2}}\left(\frac{\beta_j^2}{(\gamma^2 + \beta_j^2)} - 1\right). \tag{3.5.6}$$

The development of the calculations to prove Equations 3.5.4, 3.5.5 and 3.5.6 are left to the reader, they follow on from combinations of standard derivative properties. That $\nabla^3 f_\gamma$ is well defined is of interest for one who wants to study theoretical properties of other discretization schemes than the Euler Maruyama in order to approximate the Langevin Monte Carlo process. In particular, it is of interest for higher order numerical schemes. For example, one could consider the Ozaki discretization as defined in Ozaki (1992) that has been proposed to be used in Stramer and Tweedie (1999) to approximate the Langevin Monte Carlo algorithm.

Now that these results have been mentioned, we can study the properties of L and g_γ.

Lemma 3.5.1. *The function L is $\widehat{\sigma}_{\min}$-strongly convex and there is no $m > \widehat{\sigma}_{\min}$ such that L is m-strongly convex.*

This Lemma holds true from the definition of $\widehat{\sigma}_{\min}$ as the smallest eigen value of the Gram matrix $\mathbf{X}^\top\mathbf{X}/n$. Indeed, from Equation 3.5.2, since for any $\boldsymbol{\beta} \in \mathbb{R}^p$, $\nabla^2 L(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{X}/n$, it implies that

$$\nabla^2 L(\boldsymbol{\beta}) - \widehat{\sigma}_{\min}\mathbf{I}_p(\boldsymbol{\beta}) = \Big(\frac{\mathbf{X}^\top\mathbf{X}}{n} - \widehat{\sigma}_{\min}\mathbf{I}_p\Big)\boldsymbol{\beta} \succcurlyeq 0. \tag{3.5.7}$$

Therefore, L is $\widehat{\sigma}_{\min}$-strongly convex.
Moreover, since $\widehat{\sigma}_{\min}$ is the smallest eigen value, there is no $m > \widehat{\sigma}_{\min}$ such that $\nabla^2 L(\boldsymbol{\beta}) \succcurlyeq m\mathbf{I}_p(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, otherwise the smallest eigen-value of the Gram matrix would be greater than m which is contradictory with $m > \widehat{\sigma}_{\min}$. Therefore, there is no $m > \widehat{\sigma}_{\min}$ such that L is m-strongly convex and consequently, L is $\widehat{\sigma}_{\min}$-strongly convex with $\widehat{\sigma}_{\min}$ being the optimal constant. □

Lemma 3.5.2. *The smooth function g_γ defined in Equation 3.3.7 is strictly convex but not strongly convex.*

We remind that the non-diagonal elements of $\nabla^2 g_\gamma$ are null and that the diagonal terms are defined by Equation 3.5.5. As a consequence, proving that for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and any $j \in [p]$, $\frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) > 0$ is equivalent to prove the strict convexity of g_γ. Moreover, if for any $\iota > 0$ there exists $\boldsymbol{\beta} \in \mathbb{R}^p$ and $j \in [p]$, such that

$$\frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) < \iota,$$

then, g_γ is not strongly convex.
From Equation 3.5.5,

$$\frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) = \left(1 - \frac{\beta_j^2}{(\gamma^2 + \beta_j^2)}\right)(\gamma^2 + \beta_j^2)^{-1/2}.$$

Therefore, for any $\beta_j \in \mathbb{R}$, $\beta_j^2 < \beta_j^2 + \gamma^2$, implies that $1 - (\beta_j^2/\big(\beta_j^2 + \gamma^2\big)) > 0$. Therefore, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) > 0,$$

which proves the strict convexity of g_γ. However,

$$\lim_{\beta_j \to +\infty} \frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) = 0.$$

Therefore, for any $\iota > 0$, there exits $\beta \in \mathbb{R}^p$ such that $\nabla^2 g_\gamma(\boldsymbol{\beta}) \prec \iota\mathbf{I}_p$. Consequently, for any $\iota > 0$, g_γ is not ι-strongly convex. We have proved that g_γ is strictly convex and not strongly

convex and have therefore proved Lemma 3.5.2. □

The sum of a m-strongly convex function with a strictly but not strongly convex function is a m-strongly convex function. Thus, f_γ is $(\widehat{\sigma}_{\min}/\tau)$-strongly convex from Lemmas 3.5.1 and 3.5.2 and $\widehat{\sigma}_{\min}$ is optimal.

The strongly convex property has been proven. In order to conclude the proof of Proposition 3.3.1 we study the Lipschitz property of the gradient ∇f_γ.

Lemma 3.5.3. *Let L be defined in Equation 3.3.3 where $\mathbf{X}$ is a matrix of data with the greatest eigenvalue $\widehat{\sigma}_{\max}$ of the Gram matrix $\mathbf{X}^\top\mathbf{X}/n$ known. Then, the gradient ∇L with respect to $\boldsymbol{\beta}$ is $\widehat{\sigma}_{\max}$-Lipschitz and there is no $M < \widehat{\sigma}_{\max}$ such that ∇L is M-Lipschitz.*

To prove Lemma 3.5.3, we remark that $\nabla^2 L$ exists and is defined in Equation 3.5.2. Therefore, proving Lemma 3.5.3 is equivalent to prove that $\widehat{\sigma}_{\max}$ is the smallest quantity such that for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\|n^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\|_2 \leq \widehat{\sigma}_{\max}\|\boldsymbol{\beta}\|_2,$$

which is true from the definition of $\widehat{\sigma}_{\max}$ as the largest eigenvalue of the Gram matrix $\mathbf{X}^\top\mathbf{X}/n$. □

Lemma 3.5.4. *The gradient ∇g_γ is γ^{-1}-Lipschitz and there is no $M < \gamma^{-1}$ such that ∇L is M-Lipschitz.*

From equation 3.5.5, the Hessian $\nabla^2 g_\gamma$ exists. Therefore, proving that ∇g_γ is γ^{-1}-Lipschitz and that there is no smaller constant than γ^{-1} for which ∇g_γ is Lipschitz is equivalent to guarantee that

$$\gamma^{-1}\mathbf{I}_p \succcurlyeq \nabla^2 g_\gamma(\boldsymbol{\beta}),$$

for any $\boldsymbol{\beta} \in \mathbb{R}^p$.

The third derivative operator $\nabla^3 g_\gamma$ is also diagonal and we deduce from Equations 3.5.5 and 3.5.6 that for any $j \in [p]$, $\frac{\partial^2 g_\gamma}{\partial \beta_j^2}$ reaches its maximum in any $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\beta_j = 0$. Indeed, $\nabla^2 g_\gamma$ is continuous and $\nabla^3 g_\gamma(\boldsymbol{\beta}) = 0$ implies that $\boldsymbol{\beta} = \mathbf{0_p}$. Moreover, for any $j \in [p]$, and for any $\beta_j \in \mathbb{R}$,

$$\lim_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ |\beta_j|\to+\infty}} \frac{\partial^2 g_\gamma}{\partial \beta_j^2}(\boldsymbol{\beta}) = 0 < \frac{\partial^2 g_\gamma}{\partial \widetilde{\beta}_j^2}(\widetilde{\boldsymbol{\beta}}) = \gamma^{-1},$$

for any $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that $\widetilde{\beta}_j = 0$.

Therefore, the diagonal elements of $\nabla^2 g_\gamma(\boldsymbol{\beta})$ are not greater than the diagonal elements of

$\nabla^2 g_\gamma(\mathbf{0}_p)$, which are equal to γ^{-1}. Thus, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\gamma^{-1}\mathbf{I}_p = \nabla^2 g_\gamma(\mathbf{0}_p) \succcurlyeq \nabla^2 g_\gamma(\boldsymbol{\beta}). \tag{3.5.8}$$

The inequality in Equation 3.5.8 proves that $\nabla^2 g_\gamma$ is γ^{-1}-Lipschitz, while the left hand equality proves the optimality of γ^{-1} as the Lipschitz parameter. Hence, we have concluded the proof of Lemma 3.5.4. □

Lemmas 3.5.3 and 3.5.4 combined prove that the optimal M_γ-Lipschitz property of ∇f_γ is

$$M_\gamma = \frac{1}{\tau}\big(\widehat{\sigma}_{\max} + \frac{\lambda}{\gamma}\big).$$

We have calculated the gradients and Hessians of L and g_γ and have shown the optimal m_γ-strong convexity of f_γ as well as the optimal M_γ-Lipschitz property of ∇f_γ. Therefore, $f_\gamma \in \mathcal{F}_{M_\gamma, m_\gamma}$ and the couple (M_γ, m_γ) is optimal, which concludes the proof of Proposition 3.3.1. □

3.5.2 Proof of Proposition 3.3.2

Let μ^γ be the measure associated with f_γ defined in Equation 3.3.8. Proposition 3.3.2 upper bounds the Wasserstein distance $W_2(\mu, \mu^\gamma)$. To prove this upper bound in the clearest possible way, we define the quantities $a_\gamma = \widehat{\sigma}_{\max} + \frac{\lambda}{\gamma}$, $b_\gamma = \frac{\widehat{\sigma}_{\min} + a_\gamma}{\widehat{\sigma}_{\min}}$ and $c_\gamma = a_\gamma b_\gamma$.
With these notations, and using the fact that we assume that $\tau = np/\sigma^2$,

$$M_\gamma = \frac{1}{\tau}\big(\widehat{\sigma}_{\max} + \frac{\lambda}{\gamma}\big) = \frac{np\ a_\gamma}{\sigma^2}, \tag{3.5.9}$$

and,

$$m_\gamma = \frac{\widehat{\sigma}_{\min}}{\tau} = \frac{np\ \widehat{\sigma}_{\min}}{\sigma^2}. \tag{3.5.10}$$

Hence,

$$\frac{\big(\gamma p \lambda\big)^2}{m_\gamma \tau^2} = \frac{\sigma^2\big(np\ \gamma p\lambda\big)^2}{np\ \widehat{\sigma}_{\min}\sigma^4} = \frac{np^3\lambda^2\gamma^2}{\sigma^2\ \widehat{\sigma}_{\min}}. \tag{3.5.11}$$

Moreover, from Proposition 3.2.2, the inequality

$$KL(\mu \| \mu^\gamma) \leq \frac{1}{2}\int_{\mathbb{R}^p} (f(\boldsymbol{u}) - f_\gamma(\boldsymbol{u}))^2 \pi(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$$

holds. Therefore, remarking that $(f(\boldsymbol{\beta}) - f_\gamma(\boldsymbol{\beta}))^2 < (\lambda p\gamma/\tau)^2$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and averaging this upper bound with respect to the density measure π, we deduce that

$$KL(\mu\|\mu^\gamma) \leq \frac{1}{2}\Big(\frac{p\lambda\gamma}{\tau}\Big)^2. \tag{3.5.12}$$

Finally, applying Proposition 3.2.1 and the result of Equation 3.5.12 implies that if $f_\gamma \in \mathcal{F}_{M_\gamma,m_\gamma}$, then

$$W_2(\mu,\mu^\gamma) \leq \frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}}.$$

This upper bound concludes the proof by remarking that $m_\gamma = \frac{\widehat{\sigma}_{\min}}{\tau}$. □

3.5.3 Proof of Corollary 3.3.2 and Remark 3.3.1

In order to prove Corollary 3.3.2 and Remark 3.3.1, we consider a targeted accuracy $\epsilon > 0$, such that $\widehat{\mu}_K^{s,h}$ is a good approximation of the measure μ in the sense that $W_2(\mu,\widehat{\mu}_K^{s,h}) < \epsilon$. The goal of this corollary is to provide explicit values of the parameters that one has to choose in practical situations, namely the stepsize h, the smoothing parameter γ and the number of iterations K needed to reach the accuracy level ϵ. To do so, we approach the problem by splitting the upper bound of the quantity $W_2(\widehat{\mu}_K^{s,h},\mu)$ in three terms as in Inequalities 3.3.12 or 3.3.13. We will upper bound the three quantities by $(\epsilon/3)$ so that the sum is upper bound by the targeted accuracy ϵ such that,

$$\begin{cases} (1-m_\gamma h)^K A \leq \dfrac{\epsilon}{3}, & (3.5.13)\\ 1.82\dfrac{M_\gamma}{m_\gamma}(hp)^{1/2} \leq \dfrac{\epsilon}{3}, & (3.5.14)\\ W_2(\mu,\mu^\gamma) \leq \dfrac{\epsilon}{3}, & (3.5.15) \end{cases}$$

where

$$A \triangleq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_s\|_2 + \Big(\frac{p}{m_\gamma}\Big)^{1/2}. \tag{3.5.16}$$

As a consequence, Inequality 3.3.17 would hold and Corolarry 3.3.2 would be proven. In order to prove Remark 3.3.1 Inequality 3.5.14 has to be substituted with

$$1.82\frac{M_\gamma h}{2-M_\gamma h}(hp)^{1/2} \leq \frac{\epsilon}{3}, \tag{3.5.17}$$

and Inequality 3.5.13 has to be substituted with

$$(M_\gamma h - 1)^K A \leq \frac{\epsilon}{3}. \tag{3.5.18}$$

While Lemma 3.5.5 will provide explicit condition on γ to guarantee Inequality 3.5.15, Lemmas 3.5.6 and 3.5.7 (respectively Lemmas 3.5.8 and 3.5.9) will provide explicit requirements on K and h that guarantee the remainings inequalities.

For clarity sake, we assume

$$\tau = \frac{np}{\sigma^2}, \tag{3.5.19}$$

and

$$\lambda = \sigma[\log(p)/n]^{1/2}. \tag{3.5.20}$$

Of course, one could adjust this proof with specific values of τ and γ. As we will see, the value of h interfers in the result and the convergence rate. We consider h to be small when h is such that $h \leq \frac{2}{M_\gamma + m_\gamma}$, we consider h to be relatively large when $\frac{2}{M_\gamma + m_\gamma} \leq h < \frac{2}{M_\gamma}$. Inequality 3.3.12 applies when h is small and Inequality 3.3.13 applies when h is large. However, the last term of both inequalities is the same. As a consequence, we start the proofs of Corollary 3.3.2 and Remark 3.3.1 with Lemma 3.5.5, which provides a first condition on how small should be γ.

Lemma 3.5.5. *The following inequality*

$$\frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}} \leq \frac{\epsilon}{3} \tag{3.5.21}$$

holds if and only if

$$\gamma \leq \frac{\epsilon\widehat{\sigma}_{\min}^{1/2}}{3p^{3/2}\log(p)^{1/2}}. \tag{3.5.22}$$

Indeed, from Equations 3.5.19 and 3.5.20, the inequality

$$\frac{p\lambda\gamma}{\tau\sqrt{m_\gamma}} \leq \frac{\epsilon}{3}$$

is equivalent to

$$\frac{p\sigma[\log(p)/n]^{1/2}\gamma}{\sigma^2/(np)\sqrt{\frac{np\ \widehat{\sigma}_{\min}}{\sigma^2}}} \leq \frac{\epsilon}{3}.$$

Therefore, it is equivalent to

$$\frac{p^{3/2}\log(p)^{1/2}\gamma}{\widehat{\sigma}_{\min}^{1/2}} \leq \frac{\epsilon}{3}.$$

We conclude that Inequality 3.5.21 holds if and only if

$$\gamma \leq \frac{\epsilon\widehat{\sigma}_{\min}^{1/2}}{3p^{3/2}\log(p)^{1/2}}.$$

□

The choice of γ is of paramount importance, it determines how rough is the smoothing approximation of μ. Lemma 3.5.5 guarantees an upper bound of the quality as long as γ is smaller than a specific threshold. A small value of γ guarantee a good approximation of the target measure μ by μ^γ. Not only the parameter γ plays a role in the smooth approximation of the measure, it also has an impact on the Lipschitz property of the gradient ∇f_γ. As mentioned in Proposition 3.3.1, a small value of γ as a negative impact on the regularity of the gradient as shown in Equation 3.3.9. On the other hand, γ has no effect on the strong convexity parameter m_γ as Equation 3.3.10 shows. From this point of view, we recommend to choose the greatest value of γ that respects Inequality 3.5.22, which is

$$\gamma = \frac{\epsilon \widehat{\sigma}_{\min}^{1/2}}{3p^{3/2}\log(p)^{1/2}}. \tag{3.5.23}$$

From now on, we assume γ to be set as in Equation 3.5.23. Now, we focus on the case where h is small in the sense that $h \leq \frac{2}{M_\gamma + m_\gamma}$. In that case two lemmas are necessary to conclude the proof of Corollary 3.3.2.

Lemma 3.5.6. *If $h \leq \frac{2}{M_\gamma + m_\gamma}$, then the inequality*

$$1.82\frac{M_\gamma}{m_\gamma}(hp)^{1/2} \leq \frac{\epsilon}{3} \tag{3.5.24}$$

holds if

$$\frac{1}{h} > \frac{31}{\epsilon^2}\left(\frac{\widehat{\sigma}_{\max}\sqrt{p}}{\widehat{\sigma}_{\min}} + \frac{3\sigma\log(p)p^2}{\epsilon\sqrt{n\widehat{\sigma}_{\min}^3}}\right)^2. \tag{3.5.25}$$

In order to prove Lemma 3.5.6, let first remark that $31 > 9(1.82)^2$. Furthermore, if h is such that

$$h < \frac{\epsilon^2}{9\times(1.82)^2}\left(\frac{\widehat{\sigma}_{\max}\sqrt{p}}{\widehat{\sigma}_{\min}} + \frac{3\sigma\log(p)p^2}{\epsilon\sqrt{n\widehat{\sigma}_{\min}^3}}\right)^{-2},$$

then, it implies that

$$(1.82)^2 h\left(\frac{\widehat{\sigma}_{\max}\sqrt{p}}{\widehat{\sigma}_{\min}} + \frac{3\sigma\log(p)p^2}{\epsilon\sqrt{n\widehat{\sigma}_{\min}^3}}\right)^2 < \frac{\epsilon^2}{9}.$$

For any positive h, this inequality is equivalent to

$$1.82\sqrt{hp}\left(\frac{\widehat{\sigma}_{\max}}{\widehat{\sigma}_{\min}} + \frac{3\sigma\log(p)p^{3/2}}{\epsilon\sqrt{n\widehat{\sigma}_{\min}^3}}\right) < \frac{\epsilon}{3}. \tag{3.5.26}$$

From Equations 3.5.20 and 3.5.23, we remark that

$$\frac{M_\gamma}{m_\gamma} = \frac{\widehat{\sigma}_{\max}}{\widehat{\sigma}_{\min}} + \frac{3\sigma \log(p) p^{3/2}}{\epsilon \sqrt{n \widehat{\sigma}^3_{\min}}}, \tag{3.5.27}$$

where M_γ and m_γ are defined in Proposition 3.3.1. Therefore, the combination of Inequality 3.5.26 and Equation 3.5.27 implies the claim of Lemma 3.5.6. Namely, if h is such that

$$h < \frac{\epsilon^2}{15} \left(\frac{\widehat{\sigma}_{\max} \sqrt{p}}{\widehat{\sigma}_{\min}} + \frac{3\sigma \log(p) p^2}{\epsilon \sqrt{n \widehat{\sigma}^3_{\min}}} \right)^{-2},$$

then

$$1.82 \frac{M_\gamma}{m_\gamma} (hp)^{1/2} \leq \frac{\epsilon}{3}.$$

□

Therefore, in order to obtain Inequality 3.5.24, the stepsize h has to be small enough. When $h \leq \frac{2}{M_\gamma + m_\gamma}$, Lemmas 3.5.5 and 3.5.6 respectively provide conditions on γ and h. The last condition will focus on the number of iterations K needed to achieve a targeted accuracy. This condition is given by Lemma 3.5.7.

Lemma 3.5.7. *Let ν be a Dirac measure concentrated at $\boldsymbol{\beta}^{(0)}$ and let $\bar{\boldsymbol{\beta}}_\gamma$ be the average with respect to the density π_γ. Let A be the quantity defined by*

$$A \triangleq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2 + \frac{\sigma}{\sqrt{n \widehat{\sigma}_{\min}}}. \tag{3.5.28}$$

If $h \leq \frac{2}{M_\gamma + m_\gamma}$, then the inequality

$$(1 - m_\gamma h)^K W_2(\nu, \mu^\gamma) \leq \frac{\epsilon}{3} \tag{3.5.29}$$

holds if

$$Kh \geq \frac{\log\left(\frac{3A}{\epsilon}\right)}{m_\gamma}. \tag{3.5.30}$$

In order to prove Lemma 3.5.6, we explicit a number of iterations K that guarantees the upper bound of Inequality 3.5.29. To do so, we will use the result of Proposition 3.2.2 and we will assume ν to be a Dirac measure concentrated at $\boldsymbol{\beta}^{(0)}$. Let us define $\bar{\boldsymbol{\beta}}_\gamma$ the average over π_γ. Then, from Proposition 3.2.2, we remark that

$$W_2^2(\nu, \mu^\gamma) \leq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2^2 + \frac{p}{m_\gamma}.$$

Thus, from the definitions of m_γ and τ respectively in Equations 3.3.10 and 3.5.19,

$$W_2(\nu, \mu^\gamma) \leq \|\boldsymbol{\beta}^{(0)} - \bar{\boldsymbol{\beta}}_\gamma\|_2 + \frac{\sigma}{\sqrt{n\widehat{\sigma}_{\min}}} = A.$$

Moreover, since $0 < h < \frac{2}{m_\gamma + M_\gamma}$,

$$0 < m_\gamma h < 2\frac{\widehat{\sigma}_{\min}}{\widehat{\sigma}_{\min} + \widehat{\sigma}_{\max} + \lambda/\gamma} < 1.$$

Therefore, $0 < m_\gamma h < 1$ and consequently, for any $h < \frac{2}{M_\gamma + m_\gamma}$, $(1 - m_\gamma h) \leq \exp(-m_\gamma h)$. It implies that

$$(1 - m_\gamma h)^K \leq \exp(-K m_\gamma h).$$

As a consequence, if K is such that

$$\exp(-K m_\gamma h) A \leq \frac{\epsilon}{3}, \tag{3.5.31}$$

then Inequality 3.5.29 holds. Moreover, since Inequality 3.5.31 is equivalent to

$$Kh \geq \frac{\log(\frac{3A}{\epsilon})}{m_\gamma},$$

Lemma 3.5.7 is proven. Indeed, if Inequality 3.5.30 holds then

$$(1 - m_\gamma h)^K W_2(\nu, \mu^\gamma) \leq \exp(-K m_\gamma h) A \leq \frac{\epsilon}{3},$$

which concludes the proof. □

So far, we have proven Lemmas 3.5.5, 3.5.6 and 3.5.7. These three Lemmas combined together prove the result of Corollary 3.3.2 in the case where h is small. Indeed, applying Inequalities 3.5.29, 3.5.21 and 3.5.24 within Inequality 3.3.12 guarantees that if $0 < h < \frac{2}{m_\gamma + M_\gamma}$ then, the statement of Corollary 3.3.2 holds:

$$W_2(\widehat{\mu}_K^{s,h}, \mu) \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3},$$

as long as Inequalities 3.5.22, 3.5.25 and 3.5.30 are verified.

In order to prove Remark 3.3.1, let us now consider the second case where h is relatively large, in the sense that $\frac{2}{M_\gamma + m_\gamma} \leq h \leq \frac{2}{M_\gamma}$.

Lemma 3.5.8. *If* $\frac{2}{M_\gamma+m_\gamma} \leq h \leq \frac{2}{M_\gamma}$, *then the inequality*

$$1.82\frac{M_\gamma h}{2-M_\gamma h}(hp)^{1/2} \leq \frac{\epsilon}{3} \tag{3.5.32}$$

holds if

$$h < \frac{\epsilon^2}{11p}. \tag{3.5.33}$$

Indeed, we remark that if $\frac{2}{M_\gamma+m_\gamma} \leq h \leq \frac{2}{M_\gamma}$, then,

$$\begin{aligned} 1.82\frac{M_\gamma h}{2-M_\gamma h}(hp)^{1/2} &\leq 1.82\frac{M_\gamma \frac{2}{M_\gamma}}{2-M_\gamma/M_\gamma}(hp)^{1/2} \\ &\leq 3.64(hp)^{1/2}. \end{aligned}$$

Moreover,

$$\left\{3.64(hp)^{1/2} \leq \frac{\epsilon}{3}\right\} \Longleftrightarrow \left\{h \leq \frac{\epsilon^2}{10.92p}\right\}.$$

Therefore, Inequaliy 3.5.32 holds if

$$h \leq \frac{\epsilon^2}{11p}. \tag{3.5.34}$$

□

Lemma 3.5.9. *Let* ν *be a Dirac measure concentrated at* $\boldsymbol{\beta}^{(0)}$ *and let* $\bar{\boldsymbol{\beta}}_\gamma$ *be the average with respect to the density* π_γ *and let* A *be the quantity defined in Equation 3.5.28. If* $\frac{2}{M_\gamma+m_\gamma} \leq h \leq \frac{2}{M_\gamma}$, *then the inequality*

$$\left(M_\gamma h - 1\right)^K W_2(\nu, \mu^\gamma) \leq \frac{\epsilon}{3} \tag{3.5.35}$$

holds if

$$K \geq \frac{\log\left(\frac{3A}{\epsilon}\right)}{(2-M_\gamma h)}. \tag{3.5.36}$$

From our assumptions, we remark that $M_\gamma h \in (1,2)$. Therefore, $M_\gamma h - 1 \in (0,1)$ and

$$\exp\left(M_\gamma h - 2\right) \geq M_\gamma h - 1, \tag{3.5.37}$$

for any $h \in [M_\gamma/(M_\gamma+m_\gamma), 2]$. It implies that for any integer K,

$$\exp\left(K(M_\gamma h - 2)\right) \geq \left(M_\gamma h - 1\right)^K. \tag{3.5.38}$$

Therefore, if there exists $K \in \mathbb{N}$ such that

$$\exp\big(K(M_\gamma h - 2)\big)A \leq \frac{\epsilon}{3} \tag{3.5.39}$$

then Inequality 3.5.35 holds for this specific value of K.

The following inequality

$$K \geq \frac{\log\left(\frac{3A}{\epsilon}\right)}{(2 - M_\gamma h)},$$

is equivalent to Inequality 3.5.39. Therefore, if K is such that Inequality 3.5.36 holds, then

$$(M_\gamma h - 1)^K W_2(\nu, \mu^\gamma) \leq \exp\big(K(M_\gamma h - 2)\big)A \leq \frac{\epsilon}{3},$$

which concludes the proof of Lemma 3.5.9.

These results enable to confirm that for any $h \geq 2/(M_\gamma + m_\gamma)$, if Inequality (3.5.36) holds then Inequality 3.5.35 is guaranteed. □

Remark 3.5.1. *From the previous inequality, one remarks that in order to keep K reasonable, it is important, when $h \geq 2/(M_\gamma + m_\gamma)$ to choose a value of h not to close to $2/M_\gamma$.*

We have proven Lemmas 3.5.5, 3.5.8 and 3.5.9. These three Lemmas combined together prove the result of Corollary 3.3.2 in the case where h is large. Indeed, applying Inequalities 3.5.35, 3.5.21 and 3.5.32 within Inequality 3.3.13 guarantees that if $\frac{2}{m_\gamma + M_\gamma} < h < 2/M_\gamma$ then, the statement of Corollary 3.3.2 (respectively Remark 3.3.1) holds:

$$W_2 \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3},$$

as long as Inequalities 3.5.22, 3.5.33 and 3.5.36 are verified.

Therefore, we have proven Corollary 3.3.2 and Remark 3.3.1 statements when h is either small or relatively larger. □

Chapter 4

On the prediction loss of the lasso in the partially labeled setting

A joint work with Pierre Bellec, Arnak Dalalyan and Quentin Paris.

Contents

Abstract

In this paper we revisit the risk bounds of the lasso estimator in the context of transductive and semi-supervised learning. In other terms, the setting under consideration is that of regression with random design under partial labeling. The main goal is to obtain user-friendly bounds on the off-sample prediction risk. To this end, the simple setting of bounded response variable and bounded (high-dimensional) covariates is considered. We propose some new adaptations of the lasso to these settings and establish oracle inequalities both in expectation and in deviation. These results provide non-asymptotic upper bounds on the risk that highlight the interplay

between the bias due to the mis-specification of the linear model, the bias due to the approximate sparsity and the variance. They also demonstrate that the presence of a large number of unlabeled features may have significant positive impact in the situations where the restricted eigenvalue of the design matrix vanishes or is very small.

4.1 Introduction

We consider the problem of prediction under the quadratic loss. That is, for a random feature-label pair $(\boldsymbol{X}, Y)$ drawn from a distribution P on a product space $\mathcal{X} \times \mathcal{Y}$, we aim at predicting Y as a function of $\boldsymbol{X}$. The goal is to find a measurable function $f : \mathcal{X} \to \mathcal{Y}$ such that the expected quadratic risk,

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} (y - f(\boldsymbol{x}))^2 \, P(d\boldsymbol{x}, dy) = \mathbb{E}\big[\big(Y - f(\boldsymbol{X})\big)^2\big] \tag{4.1.1}$$

is as small as possible. When $\mathcal{Y}$ is an interval of $\mathbb{R}$ and $\mathcal{X}$ is a measurable set in $\mathbb{R}^p$—which is the setting considered in the present work—the Bayes predictor, defined as the minimizer of $\mathcal{R}(f)$ over all measurable functions $f : \mathcal{X} \to \mathcal{Y}$, is the regression function (Vapnik, 1998)

$$f^\star(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]. \tag{4.1.2}$$

Using $f^\star$, the problem can be rewritten in a form which is more familiar in Statistics, namely

$$Y = f^\star(\boldsymbol{X}) + \xi, \tag{4.1.3}$$

where the noise variable ξ satisfies $\mathbb{E}[\xi|\boldsymbol{X}] = 0$, P_X-almost surely[1]. In the present work, we tackle the prediction problem in the case where the available data $\mathcal{D}_{\text{all}}$ is of the form $\mathcal{D}_{\text{all}} = \mathcal{D}_{\text{labeled}} \cup \mathcal{D}_{\text{unlabeled}}$, where

$$\mathcal{D}_{\text{labeled}} = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\} \quad \text{and} \quad \mathcal{D}_{\text{unlabeled}} = \{\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_N\}.$$

The labeled sample $\mathcal{D}_{\text{labeled}}$ is composed of independent and identically distributed (i.i.d.) feature-label pairs with distribution P. The unlabeled sample $\mathcal{D}_{\text{unlabeled}}$ contains only i.i.d. features, with distribution P_X, and is independent of $\mathcal{D}_{\text{labeled}}$. This formal setting accounts for a number of realistic situations in which the labeling process is costly while the unlabeled data points are available in abundance (see, for instance, Balcan et al., 2005; Guillaumin et al., 2010;

[1] Notation P_X is used for the marginal distribution of $\boldsymbol{X}$.

Brouard et al., 2011), that is n may be quite small compared to N. Here, the baseline idea is to build upon the sample $\mathcal{D}_{\text{unlabeled}}$ to improve the supervised prediction process based on $\mathcal{D}_{\text{labeled}}$ alone. In this context, our study encompasses two closely related settings: semi-supervised learning and transductive learning.

In the *semi-supervised learning* setting, one aims at constructing a predictor $\widehat{f}$, based on the data $\mathcal{D}_{\text{all}}$, such that the excess risk

$$\mathcal{E}(\widehat{f}) = \mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star) = \int_{\mathbb{R}^p} \big(\widehat{f}(\boldsymbol{x}) - f^\star(\boldsymbol{x})\big)^2 P_X(d\boldsymbol{x}) = \|\widehat{f} - f^\star\|^2_{L_2(P_X)} \tag{4.1.4}$$

is as small as possible. This learning framework differs from the classical supervised learning only in that the data set is enriched by the unlabeled features.

In contrast with this, the goal of *transductive learning* is to predict solely the labels of the observed unlabeled features. This amounts to considering the same setting as above but to measure the quality of a prediction function f by the excess risk

$$\mathcal{E}_{\text{TL}}(f) = \frac{1}{N-n} \sum_{i=n+1}^{N} \big(f(\boldsymbol{X}_i) - f^\star(\boldsymbol{X}_i)\big)^2. \tag{4.1.5}$$

We refer the reader to (Chapelle et al., 2006; Zhu, 2008) and the references therein for a comprehensive survey on the topic of semi-supervised and transductive learning. Theoretical analysis of the generalisation error and the excess risk in this context can be found in (Rigollet, 2007; Wang and Shen, 2007; Lafferty and Wasserman, 2007), whereas the closely related area of manifold learning is studied in (Belkin et al., 2006; Nadler et al., 2009; Niyogi, 2013). The purpose of the present work differs from these papers in that we put the emphasis on the high-dimensional setting and the sparsity assumption. The goal is to understand whether the unlabeled data can help in predicting the unknown labels using the ℓ_1-penalized empirical risk minimizers. From another perspective—that of multi-view learning—the problem of sparse semi-supervised learning is investigated in (Sun and Shawe-Taylor, 2010).

When the feature vector is high dimensional, it is reasonable to consider prediction strategies based on "simple" functions f in order to limit the computational cost. A widely used approach is then to look for a good linear predictor

$$f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}, \qquad \boldsymbol{\beta} \in \mathbb{R}^p. \tag{4.1.6}$$

When the dimension p is of the same order as (or larger than) the size n of the labeled sample,

the simple empirical risk minimizer (*i.e.*, the least squares estimator) is a poor predictor since it suffers from the curse of dimensionality. To circumvent this shortcoming, one popular approach is to use the ℓ_1-penalised empirical risk minimizer, also known as the lasso estimator (Tibshirani, 1996a): $\widehat{f}^{\text{lasso}} = f_{\widehat{\boldsymbol{\beta}}^{\text{lasso}}}$ where[2]

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}_{\text{lab}}\boldsymbol{\beta}\|_2^2 + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{4.1.7}$$

where $\lambda > 0$ stands for a tuning parameter and

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X}_{\text{lab}} = \begin{bmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{bmatrix}. \tag{4.1.8}$$

Statistical properties of the lasso with regard to the prediction error were studied in many papers, the most relevant (to our purposes) of which will be discussed in the next section. We also refer the reader to (Bühlmann and van de Geer, 2011) for an overview of related topics. The rationale behind this approach is that (a) the term $\frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}_{\text{lab}}\boldsymbol{\beta}\|_2^2 - \mathbb{E}[\xi^2]$ is an unbiased estimator of the excess risk $\mathcal{E}(f_{\boldsymbol{\beta}})$ and (b) the ℓ_1-penalty term favors predictors $f_{\boldsymbol{\beta}}$ defined via a (nearly) sparse vector $\boldsymbol{\beta}$.

The prediction rules we are going to analyze in the present work are suitable adaptations of the (supervised) lasso to the semi-supervised and the transductive settings. More precisely, we consider the estimator

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{4.1.9}$$

where $\lambda > 0$ and $\mathbf{A} \in \mathbb{R}^{p\times p}$ are parameters to be chosen by the statistician. This definition is based on the following observation. The unlabeled sample may be used to get an improved estimator of the excess risk $\mathcal{E}(f_{\boldsymbol{\beta}}) = \mathbb{E}[f^\star(\boldsymbol{X})^2] - 2\mathbb{E}[Y\boldsymbol{X}^\top]\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}$, where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ is the $p \times p$ covariance matrix. Indeed, the population covariance matrix can be estimated using both labeled and unlabeled data. A similar observation holds for the transductive excess risk $\mathcal{E}_{\text{TL}}(f_{\boldsymbol{\beta}})$.

[2]To ease notation, we assume that both labels and features are centered, that is $\mathbb{E}[Y] = 0$ and $\mathbb{E}[\boldsymbol{X}] = 0$, so that there is no need to include an intercept in the linear combination $f_{\boldsymbol{\beta}}$.

Denoting by $\widehat{\boldsymbol{\Sigma}}_{\text{lab}}$ the empirical covariance matrix based on the labeled sample, that is

$$\widehat{\boldsymbol{\Sigma}}_{\text{lab}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

one checks that the vector $\widehat{\boldsymbol{\beta}}$ coincides with the lasso estimator (4.1.7) when $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\text{lab}}^{1/2}$. If an unlabeled sample is available, the foregoing discussion suggests a different choice for the matrix $\mathbf{A}$. This choice depends on the setting under consideration. Namely, defining the matrices

$$\widehat{\boldsymbol{\Sigma}}_{\text{all}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^\top \qquad \text{and} \qquad \widehat{\boldsymbol{\Sigma}}_{\text{unlab}} = \frac{1}{N-n}\sum_{i=n+1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

we use $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\text{all}}^{1/2}$ and $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}$ in the semi-supervised and transductive settings, respectively. The following two assumptions made on the probability distribution P will be repeatedly used throughout this work.

(A1) The random variables Y and $\boldsymbol{X}$ have zero mean and finite variance. Furthermore, all the coordinates X^j of the random vector $\boldsymbol{X}$ satisfy $\mathbb{E}[(X^j)^2] = 1$.

(A2) The random variables Y and X^j are almost surely bounded. That is, there exist constants B_Y and B_X such that $\mathbb{P}\big(|Y| \leq B_Y; \max_{j\in[p]} |X^j| \leq B_X\big) = 1$.

Assumption 4.1 is fairly mild, since one can get close to it by centering and scaling the observed labels and features. For features, the centering and the scaling may be performed using the sample mean and the sample variance computed over the whole data-set. It is however important to require this assumption, since its violation may seriously affect the quality of the ℓ_1-penalized least-squares estimator $\widehat{\boldsymbol{\beta}}$, unless the terms $|\beta_j|$ of the ℓ_1-norm are weighted according to the magnitude of the corresponding feature X^j. The second assumption is less crucial both for practical and theoretical purposes, given that its primary aim is to allow for user-friendly, easy-to-interpret theoretical guarantees. In most situations, even if assumption 4.1 is violated, the predictor $f_{\widehat{\boldsymbol{\beta}}}$ does have a fairly small prediction error rate.

The main contributions of the present work are:

- Review of the relevant recent literature on the off-sample performance of the lasso in the prediction problem.

- Non-asymptotic bounds for the prediction error of the lasso in the semi-supervised and transductive settings that guarantee the fast rate under the restricted eigenvalue condition. We did an effort for keeping the results easy to understand and to obtain small constants. These results are simple enough to be taught to graduate students.

- Oracle inequalities in expectation for the prediction error of the lasso. To the best of our knowledge, such results were not available in the literature until the very recent paper (Bellec et al., 2016b).

To give a foretaste of the results detailed in the rest of this work, let us state and briefly discuss a risk bound in the semi-supervised setting (the complete form of the result is provided in 4.5.2). For a matrix $\mathbf{A}$, we denote by $\|\mathbf{A}\|$ its largest singular value and by $\kappa_{\mathbf{A}}$ the compatibility constant (see 4.2 for a precise definition).

Theorem 4.1.1. *Let assumption 4.1 be fulfilled and let the random variables Y, X^j be bounded in absolute value by 1. For a prescribed tolerance level $\delta \in (0,1)$, assume that the overall sample size N and the tuning parameter λ satisfy $N \geq 18p\|\mathbf{\Sigma}^{-1}\| \log(3p/\delta)$ and*

$$\lambda \geq 4\left(\frac{2\log(6p/\delta)}{n}\right)^{1/2} + \frac{8\log(6p/\delta)}{3n}. \tag{4.1.10}$$

Then, for every $J \subseteq \{1,\dots,p\}$, with probability at least $1-\delta$, the estimator $\widehat{\boldsymbol{\beta}}$ defined in (4.1.9) above with $\mathbf{A} = \widehat{\mathbf{\Sigma}}_{\text{all}}^{1/2}$ satisfies

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2|J|}{2\kappa_{\widehat{\mathbf{\Sigma}}_{\text{all}}}(J,3)}\right\}. \tag{4.1.11}$$

This result follows in the footsteps of many recent papers such as (Koltchinskii et al., 2011a; Sun and Zhang, 2012b; Dalalyan et al., 2014b) among others. The term oracle inequality refers to the fact that it allows us to compare the excess risk of the predictor $f_{\widehat{\boldsymbol{\beta}}}$ to that of the best possible nearly sparse prediction function. (By nearly sparse we understand here a vector $\boldsymbol{\beta}$ such that for a set $J \subseteq \{1,\dots,p\}$ of small cardinality the entries of $\boldsymbol{\beta}$ with indices in J^c have small magnitude; that is $\|\boldsymbol{\beta}_{J^c}\|_1 = \sum_{j\notin J}|\beta_j|$ is small.) Indeed, if we denote by $\bar{\boldsymbol{\beta}}$ a nearly s-sparse vector in $\mathbb{R}^p$ such that the excess risk $\mathcal{E}(f_{\bar{\boldsymbol{\beta}}})$ is small, then the aforestated risk bound is the sum of three terms having clear interpretation. The first term, $\mathcal{E}(f_{\bar{\boldsymbol{\beta}}})$, is a bias term due to the s-sparse linear approximation. The second term, $\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|$, is the bias due to approximate s-sparsity. (Note that it vanishes if $\bar{\boldsymbol{\beta}}$ is exactly s-sparse and J is taken as its support.) Finally, the third term measures the magnitude of the stochastic error. Assuming the compatibility constant to be bounded away from 0, this last term is of the order $s\log(p)/n$, which is known to be optimal[3] over all possible estimators (Ye and Zhang, 2010; Raskutti et al., 2011; Rigollet and Tsybakov, 2011a, 2012a).

[3]More precisely, the optimal rate is $\frac{s\log(1+p/s)}{n}$, which is of the same order as $\frac{s\log(p)}{n}$ for most values of s.

Inequality (4.1.11) readily shows the advantage of using the unlabeled data: the compatibility constant involved in the last term of the right hand side is computed for the overall covariance matrix. When the size of the labeled sample is small in regard to the dimension p, the corresponding constant computed for $\widehat{\mathbf{\Sigma}}_{\text{lab}}$ may be very close (and even equal) to zero. This may downgrade the fast rate of the original lasso to the slow rate $\|\bar{\boldsymbol{\beta}}\|_1/\sqrt{n}$. Instead, if a large number of unlabeled features are used, it becomes more plausible to assume that the compatibility constant is bounded away from zero. In relation with this, it is important to underline that the unlabeled sample cannot help to improve the fast rate of convergence of the lasso, $s\log(p)/n$, which is optimal in the minimax sense. The best we can hope to achieve using the unlabeled sample is the relaxation of the conditions guaranteeing the fast rate. Another worthwhile remark is that the theorem stated above is valid when the size of the unlabeled sample is significantly larger than the dimension p. Interestingly, this condition is not required for getting the analogous result in the transductive set-up.

The rest is as follows. In 4.2, we introduce the notations used throughout the paper. 4.3 contains a review of the relevant literature and discusses the relation of the previous work with our results. 4.4 presents risk bounds for the prediction error of the lasso in the transductive setting, whereas 4.5 is devoted to the analogous results in the semi-supervised setting. Conclusions are made in 4.6. The proofs are postponed to 4.7.

4.2 Notations

In the sequel, for any integer k we denote by $[k]$ the set $\{1,\dots,k\}$. For any $q \in [1,+\infty]$ the notation $\|\boldsymbol{v}\|_q$ refers to the ℓ_q-norm of a vector $\boldsymbol{v}$ belonging to an Euclidean space $\mathbb{R}^k$ with arbitrary dimension k. Since there is no risk of confusion, we omit the dependence on k in the notation. For any square matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ we denote by $\mathbf{A}^+$ its Moore-Penrose pseudoinverse and by $\|\mathbf{A}\|$ its spectral norm defined by

$$\|\mathbf{A}\| = \max_{\|\boldsymbol{v}\|_2=1} \|\mathbf{A}\boldsymbol{v}\|_2 \tag{4.2.1}$$

We use boldface italic letters for vectors and boldface letters for matrices. Throughout the manuscript, the index j will be used for referring to p features, whereas the index i will refer to the observations ($i \in [n]$ or $i \in [N]$). For any set of indices $J \subseteq [p]$ and any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p)^\top \in \mathbb{R}^p$, we define $\boldsymbol{\beta}_J$ as the p-dimensional vector whose j-th coordinate equals $\boldsymbol{\beta}_j$ if $j \in J$ and 0 otherwise. We denote the cardinality of any $J \subseteq [p]$ by $|J|$. Also, we set $\text{supp}(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_j \neq 0\}$.

In particular, whenever $f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^\star$, we set $J^\star = \mathrm{supp}(\boldsymbol{\beta}^\star)$ and $s^\star = |J^\star|$. For $J \subseteq [p]$ and $c > 0$, we introduce the compatibility constants

$$\kappa_{\mathbf{A}}(J, c) = \inf \left\{ \frac{c^2 |J| \ \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2}{(c\|\boldsymbol{v}_J\|_1 - \|\boldsymbol{v}_{J^c}\|_1)^2} : \boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}_{J^c}\|_1 < c\|\boldsymbol{v}_J\|_1 \right\} \tag{4.2.2}$$

and

$$\bar{\kappa}_{\mathbf{A}}(J, c) = \inf \left\{ \frac{|J| \ \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2}{\|\boldsymbol{v}_J\|_1^2} : \boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}_{J^c}\|_1 < c\|\boldsymbol{v}_J\|_1 \right\}. \tag{4.2.3}$$

One easily checks that these two constants are of the same order of magnitude in the sense that

$$\frac{\bar{c}^2}{(\bar{c} + c)^2} \kappa_{\mathbf{A}}(J, \bar{c} + c) \leq \bar{\kappa}_{\mathbf{A}}(J, c) \leq \kappa_{\mathbf{A}}(J, c)$$

for every $c, \bar{c} > 0$. These constants are slightly larger[4] than the restricted eigenvalues (Bickel et al., 2009) defined by

$$\kappa_{\mathbf{A}}^{\mathrm{RE}}(J, c) = \inf \left\{ \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2 : \ \|\boldsymbol{v}_{J^c}\|_1 \leq c\|\boldsymbol{v}_J\|_1 \text{ and } \|\boldsymbol{v}_J\|_2 = 1 \right\}.$$

For more details, we refer the reader to van de Geer and Bühlmann (2009).

4.3 Brief overview of related work

The material of this paper builds on the shoulders of giants and this section aims at providing a unified overview of some of the most relevant results in our setting, without having the ambition of being exhaustive. For each of the selected papers, we will discuss its strengths and limitations in relation with the results presented further in this work.

Some recent results, obtained in the context of matrix regression, can be specialized to our problem and should be put in perspective with our contribution. For instance, a large part of Chapter 9 in (Koltchinskii, 2011) is devoted to the problem of assessing the off-sample excess risk of the trace-norm penalized empirical risk minimizer in the setting of trace regression with random design. One can arguably consider that setting as an extension of the random design regression problem by restricting attention to the set of diagonal matrices. Then the estimator studied in Koltchinskii (2011) coincides with the lasso estimator (4.1.7). With our notations, the main result of Chapter 9 in (Koltchinskii, 2011) reads as follows.

Theorem 4.3.1 (Theorem 9.3 in Koltchinskii, 2011)**.** *Assume that Assumptions 4.1 and 4.1*

[4] We recall here that a larger compatibility constant provides a better risk bound.

hold. Then there exist universal positive constants c_1 and c_2 such that, if

$$\lambda \geq c_1 B_X \max\left\{\frac{B_Y \log(2p/\delta)}{n}, \Big(\frac{B_Y \log(2p/\delta)}{n}\Big)^{1/2}\right\}$$

for some $\delta \in (0,1)$, the estimator (4.1.7) satisfies,

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{2\mathcal{E}(f_{\boldsymbol{\beta}}) + c_2\left[\frac{\|\boldsymbol{\beta}\|_0\lambda^2}{\bar{\kappa}_{\boldsymbol{\Sigma}}(\mathrm{supp}(\boldsymbol{\beta}),5)} + \Big(\|\boldsymbol{\beta}\|_1 \vee \frac{q(\lambda)}{\lambda}\Big)^2 \frac{\log(k/\delta)\log(n)}{n} + \frac{1}{n}\right]\right\}, \quad (4.3.1)$$

with probability larger than $1-\delta$, where

$$k = \log(n \vee p \vee B_Y) \vee |\log(2\lambda)| \vee 2 \quad \textit{and} \quad q(\lambda) = \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\big(\mathcal{E}(f_{\boldsymbol{\beta}}) + 2\lambda\|\boldsymbol{\beta}\|_1\big).$$

This result can be briefly compared to the risk bound in (4.1.11). The main advantages of this result is that (a) it is established under much weaker assumptions on the boundedness of the random variables $\boldsymbol{X}$ and Y than those of Assumption 4.1, (b) it holds not only for the vector regression but also for matrix regression, (c) it contains no restriction on the sample size and (d) it involves the compatibility constant of the population covariance matrix $\boldsymbol{\Sigma}$. On the negative side, the oracle inequality in 4.3.1 is not sharp since the factor in front of $\mathcal{E}(f_{\boldsymbol{\beta}})$ is not equal to one and, more importantly, the rate of convergence of the remainder term is sub-optimal in most situations. Indeed, if the best linear predictor corresponds to an s-sparse vector the nonzero entries of which are of the same order, then the term $\|\boldsymbol{\beta}\|_1^2 \log(k/\delta)\log(n)/n$, present in the right hand side, is of order $s^2\log(n)\log\log(n+p)/n$, whereas the remainder term in (4.1.11) is of smaller order $s\log(p)/n$.

On a related note, Koltchinskii et al. (2011a) establish sharp oracle inequalities for the trace-norm penalized least-squares estimator in the problem of matrix estimation and completion under low rank assumption. Using our notation, Theorem 2 in (Koltchinskii et al., 2011a) yields the following result.

Theorem 4.3.2 (Koltchinskii et al., 2011a)**.** *Assume that the matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ is known and let $\widehat{\boldsymbol{\beta}}$ be as in (4.1.9) with $\mathbf{A} = \boldsymbol{\Sigma}^{1/2}$. Suppose in addition that Assumption 4.1 holds and that, for $\delta \in (0,1)$,*

$$\lambda \geq 4B_Y\Big(\frac{\log(p/\delta)}{n}\Big)^{1/2}\left[1 + \frac{B_X}{3}\Big(\frac{\log(p/\delta)}{n}\Big)^{1/2}\right].$$

Then, with probability larger than $1-\delta$*, we have*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{J\subseteq[p]} \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda \|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2|J|}{4\kappa_{\boldsymbol{\Sigma}}(J,3)} \right\}. \tag{4.3.2}$$

The original result (Koltchinskii et al., 2011a, Theorem 2) is slightly different from the aforestated one. In particular, it is expressed in terms of the restricted eigenvalue constant with respect to the population covariance matrix $\boldsymbol{\Sigma}$. However, all these differences imply only minor modifications in the proofs. 4.3.2 is very similar to the risk bounds that we establish in the present work, but has the obvious shortcoming of requiring the covariance matrix $\boldsymbol{\Sigma}$ to be known. In fact, this corresponds to the situation in which infinitely many unlabeled feature vectors $\boldsymbol{X}_{n+1}, \boldsymbol{X}_{n+2}, \ldots$ are available, that is $N = +\infty$. To some extent, one of the purposes of the present work is to provide risk bounds analogous to the result of 4.3.2 but valid for a broad range of values of N. Note that the choice of the tuning parameter λ advocated by all the aforementioned results is of the same order of magnitude.

To the best of our knowledge, the only paper establishing risk bounds for a transductive version of the lasso is (Alquier and Hebiri, 2012). In that paper, the authors considered the problem of transductive learning in a linear model $Y = \boldsymbol{X}^\top\boldsymbol{\beta}^\star + \xi$ under the sparsity constraint. The estimator they studied is slightly different from ours and is defined by

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}\widehat{\boldsymbol{\Sigma}}_{\text{lab}}^{+}\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}. \tag{4.3.3}$$

For the predictor $f_{\widehat{\boldsymbol{\beta}}}$ based on this estimator, the authors established the following risk bound.

Theorem 4.3.3 (Theorems 4.3 and 4.4 in Alquier and Hebiri, 2012)**.** *Assume that for some* $\boldsymbol{\beta}^\star \in \mathbb{R}^p$*, the conditional distribution of* $\xi := Y - \boldsymbol{X}^\top\boldsymbol{\beta}^\star$ *given* $\boldsymbol{X}$ *is Gaussian* $\mathcal{N}(0,\sigma^2)$*. Let* $\mathscr{E}_1$ *be the event "all the unlabeled features* $\{\boldsymbol{X}_{n+i} : i \in [N-n]\}$*, belong to the linear span of the labeled features* $\{\boldsymbol{X}_i : i \in [n]\}$*" and let* $\delta \in (0,1)$*. Denote by* $a_{n,N,p}$ *the harmonic mean of the diagonal entries of the matrix* $\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}\widehat{\boldsymbol{\Sigma}}_{\text{lab}}^{+}\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}$*. Then the estimator* (4.3.3) *with* $\lambda = \sigma\sqrt{(2/n)a_{n,N,p}\log(p/\delta)}$ *satisfies*

$$\mathbb{P}\left(\mathcal{E}_{\text{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{72\sigma^2 a_{n,N,p}}{\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}}(J^\star,3)} \cdot \frac{s^\star\log(p/\delta)}{n} \;\middle|\; \mathbf{X}_{\text{all}} \right) \geq 1-\delta \quad \text{on} \quad \mathscr{E}_1.$$

This result is close in spirit to the result that we establish in this work in the setting of transductive learning. Note however that there are three main differences. First, we do not confine our study to the well-specified situation in which the Bayes predictor is linear, $f^\star(\boldsymbol{x}) =$

$\boldsymbol{x}^\top \boldsymbol{\beta}^\star$ for every $\boldsymbol{x} \in \mathbb{R}^p$, with a sparse vector $\boldsymbol{\beta}^\star$. Second, we avoid the unpleasant restriction that the unlabeled features are linear combinations of labeled features. Third, we replace the factor $a_{n,N,p}$—which may be quite large—by a more tractable quantity. This being said, the result of Alquier and Hebiri (2012)—in contrast with our results—does not require the unlabeled features to be drawn from the same distribution as the labeled features.

We also review a recent result from (Lecué and Mendelson, 2016a). In that paper, the authors consider the isotropic case $\boldsymbol{\Sigma} = \mathbf{I}_p$, where $\mathbf{I}_p$ stands for the $p \times p$ identity matrix, but impose only weak assumptions on the moments of the noise. Translated to our notations, their result can be formulated as follows.

Theorem 4.3.4 (Theorem 1.3 in Lecué and Mendelson, 2016a)**.** *Let Assumption 4.1 be satisfied and let* $\boldsymbol{\Sigma} = \mathbf{I}_p$*. Let* $f_{\bar{\boldsymbol{\beta}}}$ *be the best linear approximation in* $L^2(P_X)$ *of the regression function* $f^\star$*, that is* $\bar{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{E}(f_{\boldsymbol{\beta}})$*. Let* $\delta \in (0,1)$ *be a prescribed tolerance level. There are three constants* $c_1(\delta)$*,* $c_2(\delta, B_X)$ *and* $c_3(\delta, B_X)$ *such that, if* $\bar{\boldsymbol{\beta}}$ *is nearly* s*-sparse in the sense that*[5]

$$\sum_{j=s+1}^{p} |\bar{\beta}|_{(j)} \leq c_1(\delta) B_Y s \Big(\frac{\log(2p)}{n} \Big)^{1/2}$$

and λ *is chosen by* $\lambda = c_2(\delta, B_X) B_Y \big(\frac{\log(2p)}{n} \big)^{1/2}$*, then with probability at least* $1 - \delta$ *the lasso estimator satisfies*

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \mathcal{E}(f_{\bar{\boldsymbol{\beta}}}) + c_3(\delta, B_X) B_Y^2 \, \frac{s \log(2p)}{n}. \tag{4.3.4}$$

The principal strength of this result is that it is valid under a very weak assumption on the tails of the noise, but it has the shortcoming of requiring the minimizer of the excess risk to be nearly s-sparse with a quite precise upper bound on the authorized non-sparsity bias. From this point of view, an upper bound of the form (4.1.11) provides more information on the robustness of the prediction rule with respect to the model mis-specification.

The proofs of the results above assess the off-sample prediction error rate of the lasso by using direct arguments. An alternative approach (adopted, for example, in Raskutti et al., 2010a; Koltchinskii, 2011; Oliveira, 2013; Rudelson and Zhou, 2013) consists in taking advantage of the in-sample risk bounds in order to assess the off-sample excess risk. In short, by means of nowadays well-known techniques (developed in Bickel et al., 2009; Juditsky and Nemirovski, 2011; Bühlmann and van de Geer, 2011; Belloni et al., 2014; Dalalyan et al., 2014b, for instance)

[5] We denote by $|\bar{\beta}|_{(j)}$ the j-th largest value of the sequence $|\bar{\beta}_1|, \ldots, |\bar{\beta}_p|$, so that $|\bar{\beta}|_{(1)} \geq \cdots \geq |\bar{\beta}|_{(p)}$.

for a well-specified model[6], an upper bound on the in-sample risk,

$$\frac{1}{n}\|\mathbf{X}_{\text{lab}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 = \|\widehat{\boldsymbol{\Sigma}}_{\text{lab}}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2,$$

is obtained along with proving that the vector $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$ belongs to the dimension-reduction cone appearing in the definition of the compatibility constant. Then, using suitably chosen concentration arguments, it is shown that (with high probability) the compatibility constant $\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{lab}}}(J^\star, c)$ of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{lab}}$ is lower bounded by a (multiple of a) compatibility constant $\kappa_{\boldsymbol{\Sigma}}(J^\star, c')$ of the population covariance matrix, provided that the sparsity s is of order $n/\log(p)$. The main conceptual differences between the aforementioned papers are in the conditions on the random vectors $\boldsymbol{X}_i$. In (Raskutti et al., 2010a), it is assumed that the $\boldsymbol{X}_i$'s are Gaussian. In Rudelson and Zhou (2013) and Theorem 9.2 in Koltchinskii (2011), sub-Gaussian and bounded designs are considered, whereas only a bounded moment condition is required in Oliveira (2013). We will not reproduce their results here because (a) they do not allow to account for the robustness to the model mis-specification and, to a lesser extent, (b) the constants involved in the bounds are not explicit.

4.4 Risk bounds in transductive setting

We first consider the case of transductive learning. From an intuitive point of view, this case is simpler than the case of semi-supervised learning since a prediction needs to be carried out only for the features in $\mathcal{D}_{\text{unlabeled}}$. Indeed, recall from (4.1.5) that in this context, the excess risk of the linear predictor $f_{\boldsymbol{\beta}}$ is defined by

$$\mathcal{E}_{\text{TL}}(f_{\boldsymbol{\beta}}) = \frac{1}{N-n}\sum_{i=n+1}^{N}\big(\boldsymbol{X}_i^\top\boldsymbol{\beta} - f^\star(\boldsymbol{X}_i)\big)^2$$

and the suitably adapted lasso estimator is given by choosing $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}$ in (4.1.9), that is

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\|\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1\Big\}.$$

Note here that the role of the term $\frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\text{lab}}$ is to estimate the term $\frac{2}{N-n}\sum_{i=n+1}^{N} f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i^\top$, which appears after developing the square in the excess risk. Since the latter belongs to the image of the matrix $\mathbf{X}_{\text{unlab}}$, one can slightly improve the estimator by projecting onto the

[6]This means that for a sparse vector $\boldsymbol{\beta}^\star$, it holds that $f^\star = f_{\boldsymbol{\beta}^\star}$.

subspace of $\mathbb{R}^p$ spanned by the unlabeled vectors $\mathbf{X}_i$. This amounts to replacing the term $\boldsymbol{Y}^\top \mathbf{X}_{\text{lab}} \boldsymbol{\beta}$ by $\boldsymbol{Y}^\top \mathbf{X}_{\text{lab}} \Pi_{\text{unlab}} \boldsymbol{\beta}$, where Π_{unlab} stands for the orthogonal projector in $\mathbb{R}^p$ onto $\text{Span}(\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_N)$. However, from a theoretical point of view, this modification has no impact on the risk bound stated below. That is why we confine our attention to the lasso estimator that does not use this modification.

Theorem 4.4.1. *Let Assumptions 4.1 and 4.1 be fulfilled. Define $n_\star = n \wedge (N-n)$ and assume that, for a given $\delta \in (0,1)$, the tuning parameter λ satisfies*

$$\lambda \geq 4B_Y \Big(\frac{\log(2p/\delta)}{n_\star} \Big)^{1/2} \Big[1 + \frac{B_X}{3} \Big(\frac{\log(2p/\delta)}{n_\star} \Big)^{1/2} \Big]. \tag{4.4.1}$$

Then, with probability at least $1-\delta$, the predictor $f_{\widehat{\boldsymbol{\beta}}}$ satisfies

$$\mathcal{E}_{\text{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ J \subseteq [p]}} \Big\{ \mathcal{E}_{\text{TL}}(f_{\boldsymbol{\beta}}) + 4\lambda \|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{4\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}}(J,3)} \Big\}. \tag{4.4.2}$$

A few comments are in order. First, 4.4.1 holds for any pair of integers n and N larger than 1. However, it is especially relevant when the number $N-n$ of unlabeled features is larger than the number n of labeled ones. As already mentioned, this kind of situation is frequent in applications where the labeling procedure is expensive. In this case, $n_\star = n$ and 4.4.1 takes the same form as (4.1.11) with the notable advantage that the size of the unlabeled sample does not need to be of larger order than the dimension p. Let us present a few implications of this result in the well-specified case.

Well-specified case. Recall that the well-specified case refers to the situation where there exists $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that the Bayes predictor $f^\star$ satisfies $f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^\star$, P_X-almost surely. In this case, the excess risk of a predictor $f_{\boldsymbol{\beta}}$ can be written as $\mathcal{E}_{\text{TL}}(f_{\boldsymbol{\beta}}) = \|\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2$. In this form, the technical tractability of the transductive learning problem appears clearly since the matrix $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\text{unlab}}^{1/2}$ used in the definition of the estimator $\widehat{\boldsymbol{\beta}}$ coincides with the one appearing in the excess loss. As we shall see later, this is indeed not the case for semi-supervised learning. Now, the choice of $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ and $J = J^\star$ in the right hand side of inequality (4.4.2) yields

$$\mathcal{E}_{\text{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{9\lambda^2 s^\star}{4\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{unlab}}}(J^\star, 3)}.$$

The choice of λ provided by the right hand side of inequality (4.4.1), along with the condition $n_\star \geq B_X^2 \log(2p/\delta)$, leads to the bound

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{64 B_Y^2}{\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}}(J^\star, 3)} \cdot \frac{s^\star \log(p/\delta)}{n_\star},$$

with probability at least $1-\delta$. Comparing our result with that of Alquier and Hebiri (2012) (cf. 4.3.3 above), we can note that 4.4.1 holds without the assumption that the unlabeled features belong to the linear span of the labeled ones. On the other hand, Alquier and Hebiri (2012) do not require the labeled and the unlabeled features to be drawn from the same distribution.

4.5 Risk bounds in semi-supervised setting

We now turn to the more challenging problem of semi-supervised learning. In this subsection, we first consider the well-specified setting in which the Bayes predictor $f^\star$ is linear. We start with risk bounds that hold with a probability close to one. Such bounds are often termed *in deviation* as opposed to those holding *in expectation*.

Well-specified case. We assume here that

$$f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^\star, \quad P_X\text{-almost surely.} \tag{4.5.1}$$

In this context, the excess risk of the linear predictor $f_{\boldsymbol{\beta}}$, defined in (4.1.4), becomes $\mathcal{E}(f_{\boldsymbol{\beta}}) = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2$. This setting is more restrictive than the mis-specified setting considered below, but it has the advantage of allowing us to obtain risk bounds that are small even if the sample size N is not necessarily larger than the dimension p. The next result assesses the performance of the predictor $f_{\widehat{\boldsymbol{\beta}}}$ where

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}^{1/2} \boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} \boldsymbol{\beta} + 2\lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{4.5.2}$$

corresponding to the choice $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}^{1/2}$ in (4.1.9). In the next result, we set

$$\kappa_{\mathbf{A}}^{\mathrm{RE}}(s, c) = \min_{J \subseteq [p] : |J| \leq s} \kappa_{\mathbf{A}}^{\mathrm{RE}}(J, c),$$

where the restricted eigenvalue $\kappa_{\mathbf{A}}^{\mathrm{RE}}(J, c)$ is defined in 4.2.

Theorem 4.5.1. *Let Assumptions 4.1, 4.1 and (4.5.1) be fulfilled. Let $\delta \in (0,1)$ be a tolerance level and let the tuning parameter λ satisfy*

$$\lambda \geq 4B_Y\Big(\frac{\log(4p/\delta)}{n}\Big)^{1/2}\Big[1+\frac{B_X}{2}\Big(\frac{\log(4p/\delta)}{n}\Big)^{1/2}\Big].$$

With probability at least $1-\delta$, it holds

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \left(\frac{6\lambda s^\star}{\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star,3)}\right)^2 \bigwedge \frac{9\|\boldsymbol{\Sigma}\|\lambda^2 s^\star}{\kappa^{\mathrm{RE}}_{\widehat{\boldsymbol{\Sigma}}_N}(s^\star,3)^2}. \tag{4.5.3}$$

In addition, if the overall sample size N is such that $16 s^\star B_X^2\sqrt{2\log(4p^2/\delta)} \leq \bar{\kappa}_{\boldsymbol{\Sigma}}(J^\star,3)\sqrt{N}$ then, with probability at least $1-\delta$, the predictor $f_{\widehat{\boldsymbol{\beta}}}$ satisfies the inequality

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \frac{9\lambda^2 s^\star}{\bar{\kappa}_{\boldsymbol{\Sigma}}(J^\star,3)}. \tag{4.5.4}$$

This theorem provides three different risk bounds, all of them being valid for the same choice of the tuning parameter λ, that clearly show the benefits of using unlabeled data. The first two bounds are stated in 4.5.3. They share the common feature of depending on a characteristic (compatibility constant or restricted eigenvalue) of the sample covariance matrix. The latter is computed using both labeled and unlabeled data. For large values of N, it is more likely that these characteristics are bounded away from zero than those of the sample covariance matrix based on the labeled data only. In the asymptotic setting where $s^\star$ goes to infinity with the sample size and the dimension, the second term in the right hand side of 4.5.3 is of smaller order than the first one and is rate optimal, provided that the restricted eigenvalue is lower bounded by a fixed positive constant. However, for finite and small values of $s^\star$ the first term in the right hand side of 4.5.3 might be smaller than the second term.

This being said, it might be more insightful to look at the non random upper bounds on the excess risk as the one stated in 4.5.4. It basically tells us that if the overall sample size is larger than a multiple of $(s^\star)^2 \log p$, then the off-sample prediction risk of the semi-supervised lasso estimator achieves the fast rate $\frac{s^\star \log p}{n}$. Note that if we use only the labeled data points, the best known results—as recalled in 4.2 above—provide the fast rate when n is larger than a multiple of $s^\star \log p$. Thus, if N is of the same order as n, our result above is not the sharpest possible, but it has the advantage of being easy to prove and, nevertheless, demonstrating the gain of using the unlabeled data. In particular, the proof of results providing the fast rate under the condition $n \geq C s^\star \log(p)$, for some $C > 0$, involve the important step of lower bounding the

compatibility constant of the sample covariance matrix by its population counterpart. This step uses concentration arguments which are often tedious and come with implicit (or unreasonably large) constants. Instead, our proof makes use of much simpler tools essentially boiling down to the classical Bernstein inequality and leads to explicit and small constants.

Mis-specified case. Mathematical analysis of the semi-supervised lasso under mis-specification is more involved, since it requires careful control of the bias terms corresponding to the non-linearity and the non-sparsity of the model. We first state results providing risk bounds in deviation, then state their counterpart in expectation.

Theorem 4.5.2. *Let Assumptions 4.1 and 4.1 be fulfilled. Fix $J \subseteq [p]$ and $\delta \in (0,1)$. Suppose in addition that*

$$N \geq 18 B_X^2 p \|\boldsymbol{\Sigma}^{-1}\| \log(3p/\delta) \tag{4.5.5}$$

and

$$\lambda \geq 8 B_X B_Y \Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2} \Big[1 + \frac{B_X}{3}\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\Big]. \tag{4.5.6}$$

Then the semi-supervised lasso estimator $\widehat{\boldsymbol{\beta}}$ defined in (4.5.2) above satisfies

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda \|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{2\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}}(J,3)} \Big\}, \tag{4.5.7}$$

with probability larger than $1-\delta$.

The novelty of 4.5.2 lies in the semi-supervised nature of the estimator (4.5.2), which involves all the unlabeled features through the matrix $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}^{1/2}$ in 4.1.9. In particular, 4.5.2 quantifies the natural intuition according to which, if N is large enough, the matrix $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}^{1/2}$ is a good estimator of $\boldsymbol{\Sigma}$ and a result similar to 4.3.2 should hold. As mentioned in the introduction, an attractive feature of the upper bound in 4.5.7 is that it is of the same form as the recent oracle inequalities established in the case of fixed design regression (see, for instance, Dalalyan et al., 2014b; Pensky, 2014, and the references therein) and quantify in an easy-to-understand manner the error terms accounting for the non-linearity and the non-sparsity of the true regression function $f^\star$.

The minimal number N of features satisfying (4.5.5) depends on $\|\boldsymbol{\Sigma}^{-1}\| = \lambda_{\min}^{-1}(\boldsymbol{\Sigma})$, reflecting the fact that the quality of approximation of the identity matrix $\mathbf{I}_p$ by $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}\boldsymbol{\Sigma}^{-1/2}$ depends on $\|\boldsymbol{\Sigma}^{-1}\|$. One can remark that under constraint (4.5.5), the lowest eigenvalue of the sample covariance matrix is close to its population counterpart (Vershynin, 2010) and provides a simple

lower bound on the compatibility constant $\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{all}}}(J,3)$ appearing in 4.5.7. These considerations lead to the following corollary.

Corollary 4.5.1. *Under the conditions of 4.5.2, with probability at least* $1-\delta$,

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \le \inf_{J\subseteq[p]} \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \tfrac{27\|\boldsymbol{\Sigma}^{-1}\|}{4}\lambda^2|J| \Big\}. \tag{4.5.8}$$

Let us also mention that the factor $B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|$ present in the right hand side of 4.5.5 is an upper bound on the norm $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|_2^2$ under assumption 4.1. Under additional assumptions on the support of the features $\boldsymbol{X}_i$, this expression may be replaced by a smaller one leading thus to a relaxation of condition (4.5.5).

Sharp oracle inequality in expectation. All the previously stated results assert that the lasso estimator has a small prediction error on an event of overwhelming probability. However, in these results, the choice of the tuning parameter λ and, therefore, the final predictor $f_{\widehat{\boldsymbol{\beta}}}$, depends on the prescribed level of tolerance. A consequence of this dependence is that one can not integrate out the bounds in deviation in order to get a bound in expectation. This is probably one of the reasons why the bounds in expectation for the lasso are scarce in the literature. To fill this caveat, we state below a risk bound in expectation that can be easily deduced from the bounds in deviation.

Theorem 4.5.3. *Let Assumptions 4.1 and 4.1 be fulfilled. Suppose that the overall sample size is such that* $N \ge 18B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(3pN^2)$. *Then, for the tuning parameter*

$$\lambda = 8B_X B_Y \Big(\frac{\log(6pN^2)}{n}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{\log(6pN^2)}{n}\Big)^{1/2}\Big] \tag{4.5.9}$$

the semi-supervised lasso estimator $\widehat{\boldsymbol{\beta}}$ *defined in* (4.5.2) *above satisfies*

$$\mathbb{E}[\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}})] \le \inf_{J\subseteq[p]} \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{27\|\boldsymbol{\Sigma}^{-1}\|}{4}\lambda^2|J| \Big\} + \frac{2B_Y^2}{N^2} + \frac{B_Y^2}{2^7 n\log^2(6pN^2)}. \tag{4.5.10}$$

The proof of this theorem is postponed to 4.7.2. The bound above is not optimal in terms of its dependence on N. In particular, it blows up when N goes to infinity and all the other parameters are fixed. However, this divergence is only logarithmic in N. The dominating term in the risk bound above is (at least in the well specified setting) of the order $\lambda^2|J| \asymp \frac{s\log(pN)}{n}$.

4.6 Conclusion

We have reviewed some recent results on the prediction accuracy of the lasso in the problem of regression with random design and have proposed their extensions to the setting where the labels of some data points are not available. Theoretical guarantees stated in previous sections are formulated as oracle inequalities that allow us to compare the excess risk of a suitable adaptation of the lasso to the best possible (nearly) sparse prediction function. We have opted for considering only those risk bounds that provide the fast rate and are valid under some conditions on the design such as the restricted eigenvalue condition or the compatibility condition. Some of the established upper bounds involve the compatibility constant of the sample covariance matrix. Using results on random matrices (Rudelson and Zhou, 2013; Oliveira, 2013; Bah and Tanner, 2014) they can be further worked out to get deterministic upper bounds. However, the evaluation of the restricted eigenvalues and related quantities of the random covariance-type matrices is a dynamically evolving research area and we expect that new advances will be made in near future.

The main high level message of the contributions of this paper is that one can take advantage of the unlabeled sample for improving the prediction accuracy of the lasso. Roughly speaking, if the size of the unlabeled sample is larger than the ambient dimension, then the modified lasso predictor has a prediction risk that converges to zero at the optimal rate even if the sample covariance matrix based only on the labeled sample does not satisfy the compatibility or the restricted eigenvalue condition. However, it should be acknowledged that when the model is well specified (that is there exists a sparse linear combination of the features with an extremely low approximation error) and the population covariance matrix is well-conditioned, then the original lasso might perform as well as, or even better than, the modified lasso proposed in this work. Therefore, one can conclude that the use of the unlabeled sample improves on the robustness of the lasso to the model mis-specification.

We would like also to emphasize that, pursuing pedagogical goals, we have restricted our attention to the simple case of bounded feature vectors and bounded labels. All the proofs presented in this paper are based on elementary arguments and are fairly simple. Using more involved arguments, they can be carried over the case of sub-Gaussian design and labels. It would be interesting to explore their extensions to other settings such as regression with structured sparsity, low rank matrix regression or matrix completion, *etc.*

4.7 Proofs

We start with a general result that holds for the penalized least squares predictor with arbitrary convex penalty. This result is of independent interest. It generalizes the corresponding result of (Koltchinskii et al., 2011a) established for the matrix trace-norm penalties. The proof that we present here is different from the one in (Koltchinskii et al., 2011a) in that it does not rely on the precise form of the sub-differential of the penalty function.

Lemma 4.7.1. *Let $n, p \geq 1$. Let $\mathrm{pen} : \mathbb{R}^p \to \mathbb{R}$ be any convex function and $\widehat{\boldsymbol{\beta}}$ be defined by*

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} \boldsymbol{\beta} + \mathrm{pen}(\boldsymbol{\beta}) \right\}, \tag{4.7.1}$$

where $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$ and $\mathbf{X}_{\mathrm{lab}} \in \mathbb{R}^{n \times p}$. Then, for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 \leq \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\widehat{\boldsymbol{\beta}}) - \|\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2. \tag{4.7.2}$$

Proof. Let us introduce the function $\Phi(\boldsymbol{\beta}) = \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \frac{2}{n} \boldsymbol{Y}^\top \mathbf{X}_{\mathrm{lab}} \boldsymbol{\beta} + \mathrm{pen}(\boldsymbol{\beta})$ for every $\boldsymbol{\beta} \in \mathbb{R}^p$, so that $\widehat{\boldsymbol{\beta}}$ is a minimum point of Φ. Since the latter is a convex function, we know that the zero vector $\mathbf{0}_p$ of $\mathbb{R}^p$ belongs to the sub-differential $\partial\Phi(\widehat{\boldsymbol{\beta}})$ of Φ at $\widehat{\boldsymbol{\beta}}$. For all $\boldsymbol{\beta} \in \mathbb{R}^p$, let

$$\psi(\boldsymbol{\beta}) = \|\mathbf{A}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2, \qquad \bar{\Phi}(\boldsymbol{\beta}) = \Phi(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta}). \tag{4.7.3}$$

The function ψ is proper and convex. It is also differentiable on $\mathbb{R}^p$ and the sub-differential of ψ at $\widehat{\boldsymbol{\beta}}$ is reduced to its gradient at $\widehat{\boldsymbol{\beta}}$, so that $\partial\psi(\widehat{\boldsymbol{\beta}}) = \{\nabla\psi(\widehat{\boldsymbol{\beta}})\} = \{\mathbf{0}_p\}$. The function $\bar{\Phi}$ defined on $\mathbb{R}^p$ is the sum of an affine function and the convex function pen, thus it is also convex. The functions $\psi, \bar{\Phi}$ are proper and convex, the function ψ is continuous on $\mathbb{R}^p$ so by the Moreau-Rochafellar Theorem,

$$\partial\Phi(\widehat{\boldsymbol{\beta}}) = \partial\psi(\widehat{\boldsymbol{\beta}}) + \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}}) = \{\mathbf{0}_p\} + \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}}) = \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}}). \tag{4.7.4}$$

Thus $\mathbf{0}_p \in \partial\bar{\Phi}(\widehat{\boldsymbol{\beta}})$, which can be rewritten as

$$\bar{\Phi}(\boldsymbol{\beta}) \geq \bar{\Phi}(\widehat{\boldsymbol{\beta}}), \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^p. \tag{4.7.5}$$

By adding $\psi(\boldsymbol{\beta})$ on both sides of the previous display, we obtain

$$\Phi(\boldsymbol{\beta}) \geq \Phi(\widehat{\boldsymbol{\beta}}) + \|\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2, \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^p. \tag{4.7.6}$$

Rearranging the terms of this inequality, we get the claim of the lemma. □

We will also repeatedly use the following result.

Lemma 4.7.2. *For any pair of vectors $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$, for any pair of scalars $\mu > 0$ and $\gamma > 1$, for any $p \times p$ symmetric matrix $\mathbf{A}$ and for any set $J \subseteq [p]$, the following inequality is true*

$$2\mu\gamma^{-1}\Big(\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\boldsymbol{\beta}'\|_1\Big) - \|\mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2 \leq 4\mu\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{(\gamma+1)^2\mu^2|J|}{\gamma^2\kappa_{\mathbf{A}^2}(J, c_\gamma)}, \quad (4.7.7)$$

where $c_\gamma = (\gamma+1)/(\gamma-1)$.

Proof. To ease notation, we set $\boldsymbol{u} = \boldsymbol{\beta} - \boldsymbol{\beta}'$. Using that $\|\boldsymbol{\beta}_J\|_1 - \|\boldsymbol{\beta}'_J\|_1 \leq \|\boldsymbol{u}_J\|_1$ and $\|\boldsymbol{\beta}_{J^c}\|_1 + \|\boldsymbol{\beta}'_{J^c}\|_1 \geq \|\boldsymbol{u}_{J^c}\|_1$, we obtain

$$\begin{aligned}
\|\boldsymbol{u}\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\boldsymbol{\beta}'\|_1 &= \|\boldsymbol{u}\|_1 + \gamma\big(\|\boldsymbol{\beta}_J\|_1 + \|\boldsymbol{\beta}_{J^c}\|_1 - \|\boldsymbol{\beta}'_J\|_1 - \|\boldsymbol{\beta}'_{J^c}\|_1\big) && (4.7.8)\\
&= \|\boldsymbol{u}\|_1 + 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + \gamma\big(\|\boldsymbol{\beta}_J\|_1 - \|\boldsymbol{\beta}'_J\|_1\big) - \gamma\big(\|\boldsymbol{\beta}'_{J^c}\|_1 + \|\boldsymbol{\beta}_{J^c}\|_1\big) && (4.7.9)\\
&\leq \|\boldsymbol{u}\|_1 + 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + \gamma\|\boldsymbol{u}_J\|_1 - \gamma\|\boldsymbol{u}_{J^c}\|_1 && (4.7.10)\\
&= 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + (\gamma+1)\|\boldsymbol{u}_J\|_1 - (\gamma-1)\|\boldsymbol{u}_{J^c}\|_1 && (4.7.11)\\
&= 2\gamma\|\boldsymbol{\beta}_{J^c}\|_1 + (\gamma+1)\big(\|\boldsymbol{u}_J\|_1 - c_\gamma^{-1}\|\boldsymbol{u}_{J^c}\|_1\big). && (4.7.12)
\end{aligned}$$

If $c_\gamma\|\boldsymbol{u}_J\|_1 < \|\boldsymbol{u}_{J^c}\|_1$, the claim of the lemma is straightforward. Otherwise, $\|\boldsymbol{u}_{J^c}\|_1 \leq c_\gamma\|\boldsymbol{u}_J\|_1$ and using the definition of the compatibility constant we get

$$\begin{aligned}
\frac{2\lambda(\gamma+1)}{\gamma}\big(\|\boldsymbol{u}_J\|_1 - c_\gamma^{-1}\|\boldsymbol{u}_{J^c}\|_1\big) - \|\mathbf{A}\boldsymbol{u}\|_2^2 &\leq \frac{2\lambda(\gamma+1)}{\gamma}\left(\frac{|J|\cdot\|\mathbf{A}\boldsymbol{u}\|_2^2}{\kappa_{\mathbf{A}^2}(J, c_\gamma)}\right)^{1/2} - \|\mathbf{A}\boldsymbol{u}\|_2^2 && (4.7.13)\\
&\leq \frac{(\gamma+1)^2\lambda^2|J|}{\gamma^2\kappa_{\mathbf{A}^2}(J, c_\gamma)}, \qquad \text{[by Cauchy-Schwarz]} && (4.7.14)
\end{aligned}$$

which completes the proof. □

To close this subsection of auxiliary results, we provide simple upper bounds on the quantiles of some random noise variables.

Proposition 4.7.1. *Let $m = N - n$ and $n_\star = n \wedge m$. Introduce the random vectors $\boldsymbol{\zeta}^{(1)} = \frac{1}{n}\sum_{i=1}^n Y_i\boldsymbol{X}_i - \mathbb{E}[Y\boldsymbol{X}]$,*

$$\boldsymbol{\zeta} = \frac{1}{n}\sum_{i=1}^n Y_i\boldsymbol{X}_i - \frac{1}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i \quad \text{and} \quad \bar{\boldsymbol{\zeta}} = \frac{1}{n}\sum_{i=1}^n Y_i\boldsymbol{X}_i - \frac{1}{N}\sum_{i=1}^N f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i. \quad (4.7.15)$$

Under Assumptions 4.1 and 4.1, and for any $\delta \in (0,1)$, each of the following inequalities

$$\|\boldsymbol{\zeta}^{(1)}\|_\infty \leq 2B_Y\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big] \tag{4.7.16}$$

$$\|\boldsymbol{\zeta}\|_\infty \leq 2B_Y\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big] \tag{4.7.17}$$

$$\|\bar{\boldsymbol{\zeta}}\|_\infty \leq 2B_Y\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big[1+\frac{B_X}{2}\Big(\frac{\log(2p/\delta)}{n}\Big)^{1/2}\Big] \tag{4.7.18}$$

holds with probability at least $1-\delta$.

Proof. We will only prove the inequality corresponding to $\boldsymbol{\zeta}$. The others being very similar are left to the reader. Denote $\boldsymbol{s}\mu = \mathbb{E}[Y\boldsymbol{X}] = \mathbb{E}[f^\star(\boldsymbol{X})\boldsymbol{X}] \in \mathbb{R}^p$, and introduce the random vectors

$$\boldsymbol{s}Z_i = \begin{cases} N(Y_i\boldsymbol{X}_i - \boldsymbol{s}\mu)/n, & i \in [n], \\ N(\boldsymbol{s}\mu - f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i)/m, & i \in [N]\setminus[n]. \end{cases}$$

The vectors $\boldsymbol{s}Z_i$ are independent, centered, bounded and satisfy

$$\boldsymbol{\zeta} = \frac{\boldsymbol{s}Z_1 + \cdots + \boldsymbol{s}Z_N}{N}.$$

Furthermore, Assumption 4.1 implies that $\|\boldsymbol{s}Z_i\|_\infty \leq 2NB_YB_X/n$ if $i \leq n$ and that $\|\boldsymbol{s}Z_i\|_\infty \leq 2NB_YB_X/m$ if $i > n$. One can also bound from above the variance of the j-th component Z_{ij} of $\boldsymbol{s}Z_i$ as follows. If $i \leq n$ then, in view of Assumptions 4.1 and 4.1, $\mathbb{E}[Z_{ij}^2] \leq (N/n)^2\mathbb{E}[Y_i^2X_{ij}^2] \leq (NB_Y/n)^2$. Similarly, if $i > n$ then $\mathbb{E}[Z_{ij}^2] \leq (NB_Y/m)^2$. Hence, we may easily deduce that, for all $j \in [p]$,

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}[Z_{ij}^2] \leq \frac{2NB_Y^2}{n_\star}.$$

Therefore, using the Bernstein inequality recalled in Proposition 4.7.4 of Appendix 4.7.3, for every $j \in [p]$ and every $\delta > 0$, we get that inequality

$$|\zeta_j| > 2B_Y\left(\frac{\log(2p/\delta)}{n_\star}\right)^{1/2} + \frac{2B_YB_X\log(2p/\delta)}{3n_\star} \tag{4.7.19}$$

holds with probability at most δ/p. The claim of Proposition 4.7.1 follows from the union bound. □

Remark 4.7.1. *One can easily check that the inequality $\mathbb{E}[Z_{ij}^2] \leq (NB_Y/n)^2$, for $i = 1,\ldots,n$, used in the previous proof can be replaced by $\mathbb{E}[Z_{ij}^2] \leq (NL_YB_X/n)^2$, where $L_Y = (\mathbb{E}[Y_i^2])^{1/2}$. This may lead to a better risk bound in the cases where the random variable Y_i is not well*

concentrated around its average value.

We are now in a position to prove the main theorems of this paper.

4.7.1 Proof of 4.4.1

The proof of Theorem 4.4.1 follows directly from 4.7.1 and 4.7.2 below. For simplicity, the parameter $\gamma > 1$ introduced in Proposition 4.7.2 is fixed at the value $\gamma = 2$ in 4.4.1.

Proposition 4.7.2. *Let $\boldsymbol{\zeta}$ be as in 4.7.1. For any $\gamma > 1$, we set $c_\gamma = (\gamma+1)/(\gamma-1)$. On the event $\mathscr{E} = \{\|\boldsymbol{\zeta}\|_\infty \leq \lambda/\gamma\}$, for every $\boldsymbol{\beta} \in \mathbb{R}^p$ and every $J \subseteq [p]$, we have*

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) \leq \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \frac{(\gamma+1)^2\lambda^2|J|}{\gamma^2\kappa_{\widehat{\boldsymbol{\Sigma}}_{\mathrm{unlab}}}(J, c_\gamma)}. \tag{4.7.20}$$

Proof. Along the proof, we will use for convenience the shorthand notations $m = N-n$ and $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}^{1/2}_{\mathrm{unlab}}$. First, notice that developing the square in the expression $\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \frac{1}{m}\sum_{i=n+1}^{N}\left(\boldsymbol{X}_i^\top\boldsymbol{\beta} - f^\star(\boldsymbol{X}_i)\right)^2$, we get

$$\mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - \left(\frac{2}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)\boldsymbol{X}_i^\top\right)\boldsymbol{\beta} + \frac{1}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)^2 \tag{4.7.21}$$

$$= \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + 2\boldsymbol{\zeta}^\top\boldsymbol{\beta} - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\mathrm{lab}}\boldsymbol{\beta} + \frac{1}{m}\sum_{i=n+1}^{n+m} f^\star(\boldsymbol{X}_i)^2. \tag{4.7.22}$$

This implies that for every $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) = \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + 2\boldsymbol{\zeta}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_{\mathrm{lab}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{4.7.23}$$

Using 4.7.1 with the convex penalty term $\mathrm{pen}(\boldsymbol{\beta}) = 2\lambda\|\boldsymbol{\beta}\|_1$, we deduce that, for every $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) \leq 2\boldsymbol{\zeta}^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) - \|\mathbf{A}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2. \tag{4.7.24}$$

On the event $\mathscr{E}$, note that $2\boldsymbol{\zeta}^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq 2\|\boldsymbol{\zeta}\|_\infty\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 \leq \frac{2\lambda}{\gamma}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1$, which leads to

$$2\boldsymbol{\zeta}^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq \frac{2\lambda}{\gamma}\left(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1\right). \tag{4.7.25}$$

Combining equations (4.7.24) and (4.7.25), we get that on the event $\mathscr{E}$, for every $\boldsymbol{\beta} \in \mathbb{R}^p$ and

every $J \subseteq [p]$,

$$\mathcal{E}_{\mathrm{TL}}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}_{\mathrm{TL}}(f_{\boldsymbol{\beta}}) \leq 2\lambda\gamma^{-1}\big(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 + \gamma\|\boldsymbol{\beta}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1\big) - \|\mathbf{A}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\|_2^2. \tag{4.7.26}$$

The claim of the proposition follows from 4.7.26 by applying 4.7.2 with $\mu = \lambda$. □

To conclude the proof of 4.4.1, it suffices to note that in view of 4.7.1, the probability of the event $\mathscr{E} = \{\|\boldsymbol{\zeta}\|_\infty \leq \lambda/\gamma\}$ is larger than $1 - \delta$ provided that

$$\lambda \geq 2\gamma B_Y \Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big[1 + \frac{B_X}{3}\Big(\frac{\log(2p/\delta)}{n_\star}\Big)^{1/2}\Big].$$

4.7.2 Proofs for the semi-supervised version of the lasso

We start this section by some arguments that are shared by the proofs of both theorems stated in 4.5. Let $J \subseteq [p]$ and let $\boldsymbol{\beta}$ be a minimizer of the right hand side of (4.5.7). Note in particular that $\boldsymbol{\beta}$ is a deterministic vector depending on the unknown distribution P of the data. In addition, if the model is well-specified and $J = J^\star$ then $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$. We will also use the notation $\boldsymbol{u} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and

$$\boldsymbol{\zeta}^{(1)} = \frac{1}{n}\sum_{i=1}^n Y_i \boldsymbol{X}_i - \mathbb{E}[Y\boldsymbol{X}] \quad \text{and} \quad \boldsymbol{\zeta}^{(2)} = \big(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}\big)\boldsymbol{\beta}. \tag{4.7.27}$$

Furthermore, to ease notation, we set $\widehat{\boldsymbol{\Sigma}}_N = \widehat{\boldsymbol{\Sigma}}_{\mathrm{all}}$, $\widehat{\boldsymbol{\Sigma}}_n = \widehat{\boldsymbol{\Sigma}}_{\mathrm{lab}}$, $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_N^{1/2}$. First, observe that the excess risk $\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \int_{\mathcal{X}} \big(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}} - f^\star(\boldsymbol{x})\big)^2 P_X(\mathrm{d}\boldsymbol{x})$ of the predictor $f_{\widehat{\boldsymbol{\beta}}}$ satisfies

$$\begin{aligned}\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) &= \int_{\mathcal{X}} \big\{(\boldsymbol{x}^\top\boldsymbol{u})^2 + 2\boldsymbol{u}^\top\boldsymbol{x}\big(\boldsymbol{x}^\top\boldsymbol{\beta} - f^\star(\boldsymbol{x})\big) + \big(\boldsymbol{x}^\top\boldsymbol{\beta} - f^\star(\boldsymbol{x})\big)^2\big\} P_X(\mathrm{d}\boldsymbol{x}) &(4.7.28)\\ &= \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 + 2\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\boldsymbol{u}^\top\mathbb{E}\left[\boldsymbol{X}f^\star(\boldsymbol{X})\right] + \mathcal{E}(f_{\boldsymbol{\beta}}). &(4.7.29)\end{aligned}$$

Next, notice that

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 = \boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} + \|\mathbf{A}\boldsymbol{u}\|_{\mathbf{2}}^{\mathbf{2}}, \tag{4.7.30}$$

and that

$$\begin{aligned}2\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} &= 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{\beta} + 2\boldsymbol{u}^\top\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\beta} &(4.7.31)\\ &= 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{\beta} + \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_{\mathbf{2}}^{\mathbf{2}} - \|\mathbf{A}\boldsymbol{u}\|_{\mathbf{2}}^{\mathbf{2}} - \|\mathbf{A}\boldsymbol{\beta}\|_{\mathbf{2}}^{\mathbf{2}}, &(4.7.32)\end{aligned}$$

where in the last line we have used the identity $2a^\top b = \|a+b\|_2^2 - \|a\|_2^2 - \|b\|_2^2$ with $a = \mathbf{A}\boldsymbol{u}$ and $b = \mathbf{A}\boldsymbol{\beta}$. Transforming 4.7.29 thanks to (4.7.30) and (4.7.32) we obtain

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) = \boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} + 2\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{\beta} + \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{A}\boldsymbol{\beta}\|_2^2 - 2\boldsymbol{u}^\top\mathbb{E}\left[Y\boldsymbol{X}\right] \quad (4.7.33)$$

$$= \boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} + 2\boldsymbol{u}^\top\boldsymbol{\zeta}^{(2)} + \|\mathbf{A}\widehat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{A}\boldsymbol{\beta}\|_2^2 + 2\boldsymbol{u}^\top\boldsymbol{\zeta}^{(1)} - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_n\boldsymbol{u}, \quad (4.7.34)$$

where we have used the identity $\mathbb{E}\left[Y\boldsymbol{X}\right] = \mathbb{E}\left[\boldsymbol{X}f^\star(\boldsymbol{X})\right]$ and the definitions of $\boldsymbol{\zeta}^{(1)}$ and $\boldsymbol{\zeta}^{(2)}$. Applying Lemma 4.7.1 with $\mathrm{pen}(\boldsymbol{\beta}) = 2\lambda\|\boldsymbol{\beta}\|_1$ and combining its result with (4.7.34), we arrive at

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \leq \underbrace{2\boldsymbol{u}^\top(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}) + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)}_{\mathbf{T}_1} + \underbrace{\boldsymbol{u}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} - \|\mathbf{A}\boldsymbol{u}\|_2^2}_{\mathbf{T}_2}. \quad (4.7.35)$$

Proof of 4.5.1.

As mentioned earlier, in the well-specified setting we have $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ and, therefore, $\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2$ and $\mathcal{E}(f_{\boldsymbol{\beta}^\star}) = 0$. Hence, (4.7.35) yields

$$2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \leq 2\boldsymbol{u}^\top\left(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\right) + 2\lambda(\|\boldsymbol{\beta}^\star\|_1 - \|\boldsymbol{\beta}^\star + \boldsymbol{u}\|_1). \quad (4.7.36)$$

Combining the duality inequality $|\boldsymbol{u}^\top\left(\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\right)| \leq \|\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\|_\infty\|\boldsymbol{u}\|_1$ with the following one $\|\boldsymbol{\beta}^\star\|_1 - \|\boldsymbol{\beta}^\star + \boldsymbol{u}\|_1 = \|\boldsymbol{\beta}^\star_{J^\star}\|_1 - \|\boldsymbol{\beta}^\star_{J^\star} + \boldsymbol{u}_{J^\star}\|_1 - \|\boldsymbol{u}_{(J^\star)^c}\|_1 \leq \|\boldsymbol{u}_{J^\star}\|_1 - \|\boldsymbol{u}_{(J^\star)^c}\|_1$, we infer from inequality (4.7.36) that on the event $\mathscr{E} = \left\{2\|\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\|_\infty \leq \lambda\right\}$, we have

$$2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \leq \lambda(3\|\boldsymbol{u}_{J^\star}\|_1 - \|\boldsymbol{u}_{(J^\star)^c}\|_1). \quad (4.7.37)$$

This implies that $\|\boldsymbol{u}_{(J^\star)^c}\|_1 \leq 3\|\boldsymbol{u}_{J^\star}\|_1$ and, therefore,

$$2\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star, 3)\|\boldsymbol{u}_{J^\star}\|_1^2 \leq 2s^\star\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \leq 3\lambda s^\star\|\boldsymbol{u}_{J^\star}\|_1. \quad (4.7.38)$$

This yields $\|\boldsymbol{u}_{J^\star}\|_1 \leq 3\lambda s^\star/(2\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star, 3))$ and, since $\max_{j,j'}|\boldsymbol{\Sigma}_{j,j'}| \leq 1$, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2 \leq \|\boldsymbol{u}\|_1 \leq 4\|\boldsymbol{u}_{J^\star}\|_1$, which implies that

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \leq \left(\frac{6\lambda s^\star}{\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_N}(J^\star, 3)}\right)^2. \quad (4.7.39)$$

On the other hand, if we denote by I the set of the $s^\star$ largest entries of the vector $|\boldsymbol{u}|$, inequality (4.7.37) implies that $2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 \le \lambda(3\|\boldsymbol{u}_I\|_1 - \|\boldsymbol{u}_{I^c}\|_1)$.

Therefore, using the definition of the restricted eigenvalue and similar arguments as above, we deduce that $\|\boldsymbol{u}_I\|_2 \le 3\lambda\sqrt{s^\star}/(2\kappa^{\mathrm{RE}}_{\widehat{\boldsymbol{\Sigma}}_N}(I,3))$. Furthermore, $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{u}_I\|_2^2 + \|\boldsymbol{u}_{I^c}\|_2^2 \le \|\boldsymbol{u}_I\|_2^2 + \|\boldsymbol{u}_{I^c}\|_\infty\|\boldsymbol{u}_{I^c}\|_1 \le \|\boldsymbol{u}_I\|_2^2 + (s^\star)^{-1}\|\boldsymbol{u}_I\|_1\|\boldsymbol{u}_{I^c}\|_1 \le \|\boldsymbol{u}_I\|_2^2 + 3(s^\star)^{-1}\|\boldsymbol{u}_I\|_1^2 \le 4\|\boldsymbol{u}_I\|_2^2$. This yields

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \|\boldsymbol{\Sigma}\|\cdot\|\boldsymbol{u}\|_2^2 \le 4\|\boldsymbol{\Sigma}\|\cdot\|\boldsymbol{u}_I\|_2^2 \le \frac{9\|\boldsymbol{\Sigma}\|\lambda^2 s^\star}{\kappa^{\mathrm{RE}}_{\widehat{\boldsymbol{\Sigma}}_N}(I,3)^2}\,. \tag{4.7.40}$$

Combining (4.7.39) and(4.7.40), we get the first claim of the theorem.

To get the second claim of the theorem, we go back to (4.7.37) and use the following inequalities:

$$\begin{aligned}
2\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 &= 2\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{u}\|_2^2 + 2\boldsymbol{u}^\top(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N)\boldsymbol{u} && (4.7.41)\\
&\le 3\lambda\|\boldsymbol{u}_{J^\star}\|_1 + 2\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty\|\boldsymbol{u}\|_1^2 && (4.7.42)\\
&\le 3\lambda\|\boldsymbol{u}_{J^\star}\|_1 + 32\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty\|\boldsymbol{u}_{J^\star}\|_1^2. && (4.7.43)
\end{aligned}$$

In the sequel, let us denote $\kappa = \bar{\kappa}_{\boldsymbol{\Sigma}}(J^\star,3)$ for brevity. Then, upper bounding the two instances of $\|\boldsymbol{u}_{J^\star}\|_1$ in (4.7.43) by $(s^\star\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2/\kappa)^{1/2}$, we infer that on $\mathscr{E}$,

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \frac{3\lambda\sqrt{s^\star}}{2\sqrt{\kappa}}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2 + \frac{16s^\star}{\kappa}\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2. \tag{4.7.44}$$

Dividing both sides by $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2$ (if this quantity vanishes then the claim of the theorem is obviously true) and after some algebra, we get the inequality

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2^2 \le \frac{9\lambda^2 s^\star\kappa}{4(\kappa - 16s^\star\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty)^2} \le \frac{9\lambda^2 s^\star}{\kappa}, \tag{4.7.45}$$

where the last inequality holds on the event $\mathscr{E}\cap\{32s^\star\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty \le \kappa\}$. In view of the union bound, Hoeffding's inequality and Assumption 4.1, we get for any $t>0$,

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty \ge t\right) \le p^2\max_{j,j'\in[p]}\mathbb{P}\left(|\sigma_{jj'}-\widehat{\sigma}_{jj'}|\ge t\right) \le 2p^2\exp\left(-2Nt^2/B_X^4\right), \tag{4.7.46}$$

where $\boldsymbol{\Sigma}=(\sigma_{ij})$ and $\widehat{\boldsymbol{\Sigma}}_N=(\widehat{\sigma}_{ij})$. Therefore, if

$$16s^\star B_X^2\left(\frac{2\log(4p^2/\delta)}{N}\right)^{1/2} \le \kappa,$$

then the event $\{32s^\star\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_N\|_\infty \le \kappa\}$ has a probability larger than $1-(\delta/2)$. To bound the

probability of $\mathscr{E}$, we use the fact that $\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)} = \bar{\boldsymbol{\zeta}}$ and the quantiles of the supremum norm of the random vector $\bar{\boldsymbol{\zeta}}$ have been assessed in 4.7.1. This implies that the choice

$$\lambda \geq 4B_Y \Big(\frac{\log(4p/\delta)}{n} \Big)^{1/2} + \frac{B_X B_Y \log(4p/\delta)}{n}$$

guarantees that $P(\mathscr{E}) = P(\|\boldsymbol{\zeta}\|_\infty \leq \lambda/2) \geq 1 - (\delta/2)$. This completes the proof.

Proof of 4.5.2.

We start by some auxiliary results before providing the proof of the theorem.

Proposition 4.7.3. *Let $J \subseteq [p]$ and let $\boldsymbol{\beta}$ be a minimizer of the right hand side of (4.5.7). On the event $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3$, where*

$$\mathscr{E}_1 = \big\{ \|\boldsymbol{\zeta}^{(1)}\|_\infty \leq \tfrac{\lambda}{4} \big\}, \quad \mathscr{E}_2 = \big\{ \|\boldsymbol{\zeta}^{(2)}\|_\infty \leq \tfrac{\lambda}{4} \big\}, \quad \textit{and} \quad \mathscr{E}_3 = \big\{ \lambda_{\min}(\boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Sigma}^{-1/2}) \geq \tfrac{2}{3} \big\}, \tag{4.7.47}$$

we have

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \leq 4\lambda \|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{2\kappa_{\widehat{\boldsymbol{\Sigma}}_{\text{all}}}(J,3)}.$$

Proof. Our starting point in this proof is (4.7.34). We first focus on bounding $\mathbf{T}_1$. On the event $\mathscr{E}_1 \cap \mathscr{E}_2$, we have

$$\mathbf{T}_1 \leq 2\|\boldsymbol{\zeta}^{(1)} + \boldsymbol{\zeta}^{(2)}\|_\infty \|\boldsymbol{u}\|_1 + 2\lambda(\|\boldsymbol{\beta}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq \lambda\big(\|\boldsymbol{u}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\widehat{\boldsymbol{\beta}}\|_1\big). \tag{4.7.48}$$

We now look for an upper bound of the term $\mathbf{T}_2$. On the event $\mathscr{E}_3$, for any $\boldsymbol{v} \in \mathbb{R}^p$,

$$\boldsymbol{v}^\top \big(2\mathbf{I}_p - 3\boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{v} \leq 0, \tag{4.7.49}$$

which leads to

$$\boldsymbol{v}^\top \big(\mathbf{I}_p - 2\boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{v} \leq -\tfrac{1}{2}(\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Sigma}^{-1/2})\boldsymbol{v}. \tag{4.7.50}$$

Therefore, applying (4.7.50) to $\boldsymbol{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}$, it follows that on the event $\mathscr{E}_3$

$$\mathbf{T}_2 = \boldsymbol{v}^\top \big(\mathbf{I}_p - 2\boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Sigma}^{-1/2}\big)\boldsymbol{v} \leq -\tfrac{1}{2}\boldsymbol{v}^\top (\boldsymbol{\Sigma}^{-1/2} \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{\Sigma}^{-1/2})\boldsymbol{v} = -\tfrac{1}{2}\|\mathbf{A}\boldsymbol{u}\|_2^2. \tag{4.7.51}$$

To sum up, equations (4.7.48) and (4.7.51) together imply that on the event $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3$,

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) - \mathcal{E}(f_{\boldsymbol{\beta}}) \leq \lambda\big(\|\boldsymbol{u}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\widehat{\boldsymbol{\beta}}\|_1\big) - \tfrac{1}{2}\|\mathbf{A}\boldsymbol{u}\|_2^2. \tag{4.7.52}$$

The desired result follows from this inequality and 4.7.2 with $\mu = \lambda$ and $\gamma = 2$. □

Note that according to 4.7.1,

$$\mathbb{P}\bigg(\|\boldsymbol{\zeta}^{(1)}\|_\infty \leq 2B_Y\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\bigg[1 + \frac{B_X}{3}\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\bigg]\bigg) \geq 1 - \frac{\delta}{3}. \tag{4.7.53}$$

The next two lemmas provide bounds for the probabilities of the events $\mathscr{E}_2$ and $\mathscr{E}_3$ introduced in 4.7.3.

Lemma 4.7.3. *Let assumption 4.1 be fulfilled. Let $J \subseteq [p]$ and let $\boldsymbol{\beta}$ be a minimizer of the right hand side of (4.5.7). Then, for all $\delta \in (0,1)$, the inequality*

$$\|\boldsymbol{\zeta}^{(2)}\|_\infty \geq B_X B_Y\Big(\frac{2\log(6p/\delta)}{N}\Big)^{1/2}\bigg[1 + \frac{B_X}{3}\Big(\frac{2p\|\boldsymbol{\Sigma}^{-1}\|\log(6p/\delta)}{N}\Big)^{1/2}\bigg] \tag{4.7.54}$$

holds with probability at most $\delta/3$, where the random vector $\boldsymbol{\zeta}^{(2)}$ is defined in 4.7.27.

Proof. Note that $\boldsymbol{\zeta}^{(2)} = (1/N)\sum_{i=1}^N \boldsymbol{s}U_i$, where $\boldsymbol{s}U_i = \boldsymbol{X}_i(\boldsymbol{X}_i^\top\boldsymbol{\beta}) - \mathbb{E}[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\beta})]$. The random vectors $\boldsymbol{s}U_i$ are independent and, for all $i \in [N]$ and all $j \in [p]$, the j-th component $U_{ij} = X_{ij}(\boldsymbol{X}_i^\top\boldsymbol{\beta}) - \mathbb{E}[X_j(\boldsymbol{X}^\top\boldsymbol{\beta})]$ of $\boldsymbol{s}U_i$ satisfies, almost surely,

$$|U_{ij}| \leq 2B_X^2\|\boldsymbol{\beta}\|_1 \leq 2B_X^2\sqrt{p}\|\boldsymbol{\beta}\|_2, \tag{4.7.55}$$

where we have used that $|\boldsymbol{X}^\top\boldsymbol{\beta}| \leq \|\boldsymbol{X}\|_\infty\|\boldsymbol{\beta}\|_1 \leq B_X\|\boldsymbol{\beta}\|_1$ with probability 1. Then, noticing that $\|\boldsymbol{\beta}\|_2 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 \leq \|\boldsymbol{\Sigma}^{-1/2}\|\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 = \|\boldsymbol{\Sigma}^{-1}\|^{1/2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2$, we deduce that

$$|U_{ij}| \leq 2B_X^2(p\|\boldsymbol{\Sigma}^{-1}\|)^{1/2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2, \tag{4.7.56}$$

almost surely. Since $\boldsymbol{\beta}$ minimizes the term on the right hand side of (4.5.7), by 4.7.5 below, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 \leq B_Y$. Thus for all $i \in [N]$ and all $j \in [p]$, $|U_{ij}| \leq 2B_X^2B_Y(p\|\boldsymbol{\Sigma}^{-1}\|)^{1/2}$. Furthermore, according to the previous lines, it holds $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[X_{ij}^2(\boldsymbol{X}_i^\top\boldsymbol{\beta})^2] \leq B_X^2B_Y^2$. 4.7.4 and the union bound complete the proof. □

Lemma 4.7.4. *Under assumption 4.1, the smallest eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2})$ of the*

matrix $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}$ *satisfies*

$$\mathbb{P}\Big\{\lambda_{\min}(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{\Sigma}^{-1/2}) \geq 1 - \Big(\frac{2B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(p/\delta)}{N}\Big)^{1/2}\Big\} \geq 1-\delta, \tag{4.7.57}$$

for all $\delta \in (0,1)$ *such that* $2B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(p/\delta) \leq N$.

Proof. For all $i \in [N]$, $\lambda_{\max}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1/2}) = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|^2 \leq pB_X^2\|\boldsymbol{\Sigma}^{-1}\|$ and the matrix $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1/2}$ is positive semi-definite. Applying the first Chernoff matrix inequality given in Remark 5.3 of Tropp (2012) to the sequence of matrices $\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1/2} : i \in [N]\}$ with

$$t = 1 - \Big(\frac{2B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(p/\delta)}{N}\Big)^{1/2}, \qquad R = pB_X^2, \qquad \delta = p\exp\Big\{-\frac{(1-t)^2 N}{2R\|\boldsymbol{\Sigma}^{-1}\|}\Big\} \tag{4.7.58}$$

yields (4.7.57). □

Lemma 4.7.5. *Let* $\mathrm{pen} : \mathbb{R}^p \to [0,+\infty)$ *be a convex function such that* $\mathrm{pen}(\mathbf{0}_p) = 0$. *Let* $\bar{\boldsymbol{\beta}}$ *be a minimizer of the function*

$$\Phi(\boldsymbol{\beta}) = \mathbb{E}[(\boldsymbol{\beta}^\top\boldsymbol{X} - Y)^2] + \mathrm{pen}(\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbb{R}^p. \tag{4.7.59}$$

Then $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \leq \mathbb{E}[Y^2]$ *and, if Assumption 4.1 is fulfilled,* $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \leq B_Y^2$.

Proof. We apply 4.7.1 with $\mathbf{A} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]^{1/2}$, $n = 1$, $\boldsymbol{Y} = 1$ and $\mathbf{X}_{\mathrm{lab}} = \mathbb{E}[Y\boldsymbol{X}]$ so that $\frac{1}{n}\boldsymbol{Y}^\top\mathbf{X}_{\mathrm{lab}} = \mathbb{E}[Y\boldsymbol{X}]$. Inequality (4.7.2) with $\boldsymbol{\beta} = \mathbf{0}_p$ yields

$$\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \leq 2\mathbb{E}[Y(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})] - \mathrm{pen}(\bar{\boldsymbol{\beta}}) - \mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]. \tag{4.7.60}$$

Rearranging the terms and using that $\mathrm{pen}(\bar{\boldsymbol{\beta}}) \geq 0$, we get $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2] \leq \mathbb{E}[Y(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})]$. In view of the Cauchy-Schwarz inequality, $(\mathbb{E}[Y(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})])^2 \leq \mathbb{E}[Y^2]\,\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]$, which implies that $(\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2])^2 \leq \mathbb{E}[Y^2]\,\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]$. It now suffices to divide both sides of the last inequality by $\mathbb{E}[(\bar{\boldsymbol{\beta}}^\top\boldsymbol{X})^2]$ to obtain the claim of the lemma. □

Proof of 4.5.2. Under the conditions of the theorem, we have

$$\Big(\frac{2B_X^2 p\|\boldsymbol{\Sigma}^{-1}\|\log(3p/\delta)}{N}\Big)^{1/2} \leq \frac{1}{3}.$$

Therefore, 4.7.4 implies that $\mathbb{P}(\mathscr{E}_3) \geq 1 - \delta/3$. On the other hand, in view of 4.7.53 and 4.7.3,

the conditions

$$\lambda \geq 8B_Y\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{\log(6p/\delta)}{n}\Big)^{1/2}\Big], \tag{4.7.61}$$

$$\lambda \geq 4B_XB_Y\Big(\frac{2\log(6p/\delta)}{N}\Big)^{1/2}\Big[1+\frac{B_X}{3}\Big(\frac{2p\|\boldsymbol{\Sigma}^{-1}\|\log(6p/\delta)}{N}\Big)^{1/2}\Big] \tag{4.7.62}$$

imply that $\mathbb{P}(\mathscr{E}_1) \geq 1-\delta/3$ and $\mathbb{P}(\mathscr{E}_2) \geq 1-\delta/3$. One can easily check that under the conditions of the theorem, the two inequalities of the last display are satisfied. Therefore, we have $\mathbb{P}(\mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3) \geq 1-\delta$. Finally, applying 4.7.3 we get the claim of the theorem. □

Proof of the oracle inequality in expectation.

Let δ be a positive number smaller than 1 to be chosen later. We have already seen in 4.5.1 that on an event $\mathscr{E}$ of probability $1-\delta$, we have

$$\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}}) \leq \inf_{J\subseteq[p]}\inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\mathcal{E}(f_{\boldsymbol{\beta}}) + 4\lambda\|\boldsymbol{\beta}_{J^c}\|_1 + \tfrac{27\|\boldsymbol{\Sigma}^{-1}\|}{4}\lambda^2|J|\Big\}. \tag{4.7.63}$$

On the other hand, using the fact that $\widehat{\boldsymbol{\beta}}$ minimises the function $\psi(\boldsymbol{\beta}) = \|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\boldsymbol{\beta}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_n\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1$, we have $\psi(\widehat{\boldsymbol{\beta}}) \leq \psi(\mathbf{0}_p)$, which yields

$$\|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\widehat{\boldsymbol{\beta}}\|_2^2 - \frac{2}{n}\boldsymbol{Y}^\top\mathbf{X}_n\widehat{\boldsymbol{\beta}} + 2\lambda\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\Sigma}}_N^{1/2}\widehat{\boldsymbol{\beta}} - \tfrac{1}{n}\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\boldsymbol{Y}\|_2^2 - \frac{1}{n^2}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\boldsymbol{Y}\|_2^2 + 2\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \leq 0.$$

Note that $\widehat{\boldsymbol{\Sigma}}_N^{-1/2}$ is understood as the Moore-Penrose pseudo-inverse and all the expressions involving this quantity are well defined since $N\widehat{\boldsymbol{\Sigma}}_N \succeq n\widehat{\boldsymbol{\Sigma}}_n = \mathbf{X}_n^\top\mathbf{X}_n$. This implies that $2\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n^2}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\boldsymbol{Y}\|_2^2 \leq \frac{1}{n^2}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\mathbf{X}_n^\top\|^2\|\boldsymbol{Y}\|_2^2 = \frac{1}{n}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\|\,\|\boldsymbol{Y}\|_2^2$, which entails

$$\|\widehat{\boldsymbol{\beta}}\|_1 \leq \frac{B_Y^2}{2\lambda}\|\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Sigma}}_N^{-1/2}\| \leq \frac{B_Y^2N}{2n\lambda}. \tag{4.7.64}$$

It is also true that for every $\boldsymbol{\beta}\in\mathbb{R}^p$,

$$\mathcal{E}(f_{\boldsymbol{\beta}}) = \mathbb{E}[(f^\star(\boldsymbol{X}) - \boldsymbol{X}^\top\boldsymbol{\beta})^2] \leq 2\mathbb{E}[f^\star(\boldsymbol{X})^2] + 2\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} \leq 2B_Y^2 + 2\|\boldsymbol{\beta}\|_1^2. \tag{4.7.65}$$

Therefore, we have $\mathbb{E}[\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}})\mathbb{1}_{\mathscr{E}^c}] \leq 2B_Y^2\mathbb{P}(\mathscr{E}^c) + 2\mathbb{E}[\|\widehat{\boldsymbol{\beta}}\|_1^2\mathbb{1}_{\mathscr{E}^c}] = 2\delta B_Y^2 + 2\mathbb{E}[\|\widehat{\boldsymbol{\beta}}\|_1^2\mathbb{1}_{\mathscr{E}^c}]$. Combining this inequality with (4.7.64), we get

$$\mathbb{E}[\mathcal{E}(f_{\widehat{\boldsymbol{\beta}}})\mathbb{1}_{\mathscr{E}^c}] \leq 2\delta B_Y^2 + \frac{\delta B_Y^4N^2}{2n^2\lambda^2}.$$

Setting $\delta = N^{-2}$, we get the claim of the theorem.

4.7.3 Bernstein inequality

The next result follows from (Massart, 2007, Proposition 2.9).

Proposition 4.7.4. *Let $Z_1, \dots, Z_N$ be independent real-valued random variables satisfying, for all $i \in [N]$ and for some constant b, $\mathbb{E}[Z_i^2] < +\infty$ and $|Z_i - \mathbb{E}Z_i| \leq b$ almost surely. Denote $\bar{Z}_N = \frac{1}{N}\sum_{i=1}^N Z_i$ and $\sigma_N^2 = (1/N)\sum_{i=1}^N \mathbb{E}\left[Z_i^2 - (\mathbb{E}Z_i)^2\right]$. Then, for all $\delta \in (0,1)$, inequality*

$$|\bar{Z}_N - \mathbb{E}[\bar{Z}_N]| \leq \sigma_N \Big(\frac{2\log(2/\delta)}{N}\Big)^{1/2} \Big[1 + \frac{b}{6N\sigma_N}\Big(\frac{2\log(2/\delta)}{N}\Big)^{1/2}\Big], \tag{4.7.66}$$

holds with probability at least $1 - \delta$.

Proof. Define, for all $i \in [N]$, the random variable $X_i = (Z_i - \mathbb{E}[Z_i])/N$. Denote as well

$$v = \sum_{i=1}^N \mathbb{E}[X_i^2] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[Z_i^2 - (\mathbb{E}Z_i)^2\right] = \frac{u}{N}.$$

For all $k \geq 3$, the assumptions imply that

$$\sum_{i=1}^N \mathbb{E}[(X_i)_+^k] \leq v\left(\frac{b}{N}\right)^{k-2} \leq \frac{k!}{2} v \left(\frac{b}{3N}\right)^{k-2},$$

where we have used the fact that $k!/3^{k-2} \geq 2$, for all $k \geq 3$. As a result, applying (Massart, 2007, Prop. 2.9), with $v = \sigma_N^2/N$ and $c = b/3N$, we get that for all $\delta \in (0,1)$, the inequality

$$\sum_{i=1}^N X_i > \sigma_N \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{b\log(2/\delta)}{3N}$$

holds with probability less than $\delta/2$. Applying the same argument to the variables $-X_i$, we infer that for all $\delta \in (0,1)$, the inequality

$$\sum_{i=1}^N X_i < -\sigma_N \sqrt{\frac{2\log(2/\delta)}{N}} - \frac{b\log(2/\delta)}{3N},$$

holds with probability less than $\delta/2$, which completes the proof. □

Bibliography

Bovas Abraham and Alice Chuang. Outlier detection and time series modeling. *Technometrics*, 31(2):241–248, 1989.

Eytan Adar and Lada A Adamic. Tracking information epidemics in blogspace. pages 207–214, 2005.

Shivani Agarwal, Deepak Dugar, and Shiladitya Sengupta. Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731, 2010.

Alan Agresti and Barbara Finlay. Statistical methods for the social sciences. printice hall. *Inc. NJ*, 1997.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Pierre Alquier. *Transductive and inductive adaptative inference for regression and density estimation*. PhD thesis, ENSAE ParisTech, 2006.

Pierre Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

Pierre Alquier and Gérard Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(Jan):243–280, 2013.

Pierre Alquier and Mohamed Hebiri. Transductive versions of the LASSO and the dantzig selector. *Journal of Statistical Planning and Inference*, 142(9):2485 – 2500, 2012.

Pierre Alquier and Karim Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *JMLR*, 17(239):1–41, 2016.

Christophe Andrieu, James Ridgway, and Nick Whiteley. Sampling normalizing constants in high dimensions using inhomogeneous diffusions. *arXiv preprint arXiv:1612.07583*, 2016.

Jaromır Antoch and Daniela Jarušková. Change point detection. In *FORUM STATISTICUM SLOVACUM*, page 2, 2000.

Yindalon Aphinyanaphongs, Lawrence D Fu, Zhiguo Li, Eric R Peskin, Efstratios Efstathiadis, Constantin F Aliferis, and Alexander Statnikov. A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *Journal of the Association for Information Science and Technology*, 65(10):1964–1987, 2014.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. pages 1568–1576, 2011.

Ery Arias-Castro and Karim Lounici. Estimation and variable selection with exponential weights. *Electronic Journal of Statistics*, 8(1):328–354, 2014.

Jean-Yves Audibert. Aggregated estimators and empirical complexity for least square regression. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 40, pages 685–736. Elsevier, 2004a.

Jean-Yves Audibert. A better variance control for pac-bayesian classification. *Preprint*, 905, 2004b.

Jean-Yves Audibert. Pac-bayesian statistical learning theory. 2004c.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

Jean-Yves Audibert and Olivier Catoni. Robust linear regression through pac-bayesian truncation. *Preprint*, 38:60, 2010.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Reg Austin. *Unmanned aircraft systems: UAVS design, development and deployment*, volume 54. John Wiley & Sons, 2011.

Alberto Bacci. Gabacortex, cortical inhibitory control circuits, anr. *Impact*, 2017(4):84–87, 2017.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Bubacarr Bah and Jared Tanner. Bounds of restricted isometry constants in extreme asymptotics: formulae for Gaussian matrices. *Linear Algebra Appl.*, 441:88–109, 2014.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Grundlehren der mathematischen Wissenschaften 348. Springer International Publishing, 1 edition, 2014. ISBN 978-3-319-00226-2,978-3-319-00227-9. URL http://gen.lib.rus.ec/book/index.php?md5=D7444D65A44D4F39407F88D54C90B446.

Maria-Florina Balcan, Avrim Blum, Patrick Pakyan Choi, John Lafferty, Brian Pantano, Mugizi R. Rwebangira, and Xiaojin Zhu. Person identification in webcam images: An application of semi-supervised learning. *ICML2005 Workshop on Learning with Partially Classified Training Data*, 2005.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7):1–23, 2014.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

Pierre C Bellec, Arnak S Dalalyan, Edwin Grappin, and Quentin Paris. On the prediction loss of the lasso in the partially labeled setting. *arXiv preprint arXiv:1606.06179*, 2016a.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016b.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 04 2014. doi: 10.1214/14-AOS1204. URL http://dx.doi.org/10.1214/14-AOS1204.

Aharon Ben-Tal and Arkadi Nemirovski. The conjugate barrier mirror descent method for non-smooth convex optimization. *Minerva optimization center, Technion Institute of Technology*, 1999.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

Sergey Bobkov and Mokshay Madiman. Concentration of the information in data with log-concave distributions. *The Annals of Probability*, 39(4):1528–1543, 2011.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.

Shekhar Borkar and Andrew A Chien. The future of microprocessors. *Communications of the ACM*, 54(5):67–77, 2011.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Mohamed Boukhebouze, Stéphane Mouton, and Jimmy Nsenga. Towards an on-board personal data mining framework for p4 medicine. *ERCIM NEWS*, (104):28–29, 2016.

Bruno Bouzy and Tristan Cazenave. Computer go: an ai oriented survey. *Artificial Intelligence*, 132(1):39–103, 2001.

Kevin K Bowden, Shereen Oraby, Amita Misra, Jiaqi Wu, and Stephanie Lukin. Data-driven dialogue systems for social agents. *arXiv preprint arXiv:1709.03190*, 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Max Bramer. Using j-pruning to reduce overfitting in classification trees. *Knowledge-Based Systems*, 15(5):301–308, 2002.

John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. pages 43–52, 1998.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16(3):199–231, 08 2001. doi: 10.1214/ss/1009213726. URL http://dx.doi.org/10.1214/ss/1009213726.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. *arXiv preprint arXiv:1705.08964*, 2017.

Céline Brouard, Florence d'Alché-Buc, and Marie Szafranski. Semi-supervised penalized output kernel regression for link prediction. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 593–600. Omnipress, 2011.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007a.

Florentina Bunea, Alexandre B Tsybakov, and Marten H Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007b.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, pages 1282–1309, 2011.

Robert Burbidge, Matthew Trotter, B Buxton, and Sl Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry*, 26(1): 5–14, 2001.

Jay Burmeister and Janet Wiles. The challenge of go as a domain for ai research: a comparison between go and chess. pages 181–186, 1995.

Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences*, 43(6):1882–1889, 2003.

T Tony Cai, Lie Wang, and Guangwu Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58(3):1300–1308, 2010.

Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

Murray Campbell, A Joseph Hoane, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Emmanuel J Candes. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15: 257–325, 2006.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Ismaël Castillo and Aad van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Olivier Catoni. Universal aggregation rules with exact bias bounds. *preprint*, 510, 1999.

Olivier Catoni. A pac-bayesian approach to adaptive classification. *preprint*, 840, 2003.

Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

Olivier Catoni. Pac-bayesian supervised classification. *Lecture Notes-Monograph Series. IMS*, 2007.

Gavin C Cawley and Nicola LC Talbot. Preventing over-fitting during model selection via bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8 (Apr):841–861, 2007.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3): 427–485, 1997.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

O. Chapelle, B. Shölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

Jim X Chen. The evolution of computing: Alphago. *Computing in Science & Engineering*, 18 (4):4–7, 2016.

Xi Chen, Xudong Liu, Zicheng Huang, and Hailong Sun. Regionknn: A scalable hybrid collaborative filtering algorithm for personalized web service recommendation. pages 9–16, 2010.

Anil M Cheriyadat. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):439–451, 2014.

Elena Chernousova, Yuri Golubev, and Ekaterina Krymova. Ordered smoothers with exponential weighting. *Electronic Journal of Statistics*, 7:2395–2419, 2013.

Christophe Chesneau and Guillaume Lecué. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. *Statistica Sinica*, pages 1407–1417, 2009.

Grant Clauser. What is alexa? what is the amazon echo, and should you get one? *The Wirecutter, last updated September*, 14, 2016.

Philippe Clement and Wolfgang Desch. An elementary proof of the triangle inequality for the wasserstein metric. *Proceedings of the American Mathematical Society*, 136(1):333–339, 2008.

Bruno Sielly Jales Costa, Plamen Parvanov Angelov, and Luiz Affonso Guedes. Fully unsupervised fault detection and identification based on recursive density estimation and self-evolving cloud-based classifier. *Neurocomputing*, 150:289–303, 2015.

Vincent Cottet and Pierre Alquier. 1-bit matrix completion: Pac-bayesian analysis of a variational approximation. *arXiv preprint arXiv:1604.04191*, 2016.

Kate Crawford. Artificial intelligence's white guy problem. *The New York Times*, 2016.

Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. pages 115–122, 2010.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

Dong Dai, Philippe Rigollet, Lucy Xia, and Tong Zhang. Aggregation of affine estimators. *Electronic Journal of Statistics*, 8(1):302–327, 2014.

Arnak Dalalyan and Alexandre Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. *Learning theory*, pages 97–111, 2007.

Arnak Dalalyan and Alexandre B Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1):39–61, 2008.

Arnak Dalalyan, Yuri Ingster, and Alexandre B Tsybakov. Statistical inference in compound functional models. *Probability Theory and Related Fields*, 158(3-4):513–532, 2014a.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Arnak S Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. *arXiv:1704.04752*, 2017.

Arnak S Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

Arnak S Dalalyan and Alexandre B Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012a.

Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012b.

Arnak S. Dalalyan, Mohamed Heibiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, in press, 2014b.

Arnak S Dalalyan, Edwin Grappin, and Quentin Paris. On the exponentially weighted aggregate with the laplace prior. 2016.

Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, and Blake Livingston. The youtube video recommendation system. pages 293–296, 2010.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

David Donoho. 50 years of data science. 2015.

David L Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1:32, 2000.

David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224, 1995.

Stephan Dreiseitl, Lucila Ohno-Machado, Harald Kittler, Staal Vinterbo, Holger Billhardt, and Michael Binder. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of biomedical informatics*, 34(1):28–36, 2001.

Dimiter Driankov and Alessandro Saffiotti. *Fuzzy logic techniques for autonomous vehicle navigation*, volume 61. Physica, 2013.

Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the unadjusted langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Seth Earley. Analytics, machine learning, and the internet of things. *IT Professional*, 17(1): 10–13, 2015.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Michael Elad and Alfred M Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

Stéphane Gaïffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic journal of statistics*, 1:538–573, 2007.

Stéphane Gaiffas and Guillaume Lecué. Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Transactions on Information Theory*, 57(10):6942–6957, 2011.

Chao Gao, Aad W van der Vaart, and Harrison H Zhou. A general framework for bayes structured linear models. *arXiv preprint arXiv:1506.02174*, 2015.

Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

John Gilliom. *Overseers of the poor: Surveillance, resistance, and the limits of privacy*. University of Chicago Press, 2001.

Christophe Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

Christophe Giraud. *Introduction to High-Dimensional Statistics*. CRC Press, 2015.

Yu Golubev and D Ostrovski. Concentration inequalities for the exponential weighting method. *Mathematical Methods of Statistics*, 23(1):20–37, 2014.

Fiona Graham. Wearable technology: Clothing designed to save your life. *BBC News*, 25, 2014.

Benjamin Guedj and Pierre Alquier. Pac-bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

Benjamin Guedj and Sylvain Robbiano. Une approche pac-bayésienne d'un probleme de ranking binaire en grande dimension. 2014.

Benjamin Guedj, Pierre Alquier, Gérard Biau, Éric Moulines, and Telecom ParisTech LTCI. Prévision pac-bayésienne pour le modele additif sous contrainte de parcimonie.

Matthieu Guillaumin, Jakob J. Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 902–909, 2010. URL http://dx.doi.org/10.1109/CVPR.2010.5540120.

Manish Gupta, Jing Gao, Charu C Aggarwal, and Jiawei Han. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267, 2014.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

Akhlaque Haque. *Surveillance, transparency, and democracy: Public administration in the information age*. University of Alabama Press, 2015.

Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3386–3393. IEEE, 2012.

S Hawking, E Musk, and S Wozniak. Autonomous weapons: an open letter from ai & robotics researchers. *Future of Life Institute*, 2015.

Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

Mohamed Hebiri and Johannes Lederer. How correlations influence lasso prediction. *IEEE Transactions on Information Theory*, 59(3):1846–1854, 2013.

Toby Hocking, Guillem Rigaill, Jean-Philippe Vert, and Francis Bach. Learning sparse penalties for change-point detection using max margin interval regression. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 172–180, 2013.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. 2003.

Feng-Hsiung Hsu. *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press, 2002.

Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Lucas Introna and David Wood. Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society*, 2(2/3):177–198, 2004.

Anil K Jain and Stan Z Li. *Handbook of face recognition*. Springer, 2011.

Nitin Jindal and Bing Liu. Review spam detection. pages 1189–1190, 2007.

Joyce John. Pandora and the music genome project. *Scientific Computing*, 23(10):40–41, 2006.

AB Juditsky, Alexander V Nazin, Alexandre B Tsybakov, and Nicolas Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384, 2005.

Anatoli Juditsky and Arkadi Nemirovski. Accuracy guarantees for-recovery. *Information Theory, IEEE Transactions on*, 57(12):7818–7839, 2011.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

Taghi M Khoshgoftaar and Edward B Allen. Controlling overfitting in classification-tree models of software quality. *Empirical Software Engineering*, 6(1):59–79, 2001.

Jyrki Kivinen and Manfred Warmuth. Averaging expert predictions. In *Computational Learning Theory*, pages 638–638. Springer, 1999.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Vladimir Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *The Annals of Statistics*, pages 1332–1359, 2009.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 38. Springer, 2011.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5): 2302–2329, 2011a.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5): 2302–2329, 2011b.

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016.

John D. Lafferty and Larry A. Wasserman. Statistical analysis of semi-supervised regression. In *NIPS*, pages 801–808. Curran Associates, Inc., 2007.

Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*, 19(2):646–675, 2013.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. Technical Report 1601.05584, arXiv, January 2016a.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. *arXiv preprint arXiv:1601.05584*, 2016b.

Guillaume Lecué and Philippe Rigollet. Optimal learning with q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

Gilbert Leung and Andrew R Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, and Vaughan Pratt. Towards fully autonomous driving: Systems and algorithms. pages 163–168, 2011.

Yew Jin Lim and Yee Whye Teh. Variational bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop*, volume 7, pages 15–21, 2007.

Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

Nicholas Littlestone. Mistake bounds and logarithmic linear-threshold learning algorithms. 1990.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

James NK Liu, Meng Wang, and Bo Feng. ibotguard: an internet-based intelligent robot security system using invariant face recognition against intruder. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(1):97–105, 2005.

Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2:90–102, 2008.

The Tien Mai and Pierre Alquier. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9(1):823–841, 2015.

Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

Colin L Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

John Markoff. Relax, the terminator is far away. *The New York Times*, 25, 2015.

Pascal Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*, volume 1896. Springer, 2007.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234. ACM, 1998.

John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Readings in artificial intelligence*, pages 431–450, 1969.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2015.

Raymond J Mooney and Loriene Roy. Content-based book recommending using learning for text categorization. pages 195–204, 2000.

Gordon E Moore. Progress in digital integrated electronics. 21:11–13, 1975.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems 22*, pages 1330–1338. Curran Associates, Inc., 2009.

Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Deanna Needell and Roman Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of computational mathematics*, 9(3): 317–334, 2009.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May): 1665–1697, 2012.

Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983.

Andrew Y Ng. Preventing" overfitting" of cross-validation data. 97:245–253, 1997.

Anton Nijholt. Google home: Experience, support and re-experience of social home activities. *Information Sciences*, 178(3):612–630, 2008.

Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14:1229–1250, 2013. URL http://jmlr.org/papers/v14/niyogi13a.html.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903*, 2013.

William Olmstadt. Cataloging expert systems: optimism and frustrated reality. *Journal of Southern Academic and Special Librarianship*, 1(3):n3, 2000.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Tohru Ozaki. A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach. *Statistica Sinica*, pages 113–135, 1992.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

Randy J Pell. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems*, 52(1):87–104, 2000.

David M Pennock, Eric Horvitz, Steve Lawrence, and C Lee Giles. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. pages 473–480, 2000.

M. Pensky. Solution of linear ill-posed problems using overcomplete dictionaries. Technical Report 1408.3386, Ann. Statist., to appear, arXiv, August 2014.

Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010a.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010b.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. pages 532–538, 2009.

John K Reid. A sparsity-exploiting variant of the bartels—golub decomposition for linear programming bases. *Mathematical Programming*, 24(1):55–69, 1982.

James Ridgway, Pierre Alquier, Nicolas Chopin, and Feng Liang. Pac-bayesian auc classification and scoring. In *Advances in Neural Information Processing Systems*, pages 658–666, 2014.

Philippe Rigollet. Generalized error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392, 2007.

Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011a.

Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011b.

Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012a.

Philippe Rigollet and Alexandre B Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012b.

Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

M Roux, S Asset, and S Medjebar. Tools for assisting diagnosis. *Revue de l'infirmiere*, 66(235): 26, 2017.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.

Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

Paat Rusmevichientong and David P Williamson. An adaptive algorithm for selecting profitable keywords for search-based advertising services. pages 260–269, 2006.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Friedhelm Schwenker and Edmondo Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37:4–14, 2014.

Amir Sepehri. The bayesian slope. *arXiv preprint arXiv:1608.08968*, 2016.

John Shawe-Taylor and Robert C Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9. ACM, 1997.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 853–860. IEEE, 2012.

Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, and Geraldine S Pinkus. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74, 2002.

Galit Shmueli. To explain or to predict? *Statist. Sci.*, 25(3):289–310, 08 2010. doi: 10.1214/10-STS330. URL http://dx.doi.org/10.1214/10-STS330.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529 (7587):484–489, 2016.

Anatoliy V Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4): 66, 2015.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *COLT*, volume 5, pages 545–560. Springer, 2005.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147, 1974.

O Stramer and RL Tweedie. Langevin-type models ii: self-targeting candidates for mcmc algorithms. *Methodology and Computing in Applied Probability*, 1(3):307–328, 1999.

Weijie Su and Emmanuel Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

Shiliang Sun and John Shawe-Taylor. Sparse semi-supervised learning using conjugate functions. *J. Mach. Learn. Res.*, 11:2423–2455, 2010.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012a.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012b.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012c.

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT Press Cambridge, 1998.

Hiroshi Tanaka. Stochastic differential equations with reflecting. *Stochastic Processes: Selected Papers of Hiroshi Tanaka*, 9:157, 1979.

Sebastian Thrun, Dieter Fox, Wolfram Burgard, and Frank Dellaert. Robust monte carlo localization for mobile robots. *Artificial intelligence*, 128(1-2):99–141, 2001.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996a.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996b.

Ryan J Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Alexandre B Tsybakov. Optimal rates of aggregation. In *COLT*, volume 2777, pages 303–313. Springer, 2003.

Alexandre B Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, pages 225–246, 2014.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Sara Van De Geer. The deterministic lasso. 2007.

Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

Sara van de Geer and Johannes Lederer. The lasso, correlated design, and improved oracle inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 303–316. Institute of Mathematical Statistics, 2013.

Sara A van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

SL van der Pas, J-B Salomond, and Johannes Schmidt-Hieber. Conditions for posterior contraction in the sparse normal means problem. *Electronic journal of statistics*, 10(1):976–1000, 2016.

Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

Vladimir N Vapnik and Alexey J Chervonenkis. Theory of pattern recognition. 1974.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *ArXiv e-prints*, November 2010.

Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\{\ell_1\}$-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

Junhui Wang and Xiaotong Shen. Large margin semi-supervised learning. *J. Mach. Learn. Res.*, 8:1867–1891, 2007.

Shijun Wang and Ronald M Summers. Machine learning and radiology. *Medical image analysis*, 16(5):933–951, 2012.

Manfred K Warmuth, Jun Liao, Gunnar Rätsch, Michael Mathieson, Santosh Putta, and Christian Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43(2):667–673, 2003.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Liyang Wei, Yongyi Yang, Robert M Nishikawa, and Yulei Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE transactions on medical imaging*, 24(3):371–380, 2005.

Wikipedia. History of ibm magnetic disk drives, 2017. URL https://en.wikipedia.org/wiki/History_of_IBM_magnetic_disk_drives. [Online; accessed 16 November 2017].

David Wipf, Jason Palmer, and Bhaskar Rao. Perspectives on sparse bayesian learning. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 249–256. MIT Press, 2003.

David P Wipf and Bhaskar D Rao. Sparse bayesian learning for basis selection. *IEEE Transactions on Signal processing*, 52(8):2153–2164, 2004.

Weng-Keen Wong, Andrew W Moore, Gregory F Cooper, and Michael M Wagner. Bayesian network anomaly pattern detection for disease outbreaks. pages 808–815, 2003.

Yuhong Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica*, pages 1069–1089, 2000a.

Yuhong Yang. Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74(1):135–161, 2000b.

Yuhong Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000c.

Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588, 2001a.

Yuhong Yang. Adaptive regression by mixing. 96:574–588, 02 2001b.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.*, 11:3519–3540, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Ling-Li Zeng, Hui Shen, Li Liu, and Dewen Hu. Unsupervised classification of major depression using functional connectivity mri. *Human brain mapping*, 35(4):1630–1641, 2014.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

Li Zhang, Weida Zhou, and Licheng Jiao. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):34–39, 2004.

Tong Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

Tong Zhang and Frank J Oles. Text categorization based on regularized linear classification methods. *Information retrieval*, 4(1):5–31, 2001.

Qun Zhao and Jose C Principe. Support vector machines for sar automatic target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 37(2):643–654, 2001.

Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. *Lecture Notes in Computer Science*, 5034:337–348, 2008.

X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin – Madison, 2008.

Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.

Arthur Zimek, Ricardo JGB Campello, and Jörg Sander. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter*, 15(1):11–22, 2014.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Estimateur par agrégat en apprentissage statistique en grande dimension

Mots clés : Agrégation, PAC-Bayésien, Estimation en Grande Dimension, Apprentissage Statistique.

Les travaux de cette thèse explorent les propriétés de procédures d'estimation par agrégation appliquées aux problèmes de régressions en grande dimension. Les estimateurs par agrégation à poids exponentiels bénéficient de résultats théoriques optimaux sous une approche PAC-Bayésienne. Cependant, le comportement théorique de l'agrégat avec *prior* de Laplace n'est guère connu. Ce dernier est l'analogue du Lasso dans le cadre pseudo-bayésien. Le Chapitre 2 explicite une borne du risque de prédiction de cet estimateur. Le Chapitre 3 prouve qu'une méthode de simulation s'appuyant sur un processus de Langevin Monte Carlo permet de choisir explicitement le nombre d'itérations nécessaire pour garantir une qualité d'approximation souhaitée. Le Chapitre 4 introduit des variantes du Lasso pour améliorer les performances de prédiction dans des contextes partiellement labélisés.

Model Averaging in Large Scale Learning

Key-words: Aggregation, PAC-Bayesian, High-Dimensional Estimation, Machine Learning.

This thesis explores properties of estimation procedures related to aggregation in the problem of high-dimensional regression in a sparse setting. The exponentially weighted aggregate (EWA) is well studied in the literature. It benefits from strong results in fixed and random designs with a PAC-Bayesian approach. However, little is known about the properties of the EWA with Laplace prior. Chapter 2 analyses the statistical behaviour of the prediction loss of the EWA with Laplace prior in the fixed design setting. Sharp oracle inequalities which generalize the properties of the Lasso to a larger family of estimators are established. These results also bridge the gap from the Lasso to the Bayesian Lasso. Chapter 3 introduces an adjusted Langevin Monte Carlo sampling method that approximates the EWA with Laplace prior in an explicit finite number of iterations for any targeted accuracy. Chapter 4 explores the statisctical behaviour of adjusted versions of the Lasso for the transductive and semi-supervised learning task in the random design setting.