

Contributions to the theoretical study of variational inference and robustness

Badr-Eddine Cherief-Abdellatif

▶ **To cite this version:**

Badr-Eddine Cherief-Abdellatif. Contributions to the theoretical study of variational inference and robustness. Statistics [math.ST]. Institut Polytechnique de Paris, 2020. English. NNT : 2020IPPAG001 . tel-02893465

HAL Id: tel-02893465
https://theses.hal.science/tel-02893465

Submitted on 8 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAG001

Thèse de doctorat

ENSAE

IP PARIS

Contributions to the theoretical study of variational inference and robustness

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

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

Thèse présentée et soutenue à Palaiseau, le 23/06/2020, par

BADR-EDDINE CHÉRIEF-ABDELLATIF

Composition du Jury :

Elisabeth Gassiat Université Paris-Saclay	Président
Arnaud Doucet University of Oxford	Rapporteur
Ismaël Castillo Sorbonne Université	Rapporteur
Arnak Dalalyan ENSAE	Examinateur
Aurélien Garivier ENS Lyon	Examinateur
Vincent Rivoirard Université Paris-Dauphine	Examinateur
Pierre Alquier RIKEN AIP	Directeur de thèse

Remerciements

En ouverture de ce manuscrit, je voudrais avant tout te remercier, Pierre. Durant mes années de thèse, tu auras été un vrai modèle et une source d'inspiration. J'ai été réellement impressionné par tes remarques et tes idées souvent lumineuses, par tes connaissances nombreuses et variées en mathématiques, et surtout par ta curiosité et ta soif d'apprendre toujours intarissables. Tu as su me proposer des sujets intéressants et tu as su m'y intéresser. Tu m'as apporté dès le début tout le soutien et l'aide dont j'avais besoin quand j'en avais besoin. Pendant toutes nos collaborations, tu as établi entre nous un rapport de chercheur à chercheur et non pas de professeur à élève (même si j'ai bien trop souvent été rappelé à cette réalité), et tu as su me laisser progressivement cette liberté intellectuelle dont je ferai usage pendant toute ma carrière. J'ai également été très ému lorsque tu m'as annoncé que tu comptais t'installer à Tokyo. A la fois attristé (comme tout le monde au labo) par ton départ et touché par la confiance que tu m'as accordée en prenant cette décision. Malgré les 10 000 km qui nous séparent depuis bientôt un an, j'estime que notre collaboration à distance a été un succès (comme peut en témoigner le fameux Skype 11h-18h du Vendredi), et j'espère, sans en douter vraiment, que nous continuerons à travailler ensemble.

Cette thèse ne serait rien sans le travail des rapporteurs et des membres du jury. Arnaud et Ismaël, j'espère que la lecture de ce manuscrit ne vous sera pas trop pénible, et je vous remercie chaleureusement d'avoir accepté de le rapporter. Arnaud, j'espère que notre collaboration future sera fructueuse et dépassera très largement le cadre académique. Je remercie également les autres membres constituant le jury: Arnak, Aurélien, Elisabeth et Vincent. Je ne saurais assez vous remercier.

J'ai été réellement ravi d'effectuer ma thèse au département de statistique du CREST. Un grand merci à Arnak pour avoir été un modèle de chercheur, d'enseignant et de footballeur pour nous tous, ainsi que pour toutes les discussions scientifiques et non scientifiques que nous avons pu avoir pendant ces quelques années. Ta plus grande trouvaille en tant que chercheur restera bien sûr le nom de notre équipe de foot. Un grand merci à Mohamed Vershynin pour sa joie de vivre et ses talonnades. Un grand merci également à Nicolas, Guillaume et Matthieu dont les discussions en pauses cafés et pendant les déjeuners auront égayé mes journées. Un grand merci à Sacha et à Cristina pour leur gentillesse et pour m'avoir fait profiter de nombreuses conférences et de séminaires divers durant ma thèse au CREST, à Marco pour ses visites éclair, à Vianney pour m'avoir permis de m'inspirer des sages paroles de *Maître* Gims pendant la rédaction de ce manuscrit, et à Victor pour son sacrifice *Adumien* pour que tous les doctorants puissent avoir leur diplôme. A nouveau, un grand merci à Pierre, pour avoir fait vivre le labo jusque dans

les pho du treizième, et pour m'avoir fait profiter de la vie nippone à plusieurs reprises. Also, I would like to thank you, Emti, for welcoming me in Tokyo. Working with you was extremely valuable. I hope we will have the opportunity to collaborate again in the future.

Bien entendu, je n'oublie pas mes chers co-doctorants. Geoffrey, comme tu l'as si bien dit, nous avons commencé cette aventure ensemble et la finissons ensemble. Ta bonne humeur, ton énergie et ton sens de l'humour n'ont d'égal que ta grinta et tes coups de gueule après Cormac sur le terrain. Je te souhaite le meilleur à Zurich, et j'espère que j'aurai l'occasion de revoir à de nombreuses reprises tes petits abdos et tes blagues sur les mamans en séminaires. Bisous à Lionel, Simo et Gautier, qui auront été des saumons exemplaires sur qui compter pendant mes deux premières années. A Lucie, qui aura courageusement préféré ses pamplemousses à mes pastèques pendant trois longues années. Heureux d'apprendre que tu pourras profiter de vrais fruits l'an prochain. Aux anciens: à Jeremy, ses punchlines de Booba et ses 200 kgs au deadlift, à Yannick, le plus stateux des économistes, à Léna, la minione de Pierre, et à Alexander, qui m'a laissé la lourde responsabilité de chef du fameux Bureau 3032 dès la fin de ma première année. A mes co-bureaux Nicolas, Philip, Gabriel, Aurélien et Jules. Nicolas, c'est toi le prochain chef de bureau, sois-en digne. A Avo, son flegme et ses friandises arméniennes. Et bien sûr à tous ceux qui ont contribué à la vie du labo et à ses fous rires, que ce soit en étant membre de la tristement célèbre conversation *CREST qu'on boit* ou bien simple résident des troisième et quatrième étages: Alexis, Amir, Arshak, Arya, Boris, Christophe, Dang, Fabien, Flore, François-Pierre, Gwen, Julien, Lucas, Martin, Meyer, Morgane, Rémi, Solenne, Suzanne, Thomas, Tom, Vincent, Younès.

On dit bien souvent que la thèse est une aventure collective, mais on ne le répète jamais assez. Je voudrais donc remercier Edith, Pascale et Arnaud pour leur gentillesse, leur disponibilité et leur bonne humeur. Wasfe, Patrick, Michèle, Marie-Christine, tous les membres de la liste des footeux du CSX, les baby-foot de la kfet et les étudiants qui l'habitent, l'ensemble du personnel de l'école et tous ceux qui ont rendu les journées sur un plateau perdu à Palaiseau agréables. Je pense qu'ils n'imaginent pas l'importance qu'ils ont eu au quotidien dans le bon déroulement de ma thèse.

Petite dédicace à mes amis qui m'ont soutenu de diverses manières pendant cette aventure. Merci à Adel pour son soutien vital, à Nagy et Madjer les DZ légendaires, à Ismaël et Yacouba pour les soirées Champions League et les discussions à propos de Chat, à Yoon, Pastou, Emeric, Louise, Pernin, Jordan, toujours là malgré l'éloignement et le temps qui passe, et à Julien pour notre longue amitié.

Enfin, je voudrais remercier mon plus grand fan, ma mère, MarseilleZidane, bakhatitis, ainsi que tous les membres de ma famille. Chacun d'entre vous sait ce que je lui dois sans que je n'aie besoin d'en dire plus.

Contents

I Introduction **1**

1 Context **3**

1.1 Bayesian inference: computation and theory 3

1.2 Online learning: convexity and Bayes' rule 13

1.3 Robustness to misspecification 23

2 Contributions **31**

2.1 Consistency of variational inference 31

2.2 Online variational inference 37

2.3 Robustness via Maximum Mean Discrepancy 41

3 Résumé substantiel **47**

3.1 Consistence de l'inférence variationnelle 47

3.2 Inférence variationnelle en ligne 53

3.3 Robustesse via Maximum Mean Discrepancy 57

II Consistency of variational inference for estimation and model selection **63**

4 Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures **65**

4.1 Introduction 65

4.2 Background and notations 67

4.3 Variational Bayes estimation of a mixture 70

4.4 Variational Bayes model selection 77

4.5 Conclusion 79

4.6 Proofs 80

5 Convergence Rates of Variational Inference in Sparse Deep Learning 99
5.1 Introduction . . . 99
5.2 Sparse deep variational inference . . . 104
5.3 Generalization of variational inference for neural networks . . . 107
5.4 Architecture design via ELBO maximization . . . 111
5.5 Discussion . . . 113
5.6 Proofs and additional results . . . 113

6 Consistency of ELBO maximization for model selection 131
6.1 Introduction . . . 131
6.2 Framework . . . 132
6.3 Consistency of the ELBO criterion . . . 134
6.4 Application to probabilistic PCA . . . 137
6.5 Proofs and additional results . . . 139

III Theoretical bounds for online variational inference algorithms 151

7 A Generalization Bound for Online Variational Inference 153
7.1 Introduction . . . 153
7.2 Generalization Properties of Bayesian Inference for Online Learning . . . 155
7.3 Online Variational Inference . . . 157
7.4 Generalization Bounds for Online VI . . . 160
7.5 Experiments . . . 163
7.6 Conclusion . . . 165
7.7 Proofs and additional results . . . 166

IV Robustness to misspecification via Maximum Mean Discrepancy 179

8 Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence 181
8.1 Introduction . . . 181
8.2 Background and definitions . . . 185
8.3 Nonasymptotic bounds in the dependent, misspecified case . . . 188
8.4 Examples . . . 196

8.5 Stochastic gradient algorithm for MMD estimation 202
8.6 Simulation study 206
8.7 Proofs 210
8.8 Conclusion 221

9 MMD-Bayes: Robust Bayesian Estimation via Maximum Mean Discrepancy 223
9.1 Introduction 223
9.2 Framework and definitions 224
9.3 Theoretical analysis of MMD-Bayes 225
9.4 Variational inference 226
9.5 Numerical experiments 227
9.6 Conclusion 227
9.7 Proofs and additional results 228

Bibliography 241

Part I

Introduction

Chapter 1

Context

This chapter introduces the different domains of statistics to which this PhD thesis contributes. The list of references is not exhaustive and may serve as an entry point to more detailed literature that will be partially explored throughout the manuscript.

This chapter begins with a general presentation of Bayesian inference and PAC-Bayes theory in Section 1.1. We address their computational aspects with a particular emphasis on variational inference, and present general theoretical results assessing frequentist guarantees to posterior and approximate distributions. In Section 1.2, we turn to algorithmic and theoretical aspects of online learning when data is not available at once but in a stream. In particular, we present several algorithms and provide their statistical analysis via regret bounds using either convex optimization tools or Bayesian theory. Finally, Section 1.3 is devoted to robustness, a field of statistics aiming at designing relevant estimators in situations where the data generating process is too complex to be approximated using usual statistical models.

1.1 Bayesian inference: computation and theory

The aim of statistical modeling is to understand a phenomenon given some observations. In frequentist statistics, the phenomenon is represented by a probability distribution $\mathbb{P}_0^{(n)}$ defined over a sample space $\mathcal{X}^{(n)}$ equipped with a σ-algebra $\mathcal{A}^{(n)}$, while the dataset $\mathbf{X}^{(n)}$ is assumed to be a random realization of the unknown phenomenon $\mathbb{P}_0^{(n)}$. Unless explicitly stated otherwise, we shall consider in the following the particular *i.i.d.* framework where the dataset $\mathbf{X}^{(n)} = (X_i)_{i=1}^n$ is composed of n independent and identically distributed random variables of a phenomenon P_0 which is defined over a measurable space $(\mathcal{X}, \mathcal{A})$.

The starting point of any statistical analysis is *the model* - a collection of probability distributions $\{P_{\boldsymbol{\theta}}/\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed by a parameter $\boldsymbol{\theta}$, where $\boldsymbol{\Theta}$ is called the parameter set. In this thesis, most examples will be taken from parametric statistics where $\boldsymbol{\Theta} \subset \mathbf{R}^d$. Assuming that the data generating process belongs to the model, i.e. that there exists a parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $P_0 = P_{\boldsymbol{\theta}_0}$, many statistical problems actually boil down to the estimation of the true parameter $\boldsymbol{\theta}_0$ using a measurable function of the data.

1.1.1 The Bayesian paradigm

The Bayesian methodology relies on a different principle. Rather than assuming the existence of a true parameter $\boldsymbol{\theta}_0$, the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is viewed as a random variable (when equipping $\boldsymbol{\Theta}$ with a suitable σ-algebra $\mathcal{T}$). Some *prior* distribution Π_0 representing a prior belief as to which parameters are likely to have generated the data is placed over $\boldsymbol{\Theta}$. This prior belief is then updated and refined using the *Bayes' rule* by conditioning on the observed data $\mathbf{X}^{(n)} = (X_i)_{i=1}^n$, giving rise to the *posterior* distribution $\Pi_n(\cdot)$ which provides a natural way of quantifying uncertainty. For simplicity, we omit the dependence in the data in the notation Π_n and just replace it by the subscript n.

Bayes' rule, which was first published in the reverend's manuscript (Bayes, 1763) posthumously, was expressed in its modern version by Laplace (1774) a few years later. Assuming that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the probability distribution $P_{\boldsymbol{\theta}}$ is dominated by some reference measure μ and that the density $p_{\boldsymbol{\theta}} = \frac{dP_{\boldsymbol{\theta}}}{d\mu}$ is such that the map $(x, \boldsymbol{\theta}) \to p_{\boldsymbol{\theta}}(x)$ is $\mathcal{X} \times \mathcal{T}$-measurable, then the log-likelihood ℓ_n is simply defined as $\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i)$, and Bayes' formula characterizing the posterior distribution $\Pi_n(\cdot)$ is defined as follows:

$$\Pi_n(d\boldsymbol{\theta}) := \frac{\exp(\ell_n(\boldsymbol{\theta}))\Pi_0(d\boldsymbol{\theta})}{\int \exp(\ell_n(\phi))\Pi_0(d\phi)} = \frac{\prod_{i=1}^n p_{\boldsymbol{\theta}}(X_i)\Pi_0(d\boldsymbol{\theta})}{\int \prod_{i=1}^n p_{\phi}(X_i)\Pi_0(d\phi)}$$

where the denominator is a normalizing constant. We refer the reader to Robert (2007) for an exhaustive monograph on Bayesian statistics.

Although Bayesian inference offers a practical probabilistic approach, a recurrent criticism is the strong dependence on the statistical modeling - in particular on the statistical model and the prior - which characterizes some subjective vision of the underlying phenomenon. In machine learning, the focus is more on the predictive performance with respect to a risk measure $R(\boldsymbol{\theta})$ than on estimation and interpretability (Breiman et al., 2001), and thus classical Bayesian modeling appears to be too much of a constraint for designing efficient learning procedures. Hence, several extensions of Bayes' rule have been considered over the years, especially in the machine learning community, and most of these efforts are now bundled under the name *generalized Bayes* where the idea is to replace the negative log-likelihood $-\ell_n(\boldsymbol{\theta})$ by an empirical risk measure $r_n(\boldsymbol{\theta})$ depending on the dataset $\mathbf{X}^{(n)}$:

$$\Pi_n(d\boldsymbol{\theta}) := \frac{\exp(-r_n(\boldsymbol{\theta}))\Pi_0(d\boldsymbol{\theta})}{\int \exp(-r_n(\phi))\Pi_0(d\phi)}. \tag{1.1}$$

As soon as the normalizing constant is finite, such a machinery provides a coherent and principled way to update beliefs on $\boldsymbol{\theta}$ (Bissiri et al., 2016), although the generalized posteriors (also referred to as *pseudo-posteriors* or *quasi-posteriors*) we obtain no longer respect the fundamental rules of probability that underlie Bayes' formula. Hence, generalized Bayes should not be regarded as the amount of knowledge obtained once the dataset is available (according to a possible interpretation of Bayesian statistics) but simply as a Bayesian-flavored estimator of the unknown. Of course, when the loss function is chosen to be the negative log-likelihood $r_n(\boldsymbol{\theta}) = -\ell_n(\boldsymbol{\theta})$, then it boils down to classical Bayesian inference. In the next paragraphs, we will focus on two other popular choices of the risk leading to important generalized posteriors: the *tempered* and the *Gibbs* posteriors.

The tempered posterior

When considering a statistical model $\{P_{\boldsymbol{\theta}}/\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, Zhang (2006) proposed to raise the likelihood to an α-power $(0 < \alpha < 1)$ in Bayes' formula:

$$\Pi_{n,\alpha}(d\boldsymbol{\theta}) := \frac{\exp(\alpha \ell_n(\boldsymbol{\theta}))\Pi_0(d\boldsymbol{\theta})}{\int \exp(\alpha \ell_n(\phi))\Pi_0(d\phi)} = \frac{\prod_{i=1}^n p_{\boldsymbol{\theta}}(X_i)^\alpha \Pi_0(d\boldsymbol{\theta})}{\int \prod_{i=1}^n p_\phi(X_i)^\alpha \Pi_0(d\phi)}, \tag{1.2}$$

giving rise to the so-called *tempered posterior*, following an idea dating back to Vovk (1990). Based on the fact that regular Bayesian inference may fail when the model is wrong, i.e. when it does not contain the data generating distribution, Grünwald et al. (2017) further investigated this idea and developed the *Safe Bayes* paradigm by proposing an automated way of selecting the temperature parameter α that yields a robust pseudo-posterior distribution. We will see in the following that the use of a temperature parameter is of great interest for theoretical (Bhattacharya et al., 2016) and robustness (Barron et al., 1999; Grünwald et al., 2017) purposes. We do not give more details right now, and we refer the reader to Section 1.3 for an introduction to the notion of *robustness*.

The PAC-Bayes framework

In a statistical learning framework, no particular model for the data generating process is assumed. In supervised learning, we observe a collection of i.i.d. random variables $\mathcal{D}_n = (Z_i, Y_i)_{i=1,\dots,n}$ from a distribution P_0 defined over the sample space $\mathcal{Z} \times \mathcal{Y}$ (equipped with some σ-algebra). Depending on the nature of the problem, e.g. classification or regression, a set of predictors $\{f_{\boldsymbol{\theta}} : \mathcal{Z} \mapsto \mathcal{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed over a parameter space $\boldsymbol{\Theta}$ is chosen, along with a loss function measuring the discrepancy between a prediction $f_{\boldsymbol{\theta}}(z)$ of an input z and its associated output y. The measure of performance is then $R(\boldsymbol{\theta}) = \mathbb{E}[\ell(Y, f_{\boldsymbol{\theta}}(Z))]$ and its empirical counterpart is $r_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f_{\boldsymbol{\theta}}(Z_i))$. As the theoretical risk function $R(\boldsymbol{\theta})$ is not available (as P_0 is unknown), classical approaches proposed in the 90s have mainly consisted in variants of the Empirical Risk Minimization (ERM) principle (Vapnik, 1992).

The PAC-Bayes approach was developed by Shawe-Taylor and Williamson (1997a), McAllester (1999) and Catoni (2007) in order to obtain PAC (Probably Approximately Correct) bounds to Bayesian-based algorithms in the current model-free setting. By tempering Formula (1.1) applied to the previous empirical risk $r_n(\boldsymbol{\theta})$, we obtain the *Gibbs posterior*:

$$\Pi_{n,\alpha}(d\boldsymbol{\theta}) := \frac{\exp(-\alpha r_n(\boldsymbol{\theta}))\Pi_0(d\boldsymbol{\theta})}{\int \exp(-\alpha r_n(\phi))\Pi_0(d\phi)} \tag{1.3}$$

which is a cornerstone of PAC-Bayes theory. Indeed, one of the key results of PAC-Bayes due to Donsker and Varadhan states that the Gibbs posterior minimizes the upper bound of some oracle inequality applied to the risk of random estimators (see Section 1.1.2 for a precise statement). More especially, the Gibbs posterior is the solution of the following variational problem:

$$\min_Q \left\{ -\alpha \int r_n(\boldsymbol{\theta}) Q(d\boldsymbol{\theta}) + \mathrm{KL}(Q\|\Pi_0) \right\} = -\log\left(\int e^{-\alpha r_n(\boldsymbol{\theta})} \Pi_0(d\boldsymbol{\theta}) \right)$$

where the infimum, taken over the whole space $\mathcal{M}_1^+(\mathbf{\Theta})$ of probability distributions over $\mathbf{\Theta}$, is reached for $Q = \Pi_{n,\alpha}$, and where KL stands for the Kullback-Leibler divergence $\mathrm{KL}(P\|R) = \int \log\left(\frac{dP}{dR}\right) dP$ if R dominates P and $+\infty$ otherwise. Using this fundamental relationship, PAC-Bayes theory provides a bunch of powerful tools to derive bounds and offers sharp theoretical guarantees to such estimators. Some recent references include Ambroladze et al. (2007); Alquier (2008b); Parrado-Hernández et al. (2012); Guedj (2019). PAC-Bayes bounds are to be connected with the literature on aggregation of estimators (Leung and Barron, 2006; Dalalyan and Tsybakov, 2007; Salmon and Dalalyan, 2011).

1.1.2 Variational inference

Unfortunately, exact Bayesian inference is often computationally challenging in practice. The most popular technique to overcome the intractability of posterior distributions is Monte Carlo sampling, including MCMC algorithms (Andrieu et al., 2003; Robert and Casella, 2013) and Sequential Monte Carlo (Doucet et al., 2001; Doucet and Johansen, 2009). Nevertheless, such sampling methods can be slow for practical uses when the dataset is very large, and fast approximation methods such as Expectation Propagation (Minka, 2001) are sometimes used in PAC-Bayes (Ridgway et al., 2014). A more and more popular and fast alternative referred to as variational inference (VI) consists in finding a deterministic approximation of the posterior called variational Bayes (VB) approximation (Jordan et al., 1999; Blei et al., 2017). This approximate inference method significantly reduces the computation cost and enables application of the Bayesian approach to many large-scale machine learning problems (Hoffman et al., 2013; Kingma and Welling, 2013).

Two equivalent definitions of variational inference have appeared in the literature: the KL minimization and the ELBO maximization versions. We first present VI principle via KL minimization along with practical examples, and then we show the relationship with ELBO maximization via Donsker and Varadhan's lemma.

KL minimization

We choose a family $\mathcal{Q}$ of tractable distributions over $\mathbf{\Theta}$, and we define the variational approximation of the posterior $\widetilde{\Pi}_n$ as the closest distribution to Π_n (with respect to the KL divergence) belonging to $\mathcal{Q}$:

$$\widetilde{\Pi}_n = \arg\min_{Q\in\mathcal{Q}} \mathrm{KL}(Q\|\Pi_n). \tag{1.4}$$

Note that the choice of the variational set $\mathcal{Q}$ is of major interest, leading to a fundamental trade-off between accuracy and tractability. Indeed, when $\mathcal{Q}$ is the whole space of probability distributions over $\mathbf{\Theta}$, then the variational approximation is the posterior itself, but is no more tractable. At the opposite, if $\mathcal{Q}$ is not large enough, it may not contain any distribution that is close to our target Π_n. Several variational families have been proposed in the literature and lead to good approximations of the posterior. Here are a few examples.

Example 1.1.1 (The mean-field approximation)**.** *The mean-field approximation is very popular in the variational inference community (Blei et al., 2017). This is a natural*

choice when the space of parameters is based on a decomposition $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_K$, *e.g. mixture models.* $\mathcal{Q}$ *is then defined as the set of product probability distributions*

$$\mathcal{Q} = \left\{ Q(\boldsymbol{\theta}) = \prod_{j=1}^{K} Q_j(\boldsymbol{\theta}_j) / Q_j \in \mathcal{M}_1^+(\boldsymbol{\Theta}_j) \; \forall j = 1, ..., K \right\}.$$

This choice is particularly convenient as it leads to a natural fixed-point algorithm. Indeed, the coordinatewise minimization of the optimization program (1.4) *gives the following fixed-point conditions (Bishop, 2006):*

$$Q_j(d\boldsymbol{\theta}_j) \propto \exp\left(\int \left\{ \alpha \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(X_i) + \log \Pi(\boldsymbol{\theta}) \right\} \prod_{k \neq j} Q_k(d\boldsymbol{\theta}_k) \right) \Pi_0(d\boldsymbol{\theta}_j) \; \forall j = 1, ..., K$$

where $\propto$ *means "up to a normalizing constant", which can be solved in practice by updating successively every* Q_j.

Example 1.1.2 (The parametric approximation)**.** *It is also possible to consider a parametric approximation where the variational family* $\mathcal{Q} = \{Q_\lambda / \lambda \in \Lambda\}$ *is indexed by a finite-dimensional parameter set* Λ. *Hence, Definition* (1.4) *becomes a simple optimization program with respect to* λ *and usual tools from optimization theory can be used. One popular example is the d-dimensional Gaussian family with the usual mean/covariance parameterization* $\lambda = (m, \Sigma)$*:*

$$\mathcal{Q} = \{\mathcal{N}(m, \Sigma) / (m, \Sigma) \in \mathbb{R}^d \times \mathcal{S}_{++}^d\},$$

where $\mathcal{N}(m, \Sigma)$ *is the Gaussian family of mean* m *and covariance matrix* Σ *and* $\mathcal{S}_{++}^d$ *is the set of symmetric positive definite matrices. It is also possible to combine parametric VI with mean-field VI, by imposing for instance a diagonal covariance matrix* Σ *to a Gaussian family.*

In such a parametric setting, learning the variational approximation, i.e. learning the associated variational parameter, can be easily conducted through classical black-box optimization techniques such as gradient descent (Blei et al., 2017; Khan and Lin, 2017; Khan and Nielson, 2018).

ELBO maximization

Another point of interest is the choice of the KL as the measure of closeness between probability distributions. The use of this statistical divergence is particularly relevant here. Indeed, even though it is not possible to minimize the KL divergence in (1.4) exactly as its expression involves an intractable normalizing constant, it is possible to minimize a function that is equal to it up to a constant. Such a function is called the *Evidence Lower Bound (ELBO)*, and is often taken as the definition of variational inference.

To begin with, let us introduce the exact formulation of Donsker and Varadhan's variational formula which is at the core of the PAC-Bayes theory. This lemma will help us rewrite Definition (1.4) via the ELBO maximization program. We refer to Catoni (2007) for a proof (Lemma 1.1.3).

Lemma 1.1.1. *For any probability* Π *on some measurable space* $(\boldsymbol{E}, \mathcal{E})$ *and any measurable function* $h : \boldsymbol{E} \to \mathbb{R}$ *such that* $\int \mathrm{e}^h \mathrm{d}\Pi < \infty$,

$$\log \int \mathrm{e}^h \mathrm{d}\Pi = \sup_{Q \in \mathcal{M}_1^+(\boldsymbol{E})} \left\{ \int h \mathrm{d}Q - \mathrm{KL}(Q \| \Pi) \right\},$$

with the convention $\infty - \infty = -\infty$. *Moreover, if* h *is upper-bounded on the support of* Π, *then the supremum on the right-hand side is reached by the distribution of the form:*

$$\Pi_h(d\beta) = \frac{\mathrm{e}^{h(\beta)}}{\int \mathrm{e}^h \mathrm{d}\Pi} \Pi(\mathrm{d}\beta).$$

As for the Gibbs posterior, this lemma leads to:

$$\Pi_n = \underset{Q \in \mathcal{M}_+^1(\boldsymbol{\Theta})}{\arg\min} \left\{ - \int \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i) Q(d\boldsymbol{\theta}) + \mathrm{KL}(Q \| \Pi_0) \right\} \tag{1.5}$$

which is to be compared to the famous reformulation of (1.4):

$$\begin{aligned} \widetilde{\Pi}_n &= \underset{Q \in \mathcal{Q}}{\arg\min} \left\{ - \int \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i) Q(d\boldsymbol{\theta}) + \mathrm{KL}(Q \| \Pi_0) \right\} \\ &= \underset{Q \in \mathcal{Q}}{\arg\max} \left\{ \int \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i) Q(d\boldsymbol{\theta}) - \mathrm{KL}(Q \| \Pi_0) \right\} \end{aligned} \tag{1.6}$$

when applied to the standard posterior. The quantity maximized in (1.6) is called the ELBO, and is often taken as the criterion to maximize, particularly for people coming from the optimization community. The ELBO also provides a nice interpretation of variational inference based on the Minimum Description Length (MDL) principle (Rissanen, 1978; Grünwald, 2000), a formalization of Occam's razor that is widely used in statistics for model selection (Hansen and Yu, 2001; Chambaz et al., 2009). We refer to the short report Jerfel (2017) for more details on the connection between MDL and the ELBO.

Remark 1.1.1. *Note that it is possible to extend* (1.4) *to any generalized posterior. This gives in particular*

$$\widetilde{\Pi}_{n,\alpha} = \underset{Q \in \mathcal{Q}}{\arg\min} \left\{ -\alpha \int \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i) Q(d\boldsymbol{\theta}) + \mathrm{KL}(Q \| \Pi_0) \right\} \tag{1.7}$$

when dealing with the tempered posterior, which will be a distribution of interest in the rest of this manuscript.

1.1.3 Theoretical guarantees

First and foremost, the theoretical analysis of Bayesian procedures depends on the criterion used to quantify the quality of the posterior distribution. The Bayesian approach does not presume the existence of a true underlying distribution, and saying whether the posterior behaves well or not is not straightforward. Hence, it raises the question of finding a universal way to assess the quality of the posterior distribution.

The most common theory adopted in the literature is the *frequentist* analysis of *Bayesian* estimators. Indeed, from a frequentist point a view, when there exists a data generating $P_0 = P_{\boldsymbol{\theta}_0}$, the data-dependent posterior related to some prior distribution Π_0 is a simple estimator and hence can be analyzed as any other one. Probably the most popular criterion encountered in the literature is the asymptotic convergence of the Bayesian estimator, usually referred to as *concentration* of the posterior, and often associated with a rate of convergence. Roughly speaking, the rate with respect to a distance d is usually thought of as the smallest value r_n such that the posterior probability of neighborhoods of radius r_n containing the true distribution tends towards 1 in P_0-probability.

In the rest of this section, we will detail the standard conditions that are required in order to obtain convergence of posteriors and of their variational approximations, and we will finally provide some recent results guaranteeing convergence given such conditions.

Recent advances

The frequentist analysis of Bayesian procedures has a long and rich history. We will focus in the next few paragraphs on recent advances, and we refer the reader to the habilitation thesis of Castillo (2014) for a more complete historical overview.

The first convergence results date back to Doob (1949); Breiman et al. (1964) but are from a prior's perspective and do not match the modern conception of posterior concentration that must hold for any value of the true distribution. A breakthrough was achieved in Schwartz (1965) where the author gave sufficient conditions for concentration in the i.i.d. case under a prior mass condition and the existence of exponential tests. This work, along with those of Ghosal et al. (1999); Barron et al. (1999) which covered different priors, paved the way to the famous *prior mass and testing* setting of the seminal papers of Ghosal et al. (2000); Shen (2002) and Ghosal et al. (2007) for obtaining convergence with rates. In particular, the prior mass condition states that the prior Π_0 must give enough mass to some neighborhood (in the Kullback-Leibler sense) of the true parameter.

A few years ago, Bhattacharya et al. (2016) revealed another benefit arising from considering tempered posteriors rather than regular ones. The main finding of this paper is that the testing conditions are no longer necessary when looking at the tempered version of the posterior $\Pi_{n,\alpha}$, and that the prior mass condition of Ghosal et al. (2000) alone is sufficient to obtain the concentration of $\Pi_{n,\alpha}$ with explicit rates of convergence. Hence, one can get concentration of the posterior to the true posterior under less stringent conditions. Note that Walker and Hjort (2001) established fifteens years before simple prior mass conditions under which the tempered posterior is consistent for $\alpha = 1/2$.

More recently, Alquier and Ridgway (2017) and Bhattacharya et al. (2018) extended the prior mass assumption in order to get the concentration of variational approximations $\widetilde{\Pi}_{n,\alpha}$ of the tempered posteriors. In addition to the previous prior mass condition, this extension requires the variational set $\mathcal{Q}$ to contain probability measures concentrated around the true parameter. Furthermore, the authors provided nonasymptotic oracle-type inequalities on $\widetilde{\Pi}_{n,\alpha}$ via PAC-Bayes theory that imply the concentration of the posterior to the true parameter $\boldsymbol{\theta}_0$ when such a $\boldsymbol{\theta}_0$ exists and make it possible to quantify the convergence of $\widetilde{\Pi}_{n,\alpha}$ in case of misspecification. In particular when $\mathcal{Q} = \mathcal{M}_1^+(\boldsymbol{\Theta})$,

i.e. when there is no approximation, the extended prior mass condition of Alquier and Ridgway (2017) simply boils down to the standard prior mass condition of Ghosal et al. (2000). With additional testing conditions to the extended prior mass assumption, Zhang and Gao (2017) showed that variational approximations of regular posteriors satisfy the same properties.

At the same time, a study of the limit posterior distribution in parametric estimation was first conducted by Laplace (1810), and then further investigated by Bernstein (1917) and Von Mises (1931). The classical Bernstein - von Mises theorem mainly says that the posterior can be approached, as n increases, by a Gaussian distribution centered at an efficient pointwise estimator of $\boldsymbol{\theta}_0$ and with variance the inverse of the Fisher information matrix of the whole sample (Van der Vaart, 2000). An extension to variational approximations has been studied in Wang and Blei (2018). We will not investigate further this notion in the following of the thesis. We also refer the reader to Banerjee et al. (2020) and Ghosal and Van der Vaart (2017) for reviews on the properties of Bayesian methods in high-dimensional and in nonparametric models respectively.

Posterior concentration

Let us now introduce the formal definition of posterior concentration. The definition is the same for the tempered versions and variational approximations.

Let r_n be a sequence decreasing to 0 as n goes to infinity. We assume that there is a true parameter $\boldsymbol{\theta}_0$ associated to the data generating distribution $P_0 = P_{\boldsymbol{\theta}_0}$. Moreover, we equip the space of distributions $\mathcal{M}_+^1(\boldsymbol{\Theta})$ with a statistical distance d.

Definition 1.1.1. *The posterior distribution Π_n is said to concentrate if for any $r > 0$*

$$\Pi_n\Big(\boldsymbol{\theta} \in \boldsymbol{\Theta} \,\Big/\, d(P_{\boldsymbol{\theta}}, P_0) > r\Big) \xrightarrow[n\to+\infty]{} 0$$

in P_0-probability as $n \to +\infty$.

For a better understanding of the asymptotic behavior of posteriors, let us now define their rates of convergence.

Definition 1.1.2. *The posterior distribution Π_n is said to concentrate at rate r_n if*

$$\Pi_n\Big(\boldsymbol{\theta} \in \boldsymbol{\Theta} \,\Big/\, d(P_{\boldsymbol{\theta}}, P_0) > M_n r_n\Big) \xrightarrow[n\to+\infty]{} 0$$

in P_0-probability as $n \to +\infty$ for any $M_n \to +\infty$.

What is generally referred to as *the* rate of convergence is the smallest value of r_n satisfying Definition (1.1.2). Note that when the posterior distribution concentrates at rate r_n, there exists a pointwise estimator $\widehat{\boldsymbol{\theta}}_n$ that converges at the same rate in P_0-probability (Ghosal et al., 2000), and hence the best rate of convergence the Bayesian posterior can achieve is the frequentist minimax one, even though additional assumptions on the prior may be required to establish a clear connection with minimaxity.

Also, it is interesting to point out the importance of the statistical distance d. Indeed, any mathematical convergence result is always obtained with respect to a metric. Of course, the choice of such a metric depends on the statistical problem, but it is not necessarily required to be a mathematical distance in the sense that it may not be symmetric nor satisfy the triangle inequality. Usual appropriate choices of such a distance include the Kullback-Leibler divergence and the Hellinger distance H defined by $H(P,R)^2 = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dR})^2 = 1 - e^{-\frac{1}{2}D_{1/2}(P,R)}$. Another statistical distance of interest is the α-Rényi divergence D_α. It is defined as follows:

$$D_\alpha(P,R) = \frac{1}{\alpha - 1}\log\int\left(\frac{dP}{d\mu}\right)^\alpha\left(\frac{dR}{d\mu}\right)^{1-\alpha}d\mu,$$

where μ is any measure which dominates both P and R, e.g. $P + R$. Of course, the definition does not depend on the reference measure μ. For fixed values of P and R, $D_\alpha(P,R)$ is an increasing function of α and is related to the KL divergence and Hellinger distance through the following inequalities (Van Erven and Harremos, 2014):

$$2H(P,R)^2 \leq D_{1/2}(P,R) \leq D_\alpha(P,R) \leq \mathrm{KL}(P\|R) \leq D_\beta(P,R) \quad \text{for} \quad 1/2 < \alpha < 1 < \beta.$$

Prior mass condition

Let us now introduce the prior mass condition and its extended version to variational inference as formulated in Alquier and Ridgway (2017):

Definition 1.1.3. *Let us define the KL-ball $\mathcal{B}(P_0, r_n)$ centered at θ_0 of radius r_n:*

$$\mathcal{B}(P_0, r_n) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} / \ \mathrm{KL}(P_0\|P_{\boldsymbol{\theta}}) \leq r_n\}.$$

Then the prior mass condition is satisfied if

$$\Pi_0(\mathcal{B}(P_0, r_n)) \geq e^{-nr_n}, \tag{1.8}$$

while the extended prior mass is satisfied if there exists a distribution $Q_n \in \mathcal{Q}$ such that:

$$\int \mathrm{KL}(P_0\|P_{\boldsymbol{\theta}})Q_n(d\boldsymbol{\theta}) \leq r_n \ \text{ and } \ \mathrm{KL}(Q_n\|\Pi_0) \leq nr_n. \tag{1.9}$$

Let us give an intuitive interpretation of Condition (1.9). First, notice that when $\mathcal{Q} = \mathcal{M}_1^+(\boldsymbol{\Theta})$, then the restriction Q_n of Π_0 to $\mathcal{B}(P_0, r_n)$ belongs to $\mathcal{Q}$, and it is easy to see that the extended prior mass condition (1.9) is equivalent to the former prior mass condition (1.8) of Ghosal et al. (2000). Moreover, each inequality in Condition (1.9) play its own role. The second one characterizes the rate of convergence of the exact posterior, while the first one characterizes the approximation error given by the variational family $\mathcal{Q}$. In particular, a large set $\mathcal{Q}$ means a high expressive power given by the variational approximation. Hence the associated integral $\int \mathrm{KL}(P_0\|P_{\boldsymbol{\theta}})Q_n(d\boldsymbol{\theta})$ is small and the rate of convergence is fully determined by the second part of Condition (1.9), i.e. the concentration of the posterior.

Note that deriving such prior mass conditions in practice is a major difficulty. More especially, it depends on the model and the prior, and must be treated on a case-by-case

basis, see for instance Ghosal and Van Der Vaart (2001) for the estimation of Gaussian mixtures or Rivoirard and Rousseau (2012) for the estimation of densities in the class of Sobolev or Besov spaces. Such computations are addressed in this thesis.

The following theorem from Alquier and Ridgway (2017) presents a result of convergence that implies the concentration of the α-tempered posterior and of its variational approximation to the true distribution in α-Rényi divergence. A similar result can be found in Bhattacharya et al. (2018). We will not present the result for regular posteriors, and we refer the reader to Zhang and Gao (2017) for a precise statement involving testing conditions.

We can now express a variant of Theorem 2.6 and 2.7 of Alquier and Ridgway (2017):

Theorem 1.1.2. *Assume that Assumption (1.9) is satisfied. Then for any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{\theta}}, P_0)\widetilde{\Pi}_{n,\alpha}(d\boldsymbol{\theta})\right] \leq \frac{1+\alpha}{1-\alpha} r_n$$

In particular, using Markov's inequality, we have the concentration of the variational approximation of the posterior:

$$\widetilde{\Pi}_{n,\alpha}\left(\boldsymbol{\theta} \in \boldsymbol{\Theta} \,\middle/\, d(P_{\boldsymbol{\theta}}, P_0) > M_n r_n\right) \xrightarrow[n\to+\infty]{} 0$$

in P_0-probability as $n \to +\infty$ for any $M_n \to +\infty$. Moreover, under a slight modification in the first inequality of Condition (1.9), we get even in case of misspecification:

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{\theta}}, P_0)\widetilde{\Pi}_{n,\alpha}(d\boldsymbol{\theta})\right] \leq \frac{\alpha}{1-\alpha} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{KL}(P_0 \| P_{\boldsymbol{\theta}}) + \frac{1+\alpha}{1-\alpha} r_n. \tag{1.10}$$

Again, we remind that the rate of convergence is the one defined in the prior mass condition (1.9). It is worth mentioning that the oracle inequality (1.10) is not sharp because the divergence inequality on the left hand side is smaller than the one on the right hand side, illustrating the fact that the convergence is obtained in Rényi divergence D_α whereas the prior mass condition is required with respect to the KL one. Nevertheless, such an asymmetry is unavoidable. For instance, Zhang and Gao (2017) obtain convergence in KL divergence, which may seem more interesting, but requires at the same time a prior mass condition with respect to a β-Rényi divergence with $\beta > 1$, which is stronger than the one presented here. Nevertheless, both divergences are close very often in practice, and hence the oracle inequality remains informative and leads to optimal rates of convergence in most situations.

Note that Alquier and Ridgway (2017) do not tackle the case of models with hidden variables, in particular mixture models, which are very popular in the variational Bayes community. Similarly, the application of their results to high dimensional models such as neural networks is not straightforward. Also, the ELBO is widely used as a numerical criterion for model selection, but it has never been justified in theory. A significant part of this thesis aims at filling this gap, and we present in Chapters 4 to 6 three papers (Chérief-Abdellatif and Alquier, 2018; Chérief-Abdellatif, 2019a,b) that extend the results of Alquier and Ridgway (2017) to such models and contexts. We refer to Section 2.1 for a more detailed overview of our contributions.

1.2 Online learning: convexity and Bayes' rule

Another popular setting in statistics and machine learning is the *online learning* framework. It deals with sequential decision making in situations where observations are revealed one after another and no specific probabilistic assumption regarding the origin of the sequence of observations is made. Contrary to the usual batch case where the best predictor is selected by learning on the whole training data at once, online learning algorithms are designed to dynamically adapt to new observations and update the predictions in a sequential manner. Due to the non-stochastic nature of the problem, most guarantees provided on these algorithms are deterministic, and thus remain available in the worst case, in the presence of an adversary or under a simpler stochastic assumption on the data.

In the rest of this section, we will formalize the online learning setting and motivate its study. Then we will highlight the importance of convexity in online learning, and briefly present and analyze the main ideas of some popular algorithms used in online convex optimization. Finally we will explain how Bayesian inference can be formulated in a sequential manner and derive the associated theoretical bounds.

1.2.1 The online learning setting

Online learning can be seen as an extension of the classical statistical learning setting, where we get rid of the usual probabilistic assumptions such as i.i.d. data and where we try to propose decisions sequentially, with only one observation being revealed at each step. This change of paradigm deeply impacted the statistics and the machine learning communities and built a bridge between learning and other mathematical fields such as convex optimization and game theory.

Formally, the online learning framework can be described as follows. Lower case notations are used to stress the absence of probabilistic assumption regarding the data generating process. At each step t:

- Choose $\widehat{\boldsymbol{\theta}}_t$ from a parameter set $\boldsymbol{\Theta}$.
- Observe a datapoint x_t.
- Suffer a loss $\ell(x_t, \widehat{\boldsymbol{\theta}}_t)$.

Let us make some remarks. First, the parameter set $\boldsymbol{\Theta}$ from which we choose $\widehat{\boldsymbol{\theta}}_t$ may be allowed to change at each step, but we do not make this assumption here for simplicity. Besides, at each step, the decision $\widehat{\boldsymbol{\theta}}_t$ is taken using the whole past dataset $\mathcal{D}_{t-1} := \{x_1, x_2, \dots, x_{t-1}\}$, and the quality of $\widehat{\boldsymbol{\theta}}_t$ is defined through a loss function $\ell_t(\widehat{\boldsymbol{\theta}}_t) := \ell(x_t, \widehat{\boldsymbol{\theta}}_t)$. We consider the full-information setting where the entire loss function ℓ_t is observed (as opposed to the bandit setting, where is it only partially observed, that will not be considered in this thesis).

As for the statistical learning protocol, this online formulation is very general and encompasses various problems of interest. We detail some of them in the following.

Example 1.2.1 (Regression and classification)**.** *Let us consider the case of supervised learning. We observe a dataset* $\mathcal{D}_T = (z_t, y_t)_{t=1,...,T}$ *sequentially, without making the assumption that they are independent realizations of some true distribution* P_0 *over* $\mathcal{Z} \times \mathcal{Y}$*. Given a set of predictors* $\{f_{\boldsymbol{\theta}} : \mathcal{Z} \mapsto \mathcal{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$*, the loss function measuring the discrepancy between each prediction* $f_{\boldsymbol{\theta}}(z_t)$ *and each observed output* y_t *depends on the nature of the problem. For instance, we may choose the hinge loss* $\ell_t(\boldsymbol{\theta}) = (1 - y_t f_{\boldsymbol{\theta}}(z_t))_+$ *in classification and the square loss* $\ell_t(\boldsymbol{\theta}) = (y_t - f_{\boldsymbol{\theta}}(z_t))^2$ *in regression.*

Example 1.2.2 (Density estimation)**.** *The goal is to estimate a probability distribution* P_0 *using independent realizations* $x_1, ..., x_T$*. The model is a set of probability distributions indexed by the parameter set* $\boldsymbol{\Theta}$*, or more precisely a set of probability densities* $\{p_{\boldsymbol{\theta}} / \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *with respect to some reference measure over* $\mathcal{X}$ *equipped with some suited* σ*-algebra* $\mathcal{A}$*. The loss function for this problem is the log-loss* $\ell_t(\boldsymbol{\theta}) := -\log p_{\boldsymbol{\theta}}(x_t)$*. Note that the maximum likelihood estimator* $\widehat{\boldsymbol{\theta}}_T$ *is the parameter which minimizes the cumulative losses* $\ell_t(\boldsymbol{\theta}) := -\log p_{\boldsymbol{\theta}}(x_t)$ *until step* $T - 1$*.*

Example 1.2.3 (Prediction with Expert Advice)**.** *In this example, the learner makes a prediction at each step from the advice of* K *given experts. Those experts may have access to other sources of information to make their own predictions* $\widehat{x}_t(k) \in \mathcal{X}$*,* $k = 1, ..., K$*. This case was the first problem studied in online learning and was widely studied in the community, see Cesa-Bianchi and Lugosi (2006) for more details. The goal is to perform at each step* t *as well as the best expert by taking a convex combination of the expert advice* $\sum_{k=1}^K \boldsymbol{\theta}_t(k)\widehat{x}_t(k)$ *where the parameter* $\boldsymbol{\theta}_t$ *can be interpreted as a level of confidence the learner has in each of the experts. Hence, the prediction with expert advice can just be seen as an online optimization problem over the* $(K-1)$*-dimensional simplex* $\boldsymbol{\Theta} := \{\boldsymbol{\theta} \in \mathbb{R}_+^K / \sum_{k=1}^K \boldsymbol{\theta}(k) = 1\}$*. Given the true observation* x_t *and a discrepancy measure* $d(a, b)$ *between points* a *and* b *in* $\mathcal{X}$*, then the loss measure is simply defined as* $\ell_t(\boldsymbol{\theta}) = d(\sum_{k=1}^K \boldsymbol{\theta}_t(k)\widehat{x}_t(k), x_t)$*.*

The goal in online learning is to design an algorithm that selects a sequence of decisions $(\widehat{\boldsymbol{\theta}}_t)_{t=1}^T$ minimizing the cumulative loss $\sum_{t=1}^T \ell_t(\widehat{\boldsymbol{\theta}}_t)$. However, the usual measure of performance that is taken in the online learning community is not exactly the cumulative loss but a slight variant of it, called *cumulative regret* or simply *regret* $\mathcal{R}_T$, and which compares the cumulative loss of the algorithm to the smallest cumulative loss that could have been reached in hindsight with a *fixed* parameter:

$$\mathcal{R}_T = \sum_{t=1}^T \ell_t(\widehat{\boldsymbol{\theta}}_t) - \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=1}^T \ell_t(\boldsymbol{\theta}). \tag{1.11}$$

Minimizing the regret takes into account the difficulty of the problem by comparing the performance of an algorithm to the best performance that it could have achieved in hindsight. A sub-linear value of the regret i.e. such that $\mathcal{R}_T/T \to 0$ as T goes to infinity, is a minimal requirement.

Obviously, even though we do not make any special assumption *on a data generating process*, it is not possible to build a general theory without any assumption at all. In such a situation, the adversary could make the cumulative loss of our algorithm arbitrarily large. Let us for instance consider the case of a game with two possible answers, where the

adversary waits for our decision and chooses the opposite answer as the correct solution. Obviously, there exists an answer that leads to less than $T/2$ mistakes over T rounds, and thus all algorithms perform poorly, even if we take into account the difficulty of the problem and compare the performance of the algorithm to the best fixed strategy in hindsight. Consequently, the power of the adversarial environment is restricted in practice, typically by making some assumptions on the loss functions ℓ_t, such as *convexity* or *Lipschitzness*.

Note that online learning can sometimes give alternative estimators for solving stochastic learning problems. For instance, when considering i.i.d. data and an algorithm $(\widehat{\boldsymbol{\theta}}_t)_{t=1}^T$ that leads to a small regret, then the *online-to-batch* conversion technique provides an estimator $\bar{\boldsymbol{\theta}}_T = \frac{1}{T}\sum_{t=1}^T \widehat{\boldsymbol{\theta}}_t$ with small generalization risk $R(\boldsymbol{\theta}) = \mathbb{E}[\ell(X, \boldsymbol{\theta})]$ as soon as losses $\ell(x, \cdot)$ are convex for any x (Shalev-Shwartz, 2012). More precisely:

$$\mathbb{E}_{\mathcal{D}_T \sim P_0}[R(\bar{\boldsymbol{\theta}}_T)] \leq \mathbb{E}_{\mathcal{D}_T \sim P_0}\left[\frac{1}{T}\sum_{t=1}^T \ell_t(\widehat{\boldsymbol{\theta}}_t)\right]$$

where for purposes of notation, $\mathcal{D}_T \sim P_0$ means that all observations are independent realizations of a variable $X \sim P_0$. Hence, studying learning algorithms from an online perspective without any stochastic assumption on the data in the worst-case scenario can lead to new estimation procedures with strong theoretical guarantees.

The aim of the next section is to present the main ideas of some popular algorithms in online convex learning and to study the value of the regret $\mathcal{R}_T$ that such algorithms can achieve in terms of the number of rounds T, the geometry of the parameter set $\boldsymbol{\Theta}$, and under different assumptions on the loss functions ℓ_t in addition to convexity. Then we will formulate Bayes' rule in an online fashion and present versions of regret bounds that are suited to the analysis of Bayesian-based online algorithms.

1.2.2 Online convex optimization

Building upon the setting of Section 1.2.2, this section is devoted to the exploration of two of the most popular algorithms in online convex learning that are follow-the-regularized-leader (FTRL) and online projected gradient descent (OPGD). In particular, the objective is to emphasize the importance of convexity for designing efficient online learning algorithms, to provide an overview of the main ideas behind such algorithms, and to give the mathematical tools and techniques in order to derive sharp regret bounds.

Convexity and online learning

Convexity is known to play a central role in optimization. Indeed, even though finding the global optimum of a function can be a very difficult task in some situations, it is possible to find efficiently such a global solution in many other cases for a special class of optimization problems referred to as convex optimization problems. Here, "efficiently" means both from a theoretical and from a practical points of view: we can solve the problem in a reasonable amount of time, and this in time depending only polynomially on the problem size.

Similarly, the online version of convex optimization as introduced by Zinkevich (2003) can be seen as a particular instance of online learning and leads to algorithms that achieve a small regret while being efficient from a computational perspective. Moreover, various problems and many algorithms used in online learning can be analyzed under this setting. *In the remainder of this section, we will consider a convex online optimization framework, which is characterized by a convex domain* $\boldsymbol{\Theta}$ *and convex losses* ℓ_t *with respect to* $\boldsymbol{\theta}$, and we will show how this setting can be used to derive efficient online algorithms.

Follow the leader ?

Let us begin with a basic algorithm called *follow-the-leader* (FTL), whose name is due to Kalai and Vempala (2005). The idea (follow the leader !) is very simple and consists in predicting the parameter $\boldsymbol{\theta}$ that minimizes the past cumulative loss $\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta})$. This strategy is very natural and follows the same idea than empirical risk minimization in batch machine learning.

Algorithm 1 Follow-The-Leader (FTL)

Initialize $\boldsymbol{\theta}_1$.
for $t = 1, \dots, T$ **do**
  The function ℓ_t is revealed,
  Update $\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{s=1}^{t} \ell_s(\boldsymbol{\theta})$.
end for

Nevertheless, this algorithm is impractical because we must store all the past data and recompute their gradients at each step to get the minimizer of the convex objective, and can perform very poorly in theory, getting the worst value of the regret. Here is a simple example taken from Shalev-Shwartz (2012) to illustrate this.

Example 1.2.4 (Failure of FTL)**.** *Let us consider a game with parameter set* $\boldsymbol{\Theta} = [-1, 1]$, *loss functions* $\ell_t(\boldsymbol{\theta}) = \boldsymbol{\theta} x_t$ *and observations* x_t *such that:*

$$x_t = \begin{cases} -0.5 & \text{if } t = 1, \\ +1 & \text{if } t \text{ is even}, \\ -1 & \text{otherwise.} \end{cases}$$

Hence, the predictions of the FTL algorithm initialized with $\boldsymbol{\theta}_1 = 0$ *are* $\boldsymbol{\theta}_t = -1$ *for odd values of* $t > 1$ *and* $\boldsymbol{\theta}_t = 1$ *for even values of* t, *and their cumulative losses will be* $T - 1$ *while the cumulative loss of the constant strategy* $\boldsymbol{\theta}_t = 0$ *gives a cumulative loss of 0 !*

The explanation of the limitation of FTL in the previous context is that the predictions are *not stable*, and may change drastically from one round to another when only one single loss function is added. A simple way to fix this is to *regularize* the objective.

Regularize the leader !

The idea of regularization is natural, and is the same than the one used in batch statistical learning. As regularization in batch statistics prevents the learning algorithm from

overfitting, regularization in online learning stabilizes the predictions and prevents the algorithm from being fooled by an adversary. Obviously, different regularizers will lead to different algorithms and different regret bounds.

The most general formulation of the RFTL algorithm is given in Algorithm 2 using a learning rate α. *The regularization function R is assumed to be positive, differentiable and σ-strongly convex on $\boldsymbol{\Theta}$*, which means that for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$:

$$R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}') - (\boldsymbol{\theta} - \boldsymbol{\theta}')^T \nabla R(\boldsymbol{\theta}') \geq \frac{\sigma \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2}.$$

Roughly speaking, there exists a quadratic lower bound on the growth of a strongly convex function. This assumption is very important, as will be seen later.

Algorithm 2 Follow-The-Regularized-Leader (FTRL)

Require: Learning rate α, Regularizer R.
  Initialize $\boldsymbol{\theta}_1$
  for $t = 1, \dots, T$ **do**
    The function ℓ_t is revealed,
    Update $\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{s=1}^{t} \ell_s(\boldsymbol{\theta}) + \frac{R(\boldsymbol{\theta})}{\alpha} \right\}$.
  end for

Remark 1.2.1. *Note that the FTRL algorithm is invariant to any positive constant added to the regularizer R, so we can choose R such that* $\inf R = 0$ *without loss of generality.*

We turn now to the theoretical analysis of FTRL. Lemma 1.2.1 gives an upper bound on the regret relative to any fixed parameter (not necessarily the best one).

Lemma 1.2.1. *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) \leq \sum_{t=1}^{T} [\ell_t(\boldsymbol{\theta}_t) - \ell_t(\boldsymbol{\theta}_{t+1})] + \frac{R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}_1)}{\alpha}.$$

Hence, the regret for a fixed strategy $\boldsymbol{\theta}$ is upper bounded by two terms, the first one being the cumulative differences between the losses in two consecutive predictions of the algorithm, while the second one comes from the regularizer. Intuitively, the first term, which alone is also an upper bound for the FTL algorithm, will be small if $\boldsymbol{\theta}_t \approx \boldsymbol{\theta}_{t+1}$ under a smoothness assumption on the loss functions ℓ_t. This quantity can be controlled for the FTRL algorithm (contrary to the FTL case) because of the regularization term that acts as a stabilizer, at the price of the additional term $(R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}_1))/\alpha$.

The following lemma quantifies our intuition on the stabilization phenomenon and ensures that for a strongly convex regularizer and a Lipschitz loss, successive points $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ are close, and then the differences between the losses in two consecutive predictions are upper bounded.

Lemma 1.2.2. *For L-Lipschitz losses ℓ_t,*

$$\ell_t(\boldsymbol{\theta}_t) - \ell_t(\boldsymbol{\theta}_{t+1}) \leq L \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\| \leq \frac{\alpha L^2}{\sigma}.$$

Hence, we have the following regret bound for the FTRL algorithm:

Theorem 1.2.3. *For L-Lipschitz losses ℓ_t,*

$$\mathcal{R}_T \leq \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R(\boldsymbol{\theta})}{\alpha} + \frac{\alpha T L^2}{\sigma}.$$

In particular, if the regularizer is upper bounded by some constant $B^2/2$, then for $\alpha = \frac{B\sqrt{\sigma}}{L\sqrt{2T}}$,

$$\mathcal{R}_T \leq \frac{BL\sqrt{2T}}{\sqrt{\sigma}}.$$

Hence, the FTRL algorithm can achieve a sub-linear worst case regret of order $\sqrt{T}$ for convex and Lipschitz loss functions, which is known to be optimal with respect to T. Note that dependence on the dimension is hidden in the Lipschitz constant L and in the upper bound B.

Online projected gradient descent

Unfortunately, as for the FTL algorithm, FTRL is impractical as solving the optimization problem at each step requires storing all the past data and recomputing their gradients. Actually, it is possible to get rid of this problem by using an approximation when the losses are differentiable. The online projected gradient descent algorithm is based on this idea: at each step, the past convex losses $\ell_s(\boldsymbol{\theta})$ are replaced by their linear approximations $\nabla_{\boldsymbol{\theta}}\ell_s(\boldsymbol{\theta}_s)^T\boldsymbol{\theta}$ at the previous points $\boldsymbol{\theta}_s$. The gradients being computed at the past $\boldsymbol{\theta}_s$, then one just needs to store the past gradients $\nabla_{\boldsymbol{\theta}}\ell_s(\boldsymbol{\theta}_s)$ only at each step, and the algorithm will be efficiently performed as soon as computing the gradient of the regularization term is cheap. Moreover, the convexity of the losses ensures that for any $\boldsymbol{\theta}$:

$$\sum_{t=1}^{T}[\ell_t(\boldsymbol{\theta}_t) - \ell_t(\boldsymbol{\theta})] \leq \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}}\ell_t(\boldsymbol{\theta}_t)^T[\boldsymbol{\theta}_t - \boldsymbol{\theta}],$$

and as the previous analysis of the regret still holds for linear losses with bounded gradients under the assumption that $\boldsymbol{\Theta}$ is bounded, then linearization exactly leads to the same regret bound than Algorithm 2.

The basic regularization for a convex set $\boldsymbol{\Theta} \subset \mathbb{R}^d$ is the Euclidean penalty $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_1\|^2/2$ for an arbitrary point $\boldsymbol{\theta}_1$. Using the linearization trick, we obtain (the lazy version of) the famous online gradient algorithm 3 (Zinkevich, 2003) which is very easy to compute in practice, with a possible projection step when $\boldsymbol{\Theta}$ is not the whole space $\mathbb{R}^d$, in which case the update may not belong to $\boldsymbol{\Theta}$. $\boldsymbol{\Theta}$ is also chosen to be closed so that the projection is well defined.

OPGD consists in updating the prediction of the algorithm at each round by moving in the negative direction of the gradient of the observed loss and projecting back onto $\boldsymbol{\Theta}$. It is similar to stochastic gradient descent albeit not exactly the same algorithm, as the loss functions are different at each step.

Under a Lipschitzness assumption on loss functions ℓ_t and a boundedness assumption on the convex set $\boldsymbol{\Theta}$, we have the following regret bound for Algorithm 3:

Algorithm 3 Online Projected Gradient Descent (OPGD) [Lazy version]

Require: Learning rate α.
  Initialize $\boldsymbol{\theta}_1$
  for $t = 1, \dots, T$ **do**
    The function ℓ_t is revealed,
    Update $\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \alpha \nabla_{\boldsymbol{\theta}} \ell_t(\boldsymbol{\theta}_t)$,
    Project $\widetilde{\boldsymbol{\theta}}_{t+1}$ onto $\boldsymbol{\Theta}$ in order to obtain $\boldsymbol{\theta}_{t+1}$.
  end for

Theorem 1.2.4. *For a closed set $\boldsymbol{\Theta}$ and differentiable loss functions f_t,*

$$\mathcal{R}_T \leq \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{\theta}\|_2^2}{2\alpha} + \alpha \sum_{t=1}^{T} \|\nabla_{\boldsymbol{\theta}} \ell_t(\boldsymbol{\theta}_t)^2\|_2^2.$$

Moreover, if f_t's are L-Lipschitz and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\|\boldsymbol{\theta}\| \leq B$, then for $\alpha = \frac{B}{L\sqrt{2T}}$,

$$\mathcal{R}_T \leq BL\sqrt{2T}.$$

Note that there exists a more popular variant of the lazy version of OGPD presented in Algorithm 3 referred to as the *agile version* of OPGD that we give in Algorithm 4. The lazy version keeps track of the point $\widetilde{\boldsymbol{\theta}}_t$ and projects onto $\boldsymbol{\Theta}$ only at prediction time, while the agile version preserves the feasible point $\boldsymbol{\theta}_t$ at all times. This agile version can be shown to give a similar regret than the lazy one, but its analysis goes beyond the scope of this thesis.

Algorithm 4 Online Projected Gradient Descent (OPGD) [Agile version]

Require: Learning rate α.
  Initialize $\boldsymbol{\theta}_1$
  for $t = 1, \dots, T$ **do**
    The function ℓ_t is revealed,
    Update $\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t - \alpha \nabla_{\boldsymbol{\theta}} \ell_t(\boldsymbol{\theta}_t)$,
    Project $\widetilde{\boldsymbol{\theta}}_{t+1}$ onto $\boldsymbol{\Theta}$ in order to obtain $\boldsymbol{\theta}_{t+1}$.
  end for

In this section, we introduced some basic tools and results from online convex optimization. Of course, most algorithms and analyses presented here can be extended. For instance, it is possible to let the learning rate α vary with the step t in order to achieve smaller regret bounds in $\mathcal{O}(\log T)$ for strongly convex losses. We refer the interested reader to Bubeck (2011); Shalev-Shwartz (2012); Hazan (2016); Cesa-Bianchi and Lugosi (2006) for more details and developments on the methods described in this section.

In the rest of this chapter, we will no longer be working in the convex online optimization setting, and we will show how to get regret bounds for Bayesian inference in online learning.

1.2.3 Bayes in online learning

An appealing property of Bayesian inference is that the natural representation of the past information provided by past data can be updated in a sequential manner using the Bayes' rule as new data become available. This approach is particularly suited to the online learning setting for designing online algorithms when no probabilistic modeling of the data is assumed, as generalization of the Bayes principle can be used via loss functions $\ell_t(\boldsymbol{\theta})$, see Section 1.1. Moreover, the theoretical analysis holds even in case of misspecification or in the presence of some adversarial data.

In this section, we will show that algorithms arising from generalized Bayesian inference, often referred to as *Exponentiated Weighted Aggregation* in the online setting, achieve low regret (Banerjee, 2006; Audibert, 2009; Gerchinovitz, 2013) and match the ones obtained by classical online learning methods presented in the previous section.

EWA: a Bayesian strategy

Exponentially Weighted Aggregation (EWA) is a well-known strategy in sequential prediction and online optimization. We adopt the same Bayesian approach and notations than in Section 1.1. We define the set $\mathcal{M}_1^+(\boldsymbol{\Theta})$ of all probability measures on $\boldsymbol{\Theta}$ (equipped with some suitable σ-algebra), and the prior distribution $\Pi_0 \in \mathcal{M}_1^+(\boldsymbol{\Theta})$. EWA is given by Algorithm 5.

Algorithm 5 Exponentiated Weighted Aggregation (EWA)

Require: Learning rate $\alpha > 0$, prior probability distribution Π_0.
  Initialize $\Pi_{0,\alpha} = \Pi_0$.
  for $t = 1, \dots, T$ **do**
    The function ℓ_t is revealed,
    Update $\Pi_{t+1,\alpha}(d\boldsymbol{\theta}) \propto \exp(-\alpha \ell_t(\boldsymbol{\theta}))\Pi_{t,\alpha}(d\boldsymbol{\theta})$.
  end for

It is important to remark that when $\alpha = 1$ and $\ell_t(\boldsymbol{\theta})$ is a log-likelihood, then this is just standard Bayesian inference. In this case we assume that $\ell_t(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(x_t)$ and the x_t's are independent realizations of a distribution with density $p_{\boldsymbol{\theta}_0}$, where $\boldsymbol{\theta}_0$ is unknown. By doing so, we have:

$$\Pi_{t+1,1}(d\boldsymbol{\theta}) = \frac{\prod_{i=1}^t p_{\boldsymbol{\theta}}(x_t)\Pi_0(d\boldsymbol{\theta})}{\int \prod_{i=1}^t p_{\vartheta}(x_t)\Pi_0(d\vartheta)},$$

which is exactly the definition of the regular posterior distribution. As mentioned in Section 1.1, the posterior $\Pi_{t,1}$ might not concentrate around the best approximation of the generating distribution of the x_t's in misspecified models, whereas a suitable $\alpha < 1$ in $\Pi_{t,\alpha}$ leads to robust estimation (Grünwald et al., 2017; Bhattacharya et al., 2016).

Regret bound for EWA

In the online optimization literature, it is actually known that such results for $\alpha < 1$ will hold even without stochastic assumption on the observations, and even in adversarial

settings.

Theorem 1.2.5. *Assume that for any* $\boldsymbol{\theta}$, $0 \leq \ell_t(\boldsymbol{\theta}) \leq B$. *Then*

$$\sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}_t) \Pi_{t,\alpha}(d\boldsymbol{\theta}_t) \leq \inf_{Q \in \mathcal{M}_1^+(\boldsymbol{\Theta})} \left\{ \sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}) Q(d\boldsymbol{\theta}) + \frac{\alpha B^2 T}{8} + \frac{\mathrm{KL}(Q \| \Pi_0)}{\alpha} \right\}.$$

The proof of Theorem 1.2.5 can be found in Cesa-Bianchi and Lugosi (2006) for a finite parameter set $\boldsymbol{\Theta}$, and its general version is a special case of Theorem 4.6 in Audibert (2009), see also Gerchinovitz (2013). Note that the integral on the left hand side is the cumulated loss of the algorithm averaged over the random EWA strategy, and is compared to an integral term on the right hand side which is the smallest cumulative loss that could have been reached in hindsight averaging over a *fixed* random strategy. Therefore, the inequality above is also referred to as a *regret bound*, as for upper bounds on the true *regret* (1.11). Hence, this regret bound is a counterpart in the online framework of the frequentist evaluation of regular and generalized Bayesian distributions, and guarantees such as concentration or generalization bounds can be recovered under additional assumptions (e.g. prior mass condition or convexity of the loss function) using online-to-batch techniques.

Example 1.2.5 (Finite case)**.** *The study of algorithm 5 in the finite case* $\mathrm{card}(\boldsymbol{\Theta}) = M$ *goes back to Vovk (1990); Littlestone and Warmuth (1994). It is for instance particularly suited to the prediction with expert advice setting presented in Example 1.2.3. Taking* Π_0 *as a uniform distribution over* $\boldsymbol{\Theta}$ *leads to the bound*

$$\sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}_t) \Pi_{t,\alpha}(d\boldsymbol{\theta}_t) \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) + \frac{\alpha B^2 T}{8} + \frac{\log(M)}{\alpha}.$$

Thus,

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{B\sqrt{T \log(M)}}{\sqrt{2}}$$

for the choice $\alpha = \sqrt{8 \log(M)/(TB^2)}$ *where the expectation is an average over the random EWA strategy.*

Example 1.2.6 (Link with Bayesian literature)**.** *Another interesting setting is when each* $\ell_{\boldsymbol{\theta}}$ *is* L*-Lipschitz. Define* $\Pi_{\tau,\varepsilon}$ *as* Π_0 *restricted to* $\mathcal{B}(\tau, \varepsilon) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\tau - \boldsymbol{\theta}\| \leq \varepsilon\}$. *Then obviously* $\int \ell_t(\boldsymbol{\theta}) \Pi_{\tau,\varepsilon}(d\boldsymbol{\theta}) \leq \ell_t(\tau) + L\varepsilon$, *and the bound becomes*

$$\sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}_t) \Pi_{t,\alpha}(d\boldsymbol{\theta}_t) \leq \inf_{\tau \in \boldsymbol{\Theta}, \varepsilon > 0} \left\{ \sum_{t=1}^{T} \ell_t(\tau) + T\left(\frac{\alpha B^2}{8} + L\varepsilon\right) + \frac{-\log \Pi_0(\mathcal{B}(\tau, \varepsilon))}{\alpha} \right\}.$$

Define $\boldsymbol{\theta}^*$ *as a minimizer with respect to* τ *of* $\sum_{t=1}^{T} \ell_t(\tau)$. *Then we make the assumption* $r(\varepsilon) \geq -\log \Pi_0(B(\boldsymbol{\theta}^*, \varepsilon))$ *which is to be compared with the usual prior mass condition* (1.8) *made in the frequentist evaluation of Bayesian methods literature. Then one has*

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \inf_{\varepsilon > 0} \left\{ T\left(\frac{\alpha B^2}{8} + L\varepsilon\right) + \frac{r(\varepsilon)}{\alpha} \right\}.$$

In a parametric model $M \subset \mathbb{R}^d$, typically $r(\varepsilon) \sim d\log(1/\varepsilon)$. The choices $\epsilon = d/(TL\alpha)$ and $\alpha = \sqrt{d/(TB^2)}$ then lead to the regret bound

$$\mathbb{E}\left[\mathcal{R}_T\right] = \mathcal{O}\left(B\sqrt{dT\log\left(\frac{LT}{Bd}\right)}\right).$$

Note that it is also possible to provide bounds with large probability instead of bounds in expectation thanks to Hoeffding-Azuma inequality. We do not detail the technique here and refer the interested reader to Cesa-Bianchi and Lugosi (2006) for further explanations. Moreover, when the ℓ_t's are convex, we have $\int \ell_t(\boldsymbol{\theta})\Pi_{t,\alpha}(d\boldsymbol{\theta}) \geq \ell_t\left(\int \boldsymbol{\theta}\Pi_{t,\alpha}(d\boldsymbol{\theta})\right)$, so replacing the random character of EWA by the average $\widehat{\boldsymbol{\theta}}_t = \int \boldsymbol{\theta}\Pi_{t,\alpha}(d\boldsymbol{\theta})$ leads to the deterministic regret bound

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{\theta}}_t) \leq \inf_{Q\in\mathcal{M}_1^+(\boldsymbol{\Theta})}\left\{\sum_{t=1}^{T}\int \ell_t(\boldsymbol{\theta})Q(d\boldsymbol{\theta}) + \frac{\alpha B^2 T}{8} + \frac{\mathrm{KL}(Q\|\Pi_0)}{\alpha}\right\}.$$

Note that in the examples above, the choice of α depends on the long-term horizon T but not on the current step t, as in batch Bayesian inference where α may possibly depend on the sample size n but is fixed at once. As for classical online learning algorithms presented in the previous section, the temperature parameter α could be thought of as a learning parameter allowed to vary with t rather than a fixed value (as it is the case in usual Bayesian inference). Similar results for α depending on t instead of T are discussed in Cesa-Bianchi and Lugosi (2006). For example, it is known that under additional assumptions on ℓ_t, regrets in $\mathcal{O}(\log T)$ or even $\mathcal{O}(1)$ can be reached by letting the learning rate vary. This is for example the case when each ℓ_t is ζ-exp-concave on $\boldsymbol{\Theta}$, that is, when $\boldsymbol{\theta} \mapsto \exp(-\zeta\ell_t(\boldsymbol{\theta}))$ is concave. Conversely, such fast rates cannot be achieved for a fixed value of α. Hence, relaxing the purely Bayesian approach may be a convenient way to design algorithms that have optimal guarantees in various situations while staying inspired by the same underlying Bayesian principles.

Unfortunately, EWA distributions are often intractable, and classical variational techniques presented in the beginning of this thesis are no longer efficient to solve this problem as it would require computing the whole variational approximation from scratch at each step and skip the online aspect of the problem. Nevertheless, applying efficient algorithms from the online convex setting to the parameter of a parametric variational family (e.g. online projected gradient descent to the mean-standard deviation parameter of a Gaussian variational family), and considering the associated distributions obtained at each step as a sequential approximation of EWA could be an interesting perspective. We present in Chapter 7 a paper (Chérief-Abdellatif et al., 2019) that proposed the first theoretical analysis of VI in the online learning framework with streaming data. The contributions of this work are detailed further in Section 2.2.

1.3 Robustness to misspecification

The last section presented the online learning paradigm which does not assume that data are generated following any particular process, but rather consider the (possibly adversarial) world as being able to change at each step. This is a very different approach from the one usually taken by statisticians, who want to represent some unknown phenomenon and assume that their statistical model contains the data generating process that models the underlying phenomenon they want to describe.

Nevertheless, real life and scientific reasoning are somewhere in between, as summarized by the famous aphorism "all models are wrong, but some are useful" commonly attributed to the statistician George Box. To quote him, "it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an ideal gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules. For such a model there is no need to ask the question *is the model true*? If *truth* is to be the *whole truth* the answer must be *no*. The only question of interest is *is the model illuminating and useful?*". To be more precise, the question could - and should - be, *is it possible to build an inference method that does not lead to poor results when such a useful but misspecified model is used*?

In this manuscript, we are interested in *robustness to misspecification* and in the design and the study of estimation procedures that satisfy in a misspecified setting (almost) the same properties than in the well-specified setting. In Section 1.3.1, we will introduce some motivating examples, in particular the case of contamination and dependency in the data. In Section 1.3.2, we will present the general setting, and review some literature dealing with the two previous notions. We will finally investigate the minimum distance estimation approach in Section 1.3.3, that will be particularly exploited in the rest of this thesis, see Chapter 8 and Chapter 9.

1.3.1 Motivations

One of the oldest problems in statistics is the design of a universal estimation method with good properties in various settings. Maximum Likelihood Estimation (MLE), introduced by Sir Donald Fisher in the 1920s, is probably the most famous and the first to give a fairly general estimation procedure with strong theoretical guarantees. Indeed, it provides a simple principle that can be applied to several models, and many other estimation procedures can be interpreted as special instances of MLE such as Gauss's least squares estimator for Gaussian errors in regression, or the sample mean for the problem of Gaussian mean estimation. Moreover, MLE comes with strong guarantees from a theoretical side. For instance, the MLE is consistent, asymptotically normal and efficient under mild regularity conditions for parametric models (Le Cam, 1970; Van der Vaart, 2000). Respectively, this means that the estimated parameter converges to the true pa-

rameter (both in probability and almost surely), that the MLE goes to an approximate normal distribution as the sample size goes to infinity, and that the large-sample variance of any other estimator will be no smaller than that of the MLE.

Sensitivity to outliers

Unfortunately, the maximum likelihood is no longer optimal nor even efficient as soon as there are outliers in the data. The following example provides a situation of failure of maximum likelihood estimation in the presence of at least one outlier in the data.

Example 1.3.1 (Instability in presence of outliers)**.** *Consider observations* $x_1, ..., x_n$ *assumed to be independent realizations of a uniform distribution on* $\mathcal{U}([0, \boldsymbol{\theta}_0])$ *where* $\{\mathcal{U}([0, \boldsymbol{\theta}])/\boldsymbol{\theta} > 0\}$ *is the model,* $\boldsymbol{\theta}_0 > 0$ *and* $\mathcal{U}([a, b])$ *is the uniform distribution between* a *and* b*. Unfortunately, there is one outlier in the data: one observation is equal to* 1000 *whereas all the others are between* 0 *and* 1*. In such a model, the maximum likelihood is the maximum of the observations* $x_{(n)} := \max(x_1, ..., x_n)$*, i.e.* $x_{(n)} = 1000$*, while the most reasonable value of* $\boldsymbol{\theta}_0$ *seems to be* 1*... The same appears when considering a univariate Gaussian model rather than uniform distributions, for which the empirical mean is no longer efficient. Hence, maximum likelihood estimation is not robust to outliers.*

In a similar example using probabilistic arguments and modeling the outliers in the data by a mixture generating process, Birgé (2006) showed that the MLE may be inconsistent.

Example 1.3.2 (Inconsistency on approximate models)**.** *Let us consider the generating mixture distribution is* $P_{0,n} = (1 - 2n^{-1})\mathcal{U}([0, 0.1]) + 2n^{-1}\mathcal{U}([0.9, 1])$ *whereas the model is still composed of uniforms* $\mathcal{U}([0, \boldsymbol{\theta}])$ *with* $\boldsymbol{\theta} > 0$*. In this example, we use a probabilistic argument, and thus we do not necessarily observe the data but rather consider a collection of i.i.d. random variables* $X_1, ..., X_n$*. As in the former situation, the MLE is* $X_{(n)}$*, and* $\mathcal{U}([0, 0.1])$ *seems to be a good approximation of the generating distribution* $P_{0,n}$*. To formalize this, notice that the squared Hellinger distance between both distributions is equal to* $H^2(P_{0,n}, \mathcal{U}([0, 1/10])) < 5/4n$ *for* $n \geq 4$*. Thus, one could expect obtaining the consistency of the uniform distribution* $\mathcal{U}([0, X_{(n)}])$ *associated with the MLE* $X_{(n)}$ *in Hellinger distance, i.e.* $\mathbb{E}[H^2(P_n^0, \mathcal{U}([0, X_{(n)}]))]$ *goes to* 0 *as* $n \to +\infty$*. Nonetheless,* $\mathbb{E}[H^2(P_n^0, \mathcal{U}([0, X_{(n)}]))] > 0.38$ *! So the MLE is not even consistent in this situation where the data generating process is very close (in Hellinger distance) to the chosen model.*

Sensitivity to dependence

In many settings, usual inference techniques are based on the i.i.d. assumption. Unfortunately, even in situations where independence seems to hold, such a condition may often be unrealistic in practice, more especially when observations are successive realizations of a stochastic process with temporal ordering. This is for instance particularly the case in social sciences and in economics when considering variables such as inflation, employment and wage that are collected one year after another, and when the past can affect the future. For instance, when dealing with an economic variable Y that we want to explain

using a dependent variable Z, then the simple linear regression model with independent Gaussian errors is widely used while the error terms are often serially correlated.

Example 1.3.3 (First-Order autoregressive error model in linear regression)**.** *Let us consider the following simple linear regression model*

$$Y_t = a + bZ_t + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ *is a noise gathering all factors but* Z*. All* ϵ_t *are independent. Actually, there may be a hidden temporal dependence in the error term* ϵ_t*, which would consist of a fraction* ρ *of the previous error term* ϵ_{t-1} *plus a new disturbance term* U_t*:*

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

with $u_t \sim \mathcal{N}(0, \sigma_u^2)$*.* $\rho \in (-1, 1)$ *is referred to as the autocorrelation parameter, and is exactly equal to the autocorrelation between two consecutive error terms* ϵ_t *and* ϵ_{t+1}*. All* u_t *are independent. Of course, when* $\rho = 0$*, we recover the simple linear regression equation and there is no correlation. In the previous example, we talk about serial correlation or say that the errors suffer from autocorrelation, as they are correlated across time.*

It is important to note that the simple linear model may seem appropriate in the previous example even in presence of serial correlation when drawing a scatterplot. Hence, to answer to Box' question: the linear model may seem illuminating and useful. However, the estimated parameters obtained by MLE (or equivalently by leasts squares) are not efficient anymore: indeed, the BLUE (Best Linear Unbiased Estimator) property stating that MLE achieves the lowest variance on the estimate, as compared to other unbiased and linear estimators, is no longer available as the uncorrelated errors assumption is violated.

Notice that Bayesian inference suffers from the same lack of robustness to misspecification. Indeed, there are situations, for instance when using a (homoskedastic) linear regression model in the presence of heteroskedasticity, where the regular posterior distribution does not concentrate and puts its mass on worse and worse models of ever higher dimension. We do not explore this further and refer the reader to Grünwald et al. (2017) for more details.

Hence, there is a need to look at more robust estimation procedures in both contamination and dependence contexts, but more generally in any misspecified model. The next paragraphs will be devoted to the description of the problem and to a discussion on the previous attempts done to provide efficient estimation techniques for the contamination and dependent settings.

1.3.2 Several notions of misspecification and robustness

In this section, we shall define a measurable space $(\mathbb{X}, \mathcal{X})$ and a collection of n random variables $X_1, ..., X_n$ generated following a stationary process of marginal distribution P_0, which implies in particular that all random variables are identically distributed. Of course, this general setting includes the situation of i.i.d. variables with data generating distribution P_0. We consider a statistical model $\{P_{\boldsymbol{\theta}}/\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed by a parameter space $\boldsymbol{\Theta}$ that may be misspecified.

Outliers contamination

A particular example of misspecification is the case where outliers are present in the dataset. The probabilistic setting is the following. The data generating distribution P_0 is a small variation of a distribution $P_{\boldsymbol{\theta}_0}$ with $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, so that only a small fraction of the observations are not realizations of $P_{\boldsymbol{\theta}_0}$. We make the assumption that the variables $X_1, \dots, X_n$ are independent. For instance, Hüber's original contamination model (Hüber, 1964) is simply described as a mixture $P_0 = (1-\epsilon)P_{\boldsymbol{\theta}_0} + \epsilon Q$ where Q may be any contaminating distribution. More generally, it is possible to define a model, usually referred to as the adversarial contamination model, which does not formulate any assumption on the outliers. More precisely, we consider a collection of i.i.d. random variables from $P_{\boldsymbol{\theta}_0}$ and we remove a fraction ϵ and replace them with any other values. Hence, the true data generating distribution is totally unknown, and the statistician only considers $X_1, \dots, X_n$ where X_i can take any arbitrary value for $i \in \mathcal{O}$, with $\mathcal{O}$ an arbitrary set such that $|\mathcal{O}| \leq \epsilon n$, and $X_i = \widetilde{X}_i$ for $i \notin \mathcal{O}$ with independent random variables $\widetilde{X}_i \sim P_{\boldsymbol{\theta}_0}$.

A more and more popular problem in the statistics and machine learning communities is the robust estimation of a mean $\boldsymbol{\theta}_0$. In particular, the quest of a statistically optimal and computationally tractable estimator of the mean of a Gaussian model $\{P_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, I_d) / \boldsymbol{\theta} \in \mathbb{R}^d\}$ has received an increased interest in the last few years. Obviously, the sample mean is a very bad estimator as it is very sensitive to any outlier, and the other basic estimators that are known to work well in small dimensions such as the coordinatewise median and the geometric median are suboptimal in high dimension, in the sense that they do not reach the minimax rate of convergence $\max(\frac{d}{n}, \epsilon^2)$ with respect to the expected squared Euclidean distance. Chen et al. (2018) even showed that the componentwise median achieves a rate of $\max(\frac{d}{n}, d\epsilon^2)$ in Hüber's contamination model which is only optimal with respect to the contamination ratio ϵ but not to d in high dimension. It is further proved that Tukey's median (Tukey, 1975) is optimal, but unfortunately this estimator is not tractable and even approximate algorithm have an $\mathcal{O}(n^d)$ complexity (Amenta et al., 2000; Chan, 2004).

In 2016, Lai et al. (2016) and Diakonikolas et al. (2016) presented at the same time two concurrent tractable procedures for robust estimation of the mean of a Gaussian distribution. Both works are different but are based on the common idea that if the empirical and the theoretical covariance matrices are close enough for a subsample of points, then the arithmetic mean of this subsample is a good estimator of the theoretical one. Furthermore, the analysis of Diakonikolas et al. (2017) suggests that the additional logarithmic term $\log(1/\epsilon)$ can not be removed in the rate of convergence when seeking a tractable procedure. Further improvements in running time for near-optimal estimation procedures can be found in Diakonikolas et al. (2018a); Diakonikolas and Kane (2019); Cheng et al. (2019). Alternative approaches providing similar or better rates for tractable estimators on weaker contamination models have been investigated in the literature. For instance, Collier and Dalalyan (2017) achieve the minimax rate in Hüber's contamination model without any extra factor when $\epsilon = \mathcal{O}(\min(d^{-1/2}, n^{-1/4}))$ and with an improved overall complexity. More recently, Gao et al. (2019) connects robust Gaussian mean estimation to Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Biau et al., 2018), leading to depth-like estimators that are tractable via stochastic gradient

descent (SGD) algorithms suited to GANs training and that satisfy (almost) the same theoretical robustness properties than Tukey's median in Hüber's contamination model. Other works investigated further the question of computational efficiency, see Hopkins (2019); Cherapanamjeri et al. (2019); Depersin and Lecué (2019), to name but a few.

Dependence in time series

As said previously in this manuscript, one of the main goals of statistics is the understanding of existing phenomena. Consequently, the conditions and the models that statisticians consider is influenced by the real world and by their ability to mimic the behavior of observed data, and it is often unrealistic to assume independence.

Hence, in situations where the goal is to recover some identifiable time invariant structure P_0 using a model $\{P_{\boldsymbol{\theta}}/\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and an i.i.d. assumption, there is a need for a convenient way to describe the data generating process that allows such sequential dependence and for which the behavior of main statistics such as sums of random variables is analogous to those of independent sequences. The *mixing conditions* have been introduced by Rosenblatt (1956) in order to meet this need. Roughly speaking, they impose that the dependence between the present and the distant future is almost zero and can be interpreted as an asymptotic independence. There are many reasons why such mixing conditions have been the dominating assumptions for imposing a restriction on the dependence between time series data. The most important one is that those mixing conditions lead to limit theorems such as the law of large numbers (LLN) and the central limit theorem (CLT), but also to Hoeffding-type inequalities such as McDiarmid's inequality (McDiarmid, 1989) which are originally only available in the i.i.d. setting and can be extended to the dependent setting under a polynomial decay assumption on some mixing coefficients, see Section 1.5 in Doukhan (1994) for LLN and CLT and Rio (2013, 2017b) for a dependent version of McDiarmid's inequality. These results are of particular interest as many inequalities and fundamental properties in independent statistics can be obtained using them, see Chapter 3 in Boucheron et al. (2012). Hence, a lot of classical asymptotic theorems such as almost sure convergence of a sequence of random variables or nonasymptotic concentration inequalities providing rates of convergence can be proved using these extensions, see Doukhan (1994).

Mixing conditions are satisfied for example for hidden Markov models (Baum and Petrie, 1966) under mild assumptions on the transition matrix and on the order of the model. Unfortunately, many other processes of interest in statistics are not mixing. For instance, this is the case of a simple autoregressive process AR(1) (Andrews, 1984). Moreover, checking mixing conditions is not always easy in practice. Hence, Doukhan and Louhichi (1999) proposed an alternative notion of dependence called *weak dependence*. Such dependence coefficients make more explicit the asymptotic independence between the *future* and the *past*, which is progressively forgotten. More precisely, it assumes that the covariance of some wisely chosen functions of the past and the future is small when the distance between the past and the future is large. Also, the notion of weak dependence is more general than mixing as it stands for a wide classes of processes. Finally, many limit theorems and moment inequalities can be obtained for weakly dependent sequences, please refer to Doukhan and Louhichi (1999) for more details.

1.3.3 Minimum distance estimation

First attempts to build a universal estimation procedure using minimum distance estimation (MDE) date back to the 50s and the early work of Wolfowitz (1957). Such estimators are obtained by minimizing some discrepancy between the data and the underlying model. The majority of these divergences are based on probability density functions or on the underlying probability distributions directly. Examples of distances between densities f and g include the famous L_2-distance $\int(f-g)^2$ while the Cramer-von Mises distance $\int(F-G)^2dF$ is a typical distance between cumulative distribution functions F and G. Some metrics, such as the Hellinger distance, is formulated using densities f and g by $\int(\sqrt{f}-\sqrt{g})^2$ but does not depend on the measure with respect to which the density is taken, and consequently can be considered as a distance between probability measures.

Nevertheless, dealing with *density models* rather than *probability models* is not recommended. Indeed, even though the choice of a dominating measure and a density model leads to a probability model, the converse is not true and there are many different representations of a probability model using a reference measure and a set of densities. Furthermore, many statistical methods such as MLE are highly sensitive to the choice of the reference measure, as shown by the example below taken from Baraud and Birgé (2016):

Example 1.3.4 (Dependence on the choice of the density model)**.** *We consider a sequence of i.i.d. random variables $X_1, ..., X_n$ with Gaussian distribution $P_{\boldsymbol{\theta}_0} = \mathcal{N}(\boldsymbol{\theta}_0, 1)$ where $\boldsymbol{\theta}_0$ is unknown. We choose the standard Gaussian $P_0 = \mathcal{N}(0,1)$ as the reference measure, and we denote the associated density $p_{\boldsymbol{\theta}}$ which is equal to*

$$p_{\boldsymbol{\theta}}(x) = \begin{cases} \exp(\boldsymbol{\theta}x - \boldsymbol{\theta}^2/2) & \text{if } x \neq \boldsymbol{\theta} \text{ or } \boldsymbol{\theta} \leq 0, \\ \exp(\boldsymbol{\theta}x - \boldsymbol{\theta}^2/2 + \exp(x^2)\boldsymbol{\theta}^2/2) & \text{otherwise.} \end{cases}$$

Then, whatever the value of $\boldsymbol{\theta}_0$, on a set of probability tending to 1 as n goes to infinity, the MLE is given by $X_{(n)} := \max(X_1, ..., X_n)$ and is consequently not consistent.

Hence, we shall not consider density-based minimum distance estimators anymore in the sequel, and formulate the principle of MDE as follows. We assume that the true distribution P_0 does not belong to the probability model $\{P_{\boldsymbol{\theta}}/\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ which is equipped with a statistical divergence d. We denote the empirical measure $\widehat{P}_n = \sum_{i=1}^n \delta_{\{X_i\}}/n$. Then the MDE $\widehat{\theta}_n$ is defined as the parameter which associated distribution minimizes the probability distance d to the empirical distribution:

$$d(\widehat{P}_n, P_{\widehat{\theta}_n}) = \inf_{\theta \in \Theta} d(\widehat{P}_n, P_\theta)$$

if it exists and is unique. In the situation where such a minimizer does not exist, one can select an ε-approximate solution $\widehat{\theta}_{n,\varepsilon}$ instead of an exact minimizer:

$$d(P_{\widehat{\theta}_{n,\varepsilon}}, \widehat{P}_n) \leq \inf_{\theta \in \Theta} d(P_\theta, \widehat{P}_n) + \varepsilon.$$

Even though MDE was pioneered by Wolfowitz in the early 50s, only a few works followed on until the late 70s, mainly due to computational concerns. In the discussion

of Bickel (1976), Holm suggested MDE as being the most natural method for some robustness problems, and the finding of Beran (1977) that Hellinger-based MDE can yield both full asymptotic efficiency and robustness stands in sharp contrast with the belief that both concepts are conflicted and cannot be achieved simultaneously. This attractive property has raised the interest of the statistical community: Parr and Schucany (1980) empirically showed that L_2-based MDE could lead to robust estimators in some location models, Millar (1981) proved local asymptotic minimaxity of Cramer-von Mises-based MDE, Donoho and Liu (1988a,b) investigated robustness properties of various minimum distance estimators and some advantages of the Cramer-Von Mises- and Hellinger-based ones, while Parr (1981) provided a comprehensive review of the works done until the beginning of the 80s.

Note that the trivial choice of the KL divergence for discrete measures gives the famous maximum likelihood estimator. Unfortunately, it also leads to various problems as already discussed in Section 1.3.1, and to the following (and natural) question: what metric should be used in MDE ? A typical minimum distance estimate is based on the total variation (TV) distance. Yatracos (1985) showed that TV-based MDE is uniformly consistent in TV distance and robust to misspecification in the i.i.d. setting without any assumption on the parameter set, with a convergence rate characterized by the entropy of the space of measures, while Devroye and Lugosi (2001) provided a theoretical analysis of the restriction, called Skeleton estimate, of the minimum TV estimator to the so-called Yatracos sets. Unfortunately, as most MDE procedures, the computation of Yatracos' and skeleton estimates is not feasible in practice. Another robust version using Wasserstein distance was recently studied in Bernton et al. (2017), in which the authors make use of the recent advances in the field of computational optimal transport (Peyré, 2019) and its efficient numerical algorithms to approximate the Wasserstein distance when the exact computation of the minimum distance estimator is intractable, especially in generative models where one can simulate data from the model but not evaluate its density.

More recently, Briol et al. (2019) introduced the Maximum Mean Discrepancy (MMD) as a minimum distance estimator. The MMD distance is associated with a positive definite kernel (Gretton et al., 2012). Given the reproducing kernel Hilbert space (RKHS) and the corresponding distance associated with the given kernel, it is possible to define a one-to-one mapping between the RKHS and the model, provided some mild conditions on the kernel. Such a mapping is called *kernel mean embedding*. The MMD distance between two probability distributions is then simply defined as the distance in the RKHS between corresponding embeddings. The MMD distance has a wide range of applications from kernel Bayesian inference (Song and Gretton, 2011) to approximate Bayesian computation (Park et al., 2016), and includes two-sample (Gretton et al., 2012) and goodness-of-fit testing (Jitkrittum et al., 2017), MMD GANs (Dziugaite et al., 2015; Li et al., 2015) and autoencoders (Zhao et al., 2017). Briol et al. (2019) proved that such estimators are consistent, asymptotically normal and robust to model misspecification, and some trade-off between statistical efficiency and robustness can be achieved through the choice of the kernel. As for the Wasserstein-based MDE, MMD is particularly suited to generative models where efficient computation can be performed using a simple gradient descent algorithm, with numerous applications such as GANs where the usual discriminator can be replaced by a two-sample test based on MMD (Dziugaite et al., 2015).

A significant part of this thesis is based on two papers investigating further the properties of MMD estimation (Chérief-Abdellatif and Alquier, 2019, 2020). In Chapter 8, we show that the MMD estimator is robust to adversarial contamination, and more generally to misspecification. Besides, we go beyond the classical i.i.d framework and study robustness to dependence between observations. We introduce a new dependence coefficient expressed as a covariance in some reproducing kernel Hilbert space and which is very simple to use in practice. We also provide an SGD algorithm for finite-dimensional models along with its theoretical analysis. Moreover, we connect the MMD estimator to minimum L_2-distance estimation when the true distribution has a density relative to the Lebesgue measure, and show that MMD estimation can be seen as a generalization of minimum L_2-distance estimation. Finally, Chapter 9 is devoted to the study of a Bayesian version of the MMD estimator of Briol et al. (2019). The main results of these papers are detailed in Section 2.3.

Chapter 2

Contributions

2.1 Consistency of variational inference

As said in the previous chapter, variational inference has been extensively studied from the computational viewpoint in the recent years, but only little attention has been put in the literature towards theoretical properties of variational approximations until very recently. In the wake of the work of Alquier and Ridgway (2017) which investigates the consistency of variational approximations in general statistical models and the general conditions that ensure such consistency, we tackle the special case of mixture models in Chapter 4 and deep neural networks in Chapter 5. We also investigate in Chapter 6 the consistency of the ELBO criterion that is used for model selection in the variational Bayes community.

2.1.1 Mixtures models

Mixture models are widely used in Bayesian statistics and machine learning, and are applied in many fields as for instance in computer vision (Ayer and Sawhney, 1995), computational biology (Pan et al., 2003), economics (Deb et al., 2011), transport data analysis (Carel and Alquier, 2017), to name but a few. They are also used for modeling population heterogeneity, hence leading to practical clustering methods (Bouveyron and Brunet-Saumard, 2014; McNicholas, 2016), and they have enough flexibility to approximate accurately almost every density (Bacharoglou, 2010; Kruijer et al., 2010). We refer the interested reader to Celeux et al. (2018) for a review on the recent advances on mixtures.

Following the spirit of Alquier and Ridgway (2017) where the theory is based on the tempered posterior, i.e., the likelihood function is raised to a certain power $\alpha \in (0, 1)$, we derive in Chapter 4 convergence rates for variational approximate posterior distributions, for both the well-specified and misspecified cases. As examples, we consider multinomial and Gaussian mixture distributions. Throughout the chapter, we also provide practical VB algorithms for approximate posterior computations and we also include a short simulation study to illustrate our theoretical results.

A mixture model of K components is composed of distributions $\sum_{j=1}^{K} \omega_j P_{\boldsymbol{\theta}_j}$ where the weight vector $(\omega_1, \cdots, \omega_K)$ belongs to the $(K-1)$-dimensional simplex $\mathcal{S}_K$ and each component $\boldsymbol{\theta}_j$ belongs to some parameter space $\boldsymbol{\Theta}$. Hence, the parameter of a mixture is a vector $\boldsymbol{\theta} = (\omega_1, ..., \omega_K, \boldsymbol{\theta}_1, ...\boldsymbol{\theta}_K)$ of size $2K$. To estimate a mixture $P_{\boldsymbol{\theta}_0} = \sum_{j=1}^{K} \omega_{0,j} P_{\boldsymbol{\theta}_{0,j}}$ using a Bayesian approach, one needs to define a prior $\Pi_0(\boldsymbol{\theta}) = \Pi_{0,\omega}(\omega) \prod_{j=1}^{K} \Pi_{0,j}(\boldsymbol{\theta}_j)$ where $\Pi_{0,\omega} \in \mathcal{M}_1^+(\mathcal{S}_K)$ is a probability distribution over the simplex $\mathcal{S}_K$ and each $\Pi_{0,j} \in \mathcal{M}_1^+(\boldsymbol{\Theta})$ is a probability distribution over the parameter set $\boldsymbol{\Theta}$ (both equipped with a suited σ-algebra).

Typically, posterior distributions for mixtures are difficult-to-compute and we must resort to approximate techniques such as variational inference. The mean-field approximation is particularly suited to mixtures as the space of parameters of the mixtures can be decomposed as $\mathcal{S}_K \times \boldsymbol{\Theta} \times \cdots \times \boldsymbol{\Theta}$. The variational family requires the parameters of the different components to be independent of each other, and also independent of the weights:

$$\mathcal{Q} = \left\{ Q(\boldsymbol{\theta}) = Q_\omega(\omega) \prod_{j=1}^{K} Q_j(\boldsymbol{\theta}_j) / Q_\omega \in \mathcal{M}_1^+(\mathcal{S}_K) \ , \ Q_j \in \mathcal{M}_1^+(\boldsymbol{\Theta}) \ \forall j = 1, ..., K \right\}.$$

The main result can be summarized as follows:

Theorem 2.1.1 (Informal)**.** *Assume that the extended prior mass condition (1.9) is satisfied for each component, i.e. that for each j there exists a distribution $Q_{n,j} \in \mathcal{Q}$ such that:*

$$\int \mathrm{KL}(P_{\boldsymbol{\theta}_{0,j}} \| P_{\boldsymbol{\theta}_j}) Q_{n,j}(d\boldsymbol{\theta}) \leq r_n \ \text{ and } \ \mathrm{KL}(Q_{n,j} \| \Pi_{0,j}) \leq n r_n.$$

Then for some Dirichlet prior on the weights and any $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha \left(\sum_{j=1}^{K} \omega_j P_{\boldsymbol{\theta}_j}, \sum_{j=1}^{K} \omega_{0,j} P_{\boldsymbol{\theta}_{0,j}} \right) \widetilde{\Pi}_{n,\alpha}(d\boldsymbol{\theta}) \right] \leq \frac{1+\alpha}{1-\alpha} 2K r_n.$$

Mainly, this theorem states that when estimation of a distribution in a model is possible at rate r_n, then it is possible to estimate a mixture of K distributions in the model with a rate of convergence equal to Kr_n. Interestingly, there is no prior mass assumption on the weights. The reason is that the prior mass assumption is satisfied for the rate $\log(nK)/n$ when choosing a Dirichlet prior $\Pi_\omega = \mathcal{D}_K(\beta_1, ..., \beta_K)$ with some minor restriction on the parameters $\beta_1, ..., \beta_K$. As in practice, the rate of convergence of the variational approximation associated with a model is often slower than $\log(nK)/n$, then the rate of convergence of the approximate posterior associated with the mixture is entirely driven by the prior mass condition on the different components. Note that this result is remarkable as there are almost no assumptions made on the mixture model.

As an application, we tackle the case of Gaussian mixtures and we show that a single Gaussian distribution can be estimated at a rate $r_n = \log(n)/n$. In this case, the rate of convergence for a Gaussian r_n is faster than the one for the weight $\log(nK)/n$, and hence the final convergence rate $K \log(nK)/n$ comes from the prior mass on the weights. This is valid when the variance of the Gaussian is known but also when we estimate both the mean and the variance.

The result stating that the extended prior mass condition is satisfied for Dirichlet priors is of interest on its own. Indeed, when dealing with mixture of V multinomials, this result is sufficient to obtain the final rate of convergence $KV\log(nV)/n$ of the tempered posterior. As already explained in Section 1.1.3, the prior mass condition is exactly $\Pi_\omega(\mathcal{B}) \geq e^{-nKr_n}$ in this case, and the proof of the extended prior mass condition for the weight vector goes back to the computation of the prior mass $\Pi_\omega(\mathcal{B})$ where $\mathcal{B}$ is the KL-ball centered of radius Kr_n at ω_0. When using Dirichlet priors, Lemma 6.1 in Ghosal et al. (2000) addresses the computation of such a prior mass for L_1-balls, that we extend here to KL-balls.

We also provide a numerical algorithm to compute the variational approximation. A popular technique used for mean-field approximations consists in optimizing iteratively in all the independent components. Nevertheless, this is actually as difficult as maximizing the log-likelihood of a mixture, which is not feasible in practice. To overcome this intractability, we incorporate variables ω_j^i's in the optimization program without modifying it thanks to Lemma 1.1.1. Such variables ω_j^i's can be interpreted as posterior means of latent variables Z_j^i's as for the EM algorithm (Dempster et al., 1977). Then, the equivalent program can be solved efficiently using coordinate descent, see Algorithm 6. Note that when $\alpha = 1$, our algorithm is exactly equivalent to the popular coordinate ascent variational inference (CAVI) algorithm which is widely used in the VB community (Blei et al., 2017). We finally compare our algorithm with both CAVI and EM, which achieves comparable performance for the estimation of Gaussian mixtures.

2.1.2 Deep neural networks

In the last decade, deep learning (DL) has made major breakthroughs among practitioners (LeCun et al., 2015; Goodfellow et al., 2016). Unfortunately, generalization properties of DL are not well understood, and a recent line of research investigates the properties of deep networks from a theoretical perspective. In particular, some recent works addressed the estimation of smooth functions in a nonparametric regression framework, using either frequentist or Bayesian tools (Schmidt-Hieber, 2017; Suzuki, 2018, 2019; Rockova and Polson, 2018). In Chapter 5, we provide the first theoretical analysis of (tempered) variational inference for Bayesian deep learning in nonparametric regression. We show that when choosing a relevant variational family, then the variational approximation retains the same properties than the posterior it approximates. We give the convergence rate in a general framework with any regression function, and we show that in particular, we recover the minimax optimal convergence rate (up to log-terms) for the case of Hölder smooth functions for some network architecture.

We consider a nonparametric regression framework with a collection of i.i.d. random variables $(X_i, Y_i) \in [-1,1]^d \times \mathbb{R}$ for $i = 1, ..., n$:

$$\begin{cases} X_i \sim \mathcal{U}([-1,1]^d), \\ Y_i = f_0(X_i) + \zeta_i \end{cases}$$

where $\mathcal{U}([-1,1]^d)$ is the uniform distribution on the interval $[-1,1]^d$, $\zeta_1, ..., \zeta_n$ are i.i.d. Gaussian random variables with mean 0 and known variance σ^2, and $f_0 : [-1,1]^d \to \mathbb{R}$

is the true unknown function. We estimate f_0 using sparse deep neural networks $f_{\boldsymbol{\theta}}$ of sparsity S, depth L and width D.

Again, we adopt a Bayesian approach. We choose a sparsity inducing spike-and-slab prior, for which a number S of selected neurons of the network follow a standard Gaussian distribution, and all the others neurons are put to 0 with probability 1. The variational set is also composed of Gaussian spike-and-slab distributions.

The main contribution of the paper is a nonasymptotic generalization error bound for variational inference in sparse DL in the nonparametric regression framework:

Theorem 2.1.2 (Informal). *For any $\alpha \in (0,1)$, for any positive number B,*

$$\mathbb{E}\left[\int \|f_{\boldsymbol{\theta}} - f_0\|_2^2 \widetilde{\Pi}_{n,\alpha}(d\boldsymbol{\theta})\right] \leq \frac{2}{1-\alpha} \inf_{\|\boldsymbol{\theta}^*\|_\infty \leq B} \|f_{\boldsymbol{\theta}^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n,$$

where the rate of convergence r_n is of order $\frac{LS}{n}\log(BD)$.

The rate of convergence in the right-hand-side of the oracle inequality, which depends linearly in the number of layers and the sparsity, is determined by the previous extended prior mass condition, and recovers exactly the rate of convergence of the empirical risk minimizer for DNNs which is obtained using different proof techniques, by computing the local covering entropy i.e. the logarithm of the number of r_n-balls needed to cover a neighborhood of the true regression function (Schmidt-Hieber, 2017; Suzuki, 2019).

In particular, when the true regression function is Hölder smooth, assuming that the sparsity, depth, and width of the network are appropriately chosen as suggested by Rockova and Polson (2018), this implies that the variational approximation concentrates towards f_0 and enjoys the same near-minimax rates of convergence than those achieved by the exact tempered and regular posteriors. The result is established by deriving the previous PAC-Bayes oracle inequality for a Hölder regression function and applying the approximation result of Schmidt-Hieber (2017).

We further consider the extension of the oracle inequality when the optimization algorithm incurs error as measured by its effect on the ELBO. More precisely, when we consider an algorithm $(\widetilde{\Pi}_{n,\alpha}^{(j)})_j$ for computing the ideal approximation $\widetilde{\Pi}_{n,\alpha}$, there is an additional term in the generalization error:

Theorem 2.1.3 (Informal). *For any $\alpha \in (0,1)$, for any positive number B and any number of iterations j,*

$$\mathbb{E}\left[\int \|f_{\boldsymbol{\theta}} - f_0\|_2^2 \widetilde{\Pi}_{n,\alpha}^{(j)}(d\boldsymbol{\theta})\right] \leq \frac{2}{1-\alpha} \cdot \inf_{\|\boldsymbol{\theta}^*\|_\infty \leq B} \|f_{\boldsymbol{\theta}^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n + \frac{\mathbb{E}[\Delta_{n,j}]}{n},$$

where $\Delta_{n,j}$ is the difference between the maximum value of the ELBO and the value of the ELBO at the j^{th} iteration of the algorithm.

Hence, the algorithm $\widetilde{\Pi}_{n,\alpha}^{(j)}$ is consistent at the same rate r_n than the ideal variational approximation it computes $\widetilde{\Pi}_{n,\alpha}$ as soon as $\mathbb{E}[\Delta_{n,j}] \lesssim LS\log(BD)$.

2.1.3 Model selection

The question that naturally arises when dealing with mixtures and deep neural networks is respectively the question of the selection of the number of components and the question of the selection of the architecture of the network. For instance, the choice of the architecture in deep learning is crucial and can lead to faster convergence and better approximation. Similarly, in Bayesian mixture modeling and in situations where the number of mixture components is unknown, full hierarchical Bayes that entails putting a prior on the number of components is often used in practice. This requires Reversible Jump Markov Chain Monte Carlo algorithms to do computation, but designing such algorithms is a delicate matter in many practical situations and they do not scale well with data size. Hence, any fast VB solution to this problem is of great interest.

More generally, this raises the question of model selection, and the design of an adaptive model selection criterion that selects an optimal model. The ELBO maximization criterion is widely used in the Variational Bayes community and is known to work well in practice. More precisely, we consider several models indexed by K and associated variational approximations $\widetilde{\Pi}^K_{n,\alpha}$, and we assign a prior weight π_K to each model. The ELBO maximization criterion consists in choosing the model that provides the closest approximation to the log-evidence $\mathrm{ELBO}(K)$, i.e. the maximum value of the ELBO associated with model K. We propose in Chérief-Abdellatif (2019a) a penalized version of the ELBO maximization criterion that is equivalent to the usual one when considering a finite number of models and uniform prior weights:

$$\widehat{K} = \arg\max_K \left\{ \mathrm{ELBO}(K) - \log\left(\frac{1}{\pi_K}\right) \right\}.$$

The ELBO maximization criterion has never been justified in theory. In Chapter 6, we show that this criterion is adaptive and selects a variational approximation that achieves the optimal convergence rate among the competing models. In particular, we show respectively in Chapter 4 and Chapter 5 that the ELBO criterion selects a number of components and a network architecture that give the optimal rate and does not lead to overfitting. We also show in Chapter 6 that the minimax convergence rate is achieved for probabilistic principal component analysis.

We know from Alquier and Ridgway (2017) that as soon as there exists a true model which satisfies the extended prior mass condition, then the variational approximation corresponding to the true model is consistent at the rate defined by the prior mass condition. The main message of Chérief-Abdellatif (2019a) is that even if we do not know which model is true, the ELBO criterion provides a model (not necessarily the true one) such that the corresponding approximation is consistent and adaptively achieves the rate of convergence associated with the true model. More precisely:

Theorem 2.1.4 (Informal)**.** *Assume that Assumption 1.9 is satisfied for the true model K_0. Then for any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{\theta}}, P^0)\widetilde{\Pi}^{\widehat{K}}_{n,\alpha}(d\boldsymbol{\theta})\right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

Actually, the overall rate is composed of the convergence rate r_n corresponding to the true model and of a complexity term which reflects the prior belief over the different models. For instance, if we range a countable number of models according to our prior belief, then the complexity term will be of order K_0/n when choosing $\pi_K = 2^{-K}$. In practice, the overall term is of order r_n.

The application of this result to both mixture models and deep neural networks can be found in Chapters 4 and 5. We also consider in Chapter 6 probabilistic Principal Component Analysis (PCA). In the next paragraphs, matrices will be denoted in bold capital letters. We assume the model

$$X_i = \boldsymbol{W}_0 Z_i + \varepsilon_i$$

with independent and identically distributed Gaussian random variables $Z_i \sim \mathcal{N}(0, \boldsymbol{I}_{K_0})$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ and $\boldsymbol{I}_{K_0}$ are respectively the d- and K_0-dimensional identity matrices ($K_0 < d$), $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K}$ is the K_0-rank matrix that contains the principal axes and σ^2 is a noisy term that is known. K_0 corresponds here to the "true dimensionality" of the data. It is unknown here and we consider several models corresponding to several values of $K = 1, ..., d$ and corresponding matrices $\boldsymbol{W} \in \mathbb{R}^{d \times K}$.

We put equal prior weights π_K over each integer $K = 1, ..., d$. Given rank K, we place a prior over the K-rank matrix $\boldsymbol{W}$ to infer a distribution over principal axes. We choose independent Gaussian priors on the columns of $\boldsymbol{W}$. We also consider independent Gaussian variational approximations on the columns of $\boldsymbol{W}$. Using a clipping operator, it is possible to obtain the consistency in Frobenius norm $\|\cdot\|_F$ of the selected variational approximation to the true covariance matrix under the classical assumption that the spectral norm $\|\cdot\|_2$ of the true matrix $\boldsymbol{W}_0$ is bounded. The symbol $\mathcal{O}$ in the following result hides universal constants that are independent of α, d, K_0 and n:

Theorem 2.1.5 (Informal). *For any $\alpha \in (0,1)$, if $\|\boldsymbol{W}_0\|_2 \leq B$, then:*

$$\mathbb{E}\left[\int \left\| clip_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0\boldsymbol{W}_0^T \right\|_F^2 \widetilde{\Pi}_{n,\alpha}^{\widehat{K}}(d\boldsymbol{W})\right] = \mathcal{O}\left(\frac{1+\alpha}{\alpha(1-\alpha)} \cdot \frac{dK_0 \log(dn)}{n}\right),$$

where $clip_B(\boldsymbol{A})$ is the matrix which (i,j)-entry is equal to $\begin{cases} \boldsymbol{A}_{i,j} & \text{if } |\boldsymbol{A}_{i,j}| \leq B^2 \\ B^2 & \text{if } \boldsymbol{A}_{i,j} \geq B^2 \\ -B^2 & \text{otherwise.} \end{cases}$

Each of the following chapters is self-contained and all of them can be read independently. A brief summary of their contributions is as follows:

- ***Chapter 5 studies the concentration of variational approximations of posteriors for general mixtures, and we derive consistency and rates of convergence. We also tackle the problem of selecting the number of components by maximizing the ELBO. The work in this chapter has been published in Chérief-Abdellatif and Alquier (2018).***

- ***Chapter 6 provides nonasymptotic generalization bounds ensuring the consistency of Bayesian DNNs when an approximation is used instead***

of the exact posterior distribution, along with rates of convergence. We show that it leads to near-minimax estimation of smooth functions for a wise choice of the architecture. We also use the ELBO criterion for selecting the architecture. The work in this chapter has been accepted at ICML 2020 (Chérief-Abdellatif, 2019b).

- ***Chapter 7 justifies the use of the ELBO maximization strategy from a theoretical perspective. We illustrate our theoretical results by an application to the selection of the number of principal components in probabilistic PCA. The work in this chapter has been published in Chérief-Abdellatif (2019a).***

2.2 Online variational inference

Following the previous analyses in the batch setting, we investigate in Chapter 7 variational inference from an online perspective. No universal definition of variational Bayes has been given in online learning. As already detailed in Section 1.2, keeping the standard definition of the variational approximation would be a simplistic solution that would remove the online aspect of the problem and lead to computational issues. In Chapter 7, we propose the first theoretical analysis of VI in the online learning framework with streaming data. We study several classes of online algorithms that are inspired from popular strategies used in online sequential optimization such as gradient descent and follow-the-regularized-leader, and we provide generalization bounds in the convex case. Our proof techniques are based on the convexity of the loss function, but we argue that our bounds should hold more generally.

2.2.1 Variational approximations of EWA

In the literature, mainly two kinds of extensions of VB to the online setting have been formulated. The first definition of online VB extends the definition of Bayes used in batch statistics through Donsker-Varadhan Lemma (see Formula 1.5), and restricts the set of minimization to a family of tractable distributions. When considering parametric approximations $\mathcal{Q} = \{Q_\mu / \mu \in \mathcal{M}\}$, we define variational approximations as $\widetilde{\Pi}_{t,\alpha} = Q_{\mu_t}$ using a sequence of parameters $(\mu_t)_t$ with $Q_{\mu_0} = \Pi_0$:

$$\mu_t = \arg\min_{\mu \in \mathcal{M}} \left\{ \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\theta} \sim Q_\mu} \left[\ell_s(\boldsymbol{\theta}) \right] + \frac{\mathrm{KL}(Q \| Q_{\mu_0})}{\alpha} \right\}.$$

The main drawback of this method is that it loses the online aspect and requires the entire set of data samples to be loaded in memory at each step, which is computationally really expensive. This is though the approach adopted in Guhaniyogi et al. (2013) where the authors propose two methods: the first one is a generalization of Equation 1.5, by minimizing over a nonparametric convex set of distributions $\mathcal{Q}$ via a functional gradient descent. The problem is that except for the trivial case where $\mathcal{Q} = \mathcal{M}_1^+(\boldsymbol{\Theta})$, they do not provide any example of such sets in practice. They also use a parametric family

$\mathcal{Q} = \{Q_\mu, \mu \in M\}$. However, they change the order in KL, which involves an intractable term that they compute using Monte-Carlo. Moreover, the regret they upper bound is not a natural quantity and does not provide any guarantees on the initial objective, that is minimizing the ℓ_t's.

We notice that our variational approximation is formulated as a follow-the-regularized-leader strategy applied to the expected loss $\mu \to \mathbb{E}_{\boldsymbol{\theta}\sim Q_\mu}[\ell_s(\boldsymbol{\theta})]$. Hence, we propose to linearize the (expected) loss function as done in Section 1.2:

$$\mu_t = \arg\min_{\mu\in\mathcal{M}} \left\{ \mu^T \sum_{s=1}^{t} \nabla_\mu \mathbb{E}_{\boldsymbol{\theta}\sim Q_{\mu_s}} \Big[\ell_s(\boldsymbol{\theta})\Big] + \frac{\mathrm{KL}(Q\|Q_{\mu_0})}{\alpha} \right\}.$$

We shall denote this algorithm Sequential Variational Approximation (SVA) in the rest of this manuscript, see Algorithm 10.

As for the gradient descent algorithm, there exists another formulation of this optimization program. Rather than considering the cumulative loss and regularizing (using KL) with respect to the first approximation, it is also possible to consider only the last loss and to regularize with respect to the last approximation. Nevertheless, contrary to gradient descent, the second formulation is not equivalent to the first one and leads to a different optimization program:

$$\mu_t = \arg\min_{\mu\in\mathcal{M}} \left\{ \mu^T \nabla_\mu \mathbb{E}_{\boldsymbol{\theta}\sim Q_{\mu_t}} \Big[\ell_s(\boldsymbol{\theta})\Big] + \frac{\mathrm{KL}(Q\|Q_{\mu_{t-1}})}{\alpha} \right\}.$$

Actually, this algorithm is in line with the other approach which extends the sequential definition of EWA given in Algorithm 5: $\Pi_{t+1,\alpha}(d\boldsymbol{\theta}) \propto \exp(-\alpha\ell_t(\boldsymbol{\theta}))\Pi_{t,\alpha}(d\boldsymbol{\theta})$. This point of view is particularly appealing as it keeps the online aspect of the problem and there is no need to store any data. The update is obtained through a formula of the form $\widetilde{\Pi}_{t,\alpha} = F(\widetilde{\Pi}_{t-1,\alpha}, \ell_t)$. For instance, this is the point of view adopted in Broderick et al. (2013), which was further explored in Nguyen et al. (2017a) and Zeno et al. (2018) where authors mixed the batch definition of VB and the online update approach to obtain the following definition in two steps:

- An approximation step: $\bar{\Pi}_{t,\alpha}(d\boldsymbol{\theta}) \propto \exp(-\alpha\ell_t(\boldsymbol{\theta}))Q_{\mu_{t-1}}(d\boldsymbol{\theta})$.
- A projection step: $\mu_t = \arg\min_{\mu\in\mathcal{M}} \mathrm{KL}(Q\|\bar{\Pi}_{t,\alpha})$.

Plugging the approximation step into the projection step leads to:

$$\mu_t = \arg\min_{Q\in\mathcal{Q}} \left\{ \mathbb{E}_{\boldsymbol{\theta}\sim Q} \Big[\ell_t(\boldsymbol{\theta})\Big] + \frac{\mathrm{KL}(Q\|Q_{\mu_{t-1}})}{\alpha} \right\}.$$

Hence, our algorithm is a linearized version of the previous approximation/projection variational approximation. We will denote this algorithm Streaming Variational Bayes (SVB) in the sequel, see Algorithm 10. As for the approach using Donsker-Varadhan Lemma (1.5), no rigorous analysis of the generalization properties of the approximations obtained using this approach has been provided and it is not clear whether the computation of $\widetilde{\Pi}_{t,\alpha}$ is feasible.

Moreover, Algorithms SVA and SVB are both tractable and have closed-form updates for variational sets used in practice. For instance, when using the Gaussian mean-field class of all Gaussian approximations $Q_{(m,\sigma)} = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ with mean m and diagonal covariance matrix which diagonal is the vector $\sigma^2 = (\sigma_1^2, \dots, \sigma_d^2)^T$, they lead to the following gradient descent like updates:

$$\text{SVA:} \begin{cases} m_{t+1} & \leftarrow m_t - \alpha\sigma_0^2 \bar{g}_{m_t}, \\ g_{t+1} & \leftarrow g_t + \bar{g}_{\sigma_t}, \\ \sigma_{t+1} & \leftarrow h\left(\frac{1}{2}\alpha\sigma_0 g_{t+1}\right)\sigma_0. \end{cases}$$

$$\text{SVB:} \begin{cases} m_{t+1} & \leftarrow m_t - \alpha\sigma_t^2 \bar{g}_{m_t}, \\ \sigma_{t+1} & \leftarrow h\left(\frac{1}{2}\alpha\sigma_t \bar{g}_{\sigma_t}\right)\sigma_t. \end{cases}$$

where $\bar{g}_{m_t}$ and $\bar{g}_{\sigma_t}$ stand respectively for the gradient at $(m, \sigma) = (m_t, \sigma_t)$ of the expected loss $(m, \sigma) \to \mathbb{E}[\ell_t(\boldsymbol{\theta})]$ with respect to m and σ, where $h(x) := \sqrt{1 + x^2} - x$ and all operations are applied componentwise for vector inputs.

2.2.2 Regret bounds

We provide in Chapter 7 regret bounds for SVA. The main required assumptions are the Lipschitzness and the convexity of the expected loss $\mu \to \mathbb{E}_{\boldsymbol{\theta}\sim Q_\mu}[\ell_t(\boldsymbol{\theta})]$, as well as the strong convexity of the KL regularizer with respect to μ.

Theorem 2.2.1 (Informal)**.** *For L-Lipschitz and convex expected losses and a σ-strongly convex KL regularizer, SVA has the following regret bound:*

$$\sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}_t) \widetilde{\Pi}_{t,\alpha}(d\boldsymbol{\theta}_t) \leq \inf_{\mu \in \mathcal{M}} \left\{ \sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}) Q_\mu(d\boldsymbol{\theta}) + \frac{\alpha L^2 T}{\sigma} + \frac{\mathrm{KL}(Q_\mu, \Pi_0)}{\alpha} \right\}.$$

As desired, this result is almost the same than the one obtained in Theorem 1.2.5 where the infimum is restricted to the variational family, the upper bound B is replaced by the Lipschitz constant L, and the factor 8 by the strong convexity parameter σ. Nevertheless, the proof is very different and relies on arguments coming from online convex optimization (Shalev-Shwartz, 2012; Hazan, 2016). Note that as in Section 1.2, online-to-batch techniques can lead to generalization error bounds under a convexity assumption on the loss function ℓ_t of order:

$$\mathbb{E}_{\mathcal{D}_T \sim P_0}[R(\bar{\boldsymbol{\theta}}_T)] \leq \inf_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) + \mathcal{O}\left(\frac{L}{\sigma}\sqrt{\frac{d\log(dT)}{T}}\right).$$

We also provide in Chapter 7 a regret bound for SVB for the Gaussian mean-field family $Q_{(m,\sigma)} = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ with mean/standard deviation parameterization with a dynamic and multidimensional learning rate $\alpha_t = (\alpha_{t,1}, \cdots, \alpha_{t,j})$. The regret bound requires a bounded parameter set $\mathcal{M} = \mathcal{M}_m \times \mathcal{M}_\sigma$ with an additive projection step where $\mathcal{M}_m$ and $\mathcal{M}_\sigma$ are closed, bounded, convex subsets of $\mathbb{R}^d$ and $\mathbb{R}^d_{\geq 0}$ respectively, with $0 \in \mathcal{M}_\sigma$. We also define the diameter $D^2 = \sup\{\|m - m'\|_2^2 + \|\sigma\|_2^2, m, m' \in \mathcal{M}_m, \sigma \in \mathcal{M}_\sigma\}$. Note that the strong convexity of the regularizer is always satisfied in this case.

Theorem 2.2.2 (Informal). *For convex losses ℓ_t and L-Lipschitz and convex expected losses, and for $\widehat{\boldsymbol{\theta}}_t \sim Q_{(m_t,\sigma_t)}$, we have for some choice of the learning rate:*

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{\theta}}_t) \leq \inf_{\theta \in \mathcal{M}_m} \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) + DL\sqrt{2T}.$$

Moreover, if the expected losses are H-strongly convex, we even have a logarithmic rate:

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{\theta}}_t) \leq \inf_{\theta \in \mathcal{M}_m} \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) + \frac{L^2(1+\log T)}{H}.$$

Once again, the results are similar to the regular Bayesian case but are directly expressed in terms of the parameters $\boldsymbol{\theta}$ instead of expectations. It is even possible, optimizing over $\mathcal{M}_m = \{m \in \mathbb{R}^d : \|m\|_2 \leq \bar{M}\}$ and $\mathcal{M}_\sigma = \{\sigma \in \mathbb{R}^d_+ : \|\sigma\|_2 \leq \bar{S}\}$, to get dimension-free generalization error bounds of order $L(4\bar{M}^2 + \bar{S}^2)^{1/2}T^{-1/2}$.

2.2.3 Going beyond convexity

Our proofs are based on a convexity argument. However, we expect our bounds to hold more generally, even for some nonconvex expected losses. This is particularly appealing as it would help us obtain generalization guarantees for the natural-gradient variational inference (NGVI) algorithm (Sato, 2001; Hoffman et al., 2013; Khan and Lin, 2017) using an exponential variational family.

This algorithm is widely used in stochastic learning but can be easily modified for the online setting. The method presented in Khan and Lin (2017) exploits the duality between the expectation parameter μ and the natural parameter λ of the exponential family, and can be written as follows:

$$\lambda_t = \lambda_{t-1} - \alpha \nabla_\mu \mathbb{E}_{\boldsymbol{\theta} \sim Q_{\mu_{t-1}}}[\ell_t(\boldsymbol{\theta})],$$

with a possible projection step.

In particular, this natural gradient algorithm can be simply written as SVB applied to an exponential variational family with an expectation parameterization. For instance, when using a Gaussian mean-field family, the expectation parameter is simply composed of the first two order moments, i.e. the mean and the correlation matrix. It is known to perform very well in practice, and it is shown in Chérief-Abdellatif et al. (2019) that it outperforms both SVA and SVB applied to the usual mean/standard deviation matrix parameterization (m, σ) in several settings and with several losses. Unfortunately, the expected loss is not convex with respect to the expectation parameter for most losses while being convex with respect to (m, σ), and thus we are not able to analyze the performance of NGVI.

One of the main empirical findings and suggestions of Chapter 7 is that the generalization properties of online VI seem to go beyond the convex assumption required to obtain the theoretical results, and that convexity of the expected loss is not the cornerstone of the generalization of online VI.

In Chapter 7, we derive the first generalization bounds for online variational inference. By using existing methods, we propose some online methods for variational inference, namely the SVA and SVB algorithms, and we provide generalization bounds for both of them. We support our theoretical findings with numerical experiments on simulated and real data. We observe that NGVI outperforms all the other methods, we conjecture that the theoretical convexity assumption on the expected loss can be relaxed in practice, and we believe that our theoretical analysis can be extended to the NGVI algorithm. The work in this chapter has been published in Chérief-Abdellatif et al. (2019).

2.3 Robustness via Maximum Mean Discrepancy

As already said in Section 1.3, most attempts to design a universal estimator which is simultaneously consistent and statistically optimal when the generating distribution of the data belongs to the model, and robust to small departures from the model assumptions are not computationally feasible. The recent minimum distance estimator introduced by Briol et al. (2019) and based on the Maximum Mean Discrepancy (MMD) is consistent at optimal rates in MMD distance when the model is well-specified, robust to misspecification, and can be easily computed using stochastic gradient descent. In Chapter 8, we show that this estimator is robust to both dependency and to the presence of outliers in the dataset. We also relate this MMD-based estimator to L_2-estimation, and we propose a theoretical analysis of the gradient algorithm used to compute the estimator. We present empirical evidence in support of this. We also propose a Bayesian version in Chapter 9, where we study its concentration properties under some prior mass condition and provide an explicit algorithm for computing the MMD-based pseudo-Bayes posterior using variational inference.

2.3.1 The MMD estimator

In the sequel, we consider a positive definite kernel function k, i.e a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for any integer $n \geq 1$, for any $x_1, ..., x_n \in \mathcal{X}$ and for any $c_1, ..., c_n \in \mathbb{R}$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0,$$

and the corresponding reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ which satisfies the reproducing property $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$ for any function $f \in \mathcal{H}_k$ and any $x \in \mathcal{X}$. We assume that the kernel is bounded by some positive constant, that will be assumed to be 1 without loss of generality.

The *kernel mean embedding* of a probability measure P is the function $\mu_P \in \mathcal{H}_k$ such that:

$$\mu_P(\cdot) := \mathbb{E}_{X \sim P}[k(X, \cdot)] \in \mathcal{H}_k.$$

We also assume that the kernel is *characteristic*, which means that the mapping $P \mapsto \mu_P$ is injective. Hence we can define the distance:

$$\mathbb{D}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$$

which is the maximum mean discrepancy between P and Q.

Given a parameter set $\boldsymbol{\Theta}$, we define the MMD estimator $\widehat{\boldsymbol{\theta}}_n$ as in Briol et al. (2019):

$$\mathbb{D}_k(P_{\widehat{\boldsymbol{\theta}}_n}, \widehat{P}_n) = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{D}_k(P_{\boldsymbol{\theta}}, \widehat{P}_n).$$

It is also possible to consider an approximate minimizer when the exact minimum does not exist.

2.3.2 Robustness to misspecification and dependence

Chapter 8 investigates the universality properties of minimum distance estimation based on the maximum mean discrepancy, particularly regarding dependence to misspecification and dependence.

A generalization bound:

The first result is an oracle inequality that still holds in the dependent setting ensuring an MMD decrease of the generalization error in $n^{-1/2}$ as $n \to +\infty$, which is known to be optimal (Tolstikhin et al., 2017):

Theorem 2.3.1 (Informal)**.** *Under relevant weak dependence assumptions involving two positive constants* Σ *and* Γ*, we have for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*,*

$$\mathbb{D}_k\left(P_{\widehat{\boldsymbol{\theta}}_n}, P_0\right) \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{D}_k\left(P_{\boldsymbol{\theta}}, P_0\right) + 2\frac{\sqrt{1 + 2\Sigma} + (1 + \Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}.$$

In particular, under the i.i.d. assumption, we have $\Sigma = \Gamma = 0$*.*

The result above is based on two main dependence assumptions. The first one is sufficient to get a result in expectation. We introduce in Chérief-Abdellatif and Alquier (2019) a new weak dependence coefficient in the wake of the works of Doukhan and Louhichi (1999) which is expressed as a covariance in the RKHS associated with the kernel. Such a coefficient is very simple to use in practice, much easier to compute than mixing coefficients and the summability of the series of these coefficients stands for a wider class of processes. When the sum of the series exists, we say that the stochastic process is *weakly dependent*. Σ denotes the sum of the series in the previous theorem. We investigate several examples of processes that are not mixing but still weakly dependent in Chapter 8.

The second assumption is essential to obtain a result in probability. Indeed, a concentration inequality upper bounding the MMD distance between the empirical and the

true distribution is required to get such an oracle inequality. For instance, the result of Briol et al. (2019) is based on McDiarmid's inequality (McDiarmid, 1989), but this Hoeffding-type inequality is only valid in the independent setting. Hence, we exploit here a version of McDiarmid's inequality designed for time series which is available under a polynomial decay assumption on some mixing dependence coefficients $(\gamma_{i,j})_{1\leq i<j}$ (Rio, 2013). Again, we adapt here the decay assumption to the RKHS $\mathcal{H}_k$, and the constant Γ in the previous theorem denotes the sum of the series of such mixing coefficients.

Bounds in L_2-distance:

We also connect our bounds in MMD to other metrics, in particular to the L_2-distance. Some previous attempts to design estimators that are robust to misspecification give bounds in TV or in Hellinger distances (Baraud and Birgé, 2016; Devroye and Lugosi, 2001), and prevent people from using that the quadratic loss as a minimum distance estimator. The first main reason is that the L_2-metric is not universal. Indeed, all probability measures do not necessarily have a density with respect to some reference measure (e.g. Lebesgue measure), and some of them may have one but that is not L_2-integrable. Moreover, such a density highly depends on the choice of the reference measure as explained in Section 1.3.

We argue in Chapter 8 that for kernels of the form $k(x,y) = F(\|x-y\|/\gamma)$ for a function $F : [0,+\infty) \to [0,1]$, the maximum mean discrepancy can be expressed as an approximation of the quadratic distance that is well-defined for any probability distribution, and that it is possible to derive oracle inequalities in L_2-distance when densities exist and are L_2-integrable. For instance, for the Gaussian kernel $k_\gamma(x,y) = \exp(-\|x-y\|^2/\gamma^2)$, $\mathbb{D}_{k_\gamma}(P,Q) \sim \pi^{\frac{d}{4}}\gamma^{\frac{d}{2}}\|p-q\|_{L_2}$ when $\gamma \to 0$ and where p and q are densities of P and Q respectively with respect to the Lebesgue measure. Hence, for γ small enough, we give a sense to L_2 estimation even for densities that are not L_2-integrable. Furthermore, the maximum mean discrepancy does not depend on any reference measure, though it depends on the choice of the kernel. This dependence in k is actually an attractive feature as it gives flexibility to take into account the underlying geometry of $\mathcal{X}$ via the choice of a distance on this space, a property that explains the popularity of the Wasserstein distance in statistics. For example, $\mathbb{D}_k(\delta_x,\delta_y) \to 0$ when $x \to y$, a property that is shared with the Wasserstein distance but that does not hold for the Hellinger nor the TV distance.

Robustness to adversarial contamination:

MMD-based estimation can also be used in the special case of robust parametric estimation with adversarial contamination, where the target distribution $P_{\boldsymbol{\theta}_0}$ belongs to the model but a fraction ε of the data are contaminated by an adversary. In this setting:

Theorem 2.3.2 (Informal)**.** *Under the same assumptions than for Theorem 2.3.1, we have for any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\mathbb{D}_k\left(P_{\widehat{\boldsymbol{\theta}}_n}, P_{\boldsymbol{\theta}_0}\right) \leq 4\left(\varepsilon + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right).$$

Again, it is possible to obtain a weaker result in expectation by relaxing the assumption involving Γ. Note that this result still holds for Hüber's contamination setting but the constant 4 is replaced by the smaller one 2.

The rate achieved by the MMD estimator (in MMD) is $\max(1/\sqrt{n}, \varepsilon)$, and has the same form than the minimax rate when estimating the mean of a Gaussian distribution. We recover the optimal rate of convergence with respect to n without contamination when $\varepsilon \lesssim 1/\sqrt{n}$, while the rate is dominated by the contamination ratio ε otherwise. Thus, the maximum number of outliers that is tolerated without breaking down the optimal rate is $\sqrt{n}$, and is independent of the dimension. In the particular setting of the estimation of a Gaussian mean with covariance matrix $\sigma^2 I_d$, we achieve via the Gaussian kernel the same rate than the coordinatewise median, i.e. $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = \mathcal{O}(\max(d^{1/2}n^{-1/2}, d^{1/2}\varepsilon))$, while the minimax optimal rate is $\max(d^{1/2}n^{-1/2}, \varepsilon)$. As a comparison, the robust Median-of-Means methodology leads to estimation in $\mathcal{O}(\max(d^{1/2}n^{-1/2}, \varepsilon^{1/2}))$, i.e. a maximum number of outliers that is tolerated without breaking down the optimal rate of order d. We believe that our rates obtained using the MMD estimator can be improved by a proper choice of the kernel.

Computational issues:

The estimator $\widehat{\boldsymbol{\theta}}_n$ can be computed using a gradient-based algorithm when $\boldsymbol{\Theta} \subset \mathbb{R}^d$ for a generative model, i.e. when it is possible to sample from any $P_{\boldsymbol{\theta}}$. The idea of exploiting stochastic gradient descent to compute $\widehat{\boldsymbol{\theta}}_n$ goes back to Dziugaite et al. (2015) who used SGD to train a generative neural network, and was discussed again in Briol et al. (2019). The algorithm is based on a U-statistic approximation of the MMD criterion and is detailed in Algorithm 12. We also provide a theoretical analysis of the algorithm and numerical simulations, where we test the robust estimation of a uni- and multidimensional univariate Gaussian, a uniform, a Cauchy, and a Gaussian mixture.

2.3.3 A Bayesian estimator

As already explained in Section 1.3, the Bayesian approach is not robust to model misspecification, and the posterior is not consistent in many situations (Barron et al., 1999; Grünwald et al., 2017). We propose again in Chérief-Abdellatif and Alquier (2020) to use the MMD distance to design a robust Bayesian estimator.

We argue in Chapter 9 that the choice of MMD distance is more suited to perform robust estimation than the KL one. To motivate our claim, we show that in Hüber's contamination model for the estimation of a Gaussian mean corrupted by another Gaussian distribution, we exactly recover the true mean when using the minimizer to the true mixture distribution with respect to the MMD distance, whereas we do not when using the KL divergence. Thus, following the core idea of the PAC-Bayes theory, we replace the log-likelihood ℓ_n by the MMD in Bayes' formula, and we call this pseudo-posterior distribution MMD-Bayes:

$$\Pi_{n,\alpha}(d\boldsymbol{\theta}) \propto \exp\left(-\alpha \cdot \mathbb{D}_k^2(P_{\boldsymbol{\theta}}, \widehat{P}_n)\right)\Pi_0(d\boldsymbol{\theta}).$$

In Chapter 9, we show that the MMD-Bayes concentrates to the true distribution when the model is well-specified under a version of the prior mass condition adapted to MMD estimation. When the metric and the radius of the MMD-neighborhoods in the prior mass condition are respectively the MMD metric and $n^{-1/2}$, then the MMD-Bayes concentrates at the optimal rate $n^{-1/2}$ in MMD distance (Tolstikhin et al., 2017).

Furthermore, when the model is misspecified, it is still possible to obtain an oracle inequality for the pseudo-posterior when the prior mass condition is based on a neighborhood of an approximation $P_{\boldsymbol{\theta}^*} = \arg\min_{P_{\boldsymbol{\theta}}} \mathbb{D}_k\left(P_{\boldsymbol{\theta}}, P_0\right)$ of the true distribution instead of the true distribution itself:

Theorem 2.3.3 (Informal). *Under the prior mass condition applied to neighborhoods of the best approximation, we have for any $\alpha \in (0,1)$:*

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_{\boldsymbol{\theta}}, P_0\right) \Pi_{n,\alpha}(\mathrm{d}\boldsymbol{\theta})\right] \leq 8 \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{D}_k^2\left(P_{\boldsymbol{\theta}}, P_0\right) + \frac{16}{n}.$$

Note that the parameter α does not appear in the right-hand side, this phenomenon will be explained in Chapter 9 and has already been encountered for example in Dalalyan et al. (2018). We also provide an example of computation of such a prior mass in this chapter.

To overcome the intractability of the MMD-Bayes in complex models, we use variational inference in Chérief-Abdellatif and Alquier (2020). We show that its variational approximation retains the same theoretical properties under an extended prior mass condition, and we support our theoretical findings with numerical simulations using an SGD algorithm as for MMD-based minimum distance estimation.

In Chapter 8, we provide a simple way to define universal estimation procedures via the MMD metric. In particular:

- ***We give oracle inequalities that imply robust estimation under the i.i.d assumption in Hüber's and in the adversarial contamination models.***
- ***We go beyond the usual i.i.d. framework. By introducing a simple and new weak dependence coefficient expressed as a covariance in the RKHS, we show that the MMD estimator is robust to dependence between observations.***
- ***We also connect our MMD estimator to minimum distance estimation using L_2-metric.***
- ***We give a theoretical analysis of an SGD algorithm used to compute this estimator for finite dimensional models, and we provide empirical evidence in support of this.***

The work in this chapter has been submitted to Bernoulli (Chérief-Abdellatif and Alquier, 2019).

In Chapter 9, we provide a Bayesian version of the MMD estimator that is consistent with optimal properties in well-specified models, and which is robust otherwise. The work in this chapter has been published in Chérief-Abdellatif and Alquier (2020).

Chapter 3

Résumé substantiel

3.1 Consistence de l'inférence variationnelle

Comme indiqué dans le premier chapitre de cette thèse, l'inférence variationnelle a été largement étudiée du point de vue computationnel ces dernières années, mais la littérature n'a accordé que peu d'attention aux propriétés théoriques des approximations variationnelles jusqu'à très récemment. Dans le sillage des travaux d'Alquier and Ridgway (2017) qui étudient la consistence des approximations variationnelles en statistique et les conditions générales qui assurent cette consistence, nous abordons le cas particulier des modèles de mélanges dans le Chapitre 4 et des réseaux de neurones profonds dans le Chapitre 5. Dans le Chapitre 6, nous examinons également la consistence du critère ELBO utilisé pour la sélection des modèles dans la communauté variationnelle Bayésienne.

3.1.1 Modèles de mélanges

Les modèles de mélange sont largement utilisés en statistique Bayésienne et en machine learning, et sont appliqués dans de nombreux domaines comme par exemple la vision par ordinateur (Ayer and Sawhney, 1995), la biologie computationnelle (Pan et al., 2003), l'économie (Deb et al., 2011), ou encore l'analyse des données de transport (Carel and Alquier, 2017), pour n'en citer que quelques-uns. Ils sont également utilisés pour modéliser l'hétérogénéité des populations, ce qui conduit à des méthodes pratiques de clustering (Bouveyron and Brunet-Saumard, 2014; McNicholas, 2016), et ils ont suffisamment de flexibilité pour approcher avec précision presque toutes les densités (Bacharoglou, 2010; Kruijer et al., 2010). Nous renvoyons le lecteur intéressé à Celeux et al. (2018) pour une revue de littérature sur les récentes avancées concernant les mélanges.

Suivant l'esprit d'Alquier and Ridgway (2017) où la théorie est basée sur la posterior tempéré, c'est-à-dire lorsque la fonction de vraisemblance est élevée à une certaine puissance $\alpha \in (0, 1)$, nous calculons dans le Chapitre 4 les vitesses de convergence pour les approximations variationnelles, pour les cas bien spécifiés et mal spécifiés. À titre d'exemple, nous considérons les distributions de mélange multinomiales et Gaussiennes. Tout au long du chapitre, nous fournissons également des algorithmes pratiques pour les

calculs d'approximations de la posterior et nous incluons également de brèves simulations pour illustrer nos résultats théoriques.

Un modèle de mélanges de K composantes est composé de distributions $\sum_{j=1}^{K} \omega_j P_{\boldsymbol{\theta}_j}$ où le vecteur des poids $(\omega_1, \cdots, \omega_K)$ appartient au simplexe $\mathcal{S}_K$ de dimension $K-1$ et où chaque composante $\boldsymbol{\theta}_j$ appartient à un espace de paramètres $\boldsymbol{\Theta}$. Ainsi, le paramètre du mélange est un vecteur $\boldsymbol{\theta} = (\omega_1, ..., \omega_K, \boldsymbol{\theta}_1, ...\boldsymbol{\theta}_K)$ de taille $2K$. Pour estimer un mélange $P_{\boldsymbol{\theta}_0} = \sum_{j=1}^{K} \omega_{0,j} P_{\boldsymbol{\theta}_{0,j}}$ en utilisant une approche Bayésienne, nous définissons une prior $\Pi_0(\boldsymbol{\theta}) = \Pi_{0,\omega}(\omega) \prod_{j=1}^{K} \Pi_{0,j}(\boldsymbol{\theta}_j)$ où $\Pi_{0,\omega} \in \mathcal{M}_1^+(\mathcal{S}_K)$ est une distribution de probabilité sur le simplexe $\mathcal{S}_K$ et où chaque $\Pi_{0,j} \in \mathcal{M}_1^+(\boldsymbol{\Theta})$ est une probabilité sur l'espace des paramètres $\boldsymbol{\Theta}$ (tous deux munis d'une tribu adéquate).

Typiquement, les distributions a posteriori pour les mélanges sont incalculables en pratique et il faut recourir à des méthodes approchées telles que l'inférence variationnelle. L'approximation à champ moyen est particulièrement adaptée aux mélanges étant donné que l'espace des paramètres des mélanges peut s'écrire $\mathcal{S}_K \times \boldsymbol{\Theta} \times \cdots \times \boldsymbol{\Theta}$. La famille variationnelle force les paramètres de chaque composante à être indépendants entre eux, et indépendants des poids:

$$\mathcal{Q} = \left\{ Q(\boldsymbol{\theta}) = Q_\omega(\omega) \prod_{j=1}^{K} Q_j(\boldsymbol{\theta}_j) / Q_\omega \in \mathcal{M}_1^+(\mathcal{S}_K) \ , \ Q_j \in \mathcal{M}_1^+(\boldsymbol{\Theta}) \ \forall j = 1, ..., K \right\}.$$

Le résultat principal est le suivant:

Theorem 3.1.1 (Informel)**.** *Supposons que la condition de prior mass étendue* (1.9) *soit satisfaite pour chaque composante, c'est-à-dire que pour chaque composante j, il existe une distribution $Q_{n,j} \in \mathcal{Q}$ telle que:*

$$\int \mathrm{KL}(P_{\boldsymbol{\theta}_{0,j}} \| P_{\boldsymbol{\theta}_j}) Q_{n,j}(d\boldsymbol{\theta}) \leq r_n \ \text{ and } \ \mathrm{KL}(Q_{n,j} \| \Pi_{0,j}) \leq n r_n.$$

Alors pour une prior de Dirichlet sur les poids, on a pour tout $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha \left(\sum_{j=1}^{K} \omega_j P_{\boldsymbol{\theta}_j}, \sum_{j=1}^{K} \omega_{0,j} P_{\boldsymbol{\theta}_{0,j}} \right) \widetilde{\Pi}_{n,\alpha}(d\boldsymbol{\theta}) \right] \leq \frac{1+\alpha}{1-\alpha} 2K r_n.$$

Ce théorème stipule principalement que lorsque l'estimation d'une distribution dans un modèle est possible à une vitesse r_n, il est alors possible d'estimer un mélange de K distributions du modèle avec une vitesse de convergence égale à Kr_n. Il est intéressant de noter qu'il n'y a pas d'hypothèse de prior mass associée aux poids. La raison est que l'hypothèse de prior mass est satisfaite à la vitesse $\log(nK)/n$ lorsque l'on choisit une prior de Dirichlet $\Pi_\omega = \mathcal{D}_K(\beta_1, ..., \beta_K)$ avec quelques restrictions mineures sur les paramètres $\beta_1, ..., \beta_K$. Comme en pratique, la vitesse de convergence de l'approximation variationnelle associée à un modèle est bien souvent plus lente que $\log(nK)/n$, alors la vitesse de convergence de l'approximation de la posterior associée au mélange est entièrement déterminée par la condition de prior mass sur les différentes composantes. Il est à noter que ce résultat est remarquable car le modèle de mélange ne comporte pratiquement aucune hypothèse.

Comme application, nous abordons le cas des mélanges de Gaussiennes et nous montrons qu'une seule distribution Gaussienne peut être estimée à vitesse $r_n = \log(n)/n$. Dans ce cas, la vitesse de convergence d'une Gaussienne r_n est plus rapide que celui du poids $\log(nK)/n$, et donc la vitesse de convergence final $K\log(nK)/n$ provient de l'estimation des poids du mélange. Ceci est valable lorsque la variance de la Gaussienne est connue mais aussi lorsque nous estimons à la fois la moyenne et la variance.

Le résultat indiquant que la condition de prior mass étendue est satisfaite pour les priors de Dirichlet est intéressant en soi. En effet, lorsqu'il s'agit d'un mélange de V multinomiales, ce résultat est suffisant pour obtenir la vitesse de convergence finale $KV\log(nV)/n$ de l'approximation. Comme déjà expliqué dans la partie 1.1.3, la condition de prior mass est exactement $\Pi_\omega(\mathcal{B}) \geq e^{-nKr_n}$ dans ce cas, et la preuve de la condition de prior mass étendue pour le vecteur de poids revient au calcul de la masse de la prior $\Pi_\omega(\mathcal{B})$ où $\mathcal{B}$ est la boule KL centrée en ω_0 et de rayon Kr_n. En utilisant une prior de Dirichlet, le Lemme 6.1 de Ghosal et al. (2000) aborde le calcul d'une telle masse de la prior pour les boules L_1, que nous étendons ici aux boules KL.

Nous fournissons également un algorithme numérique pour calculer l'approximation variationnelle. Une technique populaire utilisée pour les approximations à champ moyen consiste à optimiser de manière itérative toutes les composantes indépendamment. Néanmoins, cela est en fait aussi difficile que de maximiser la log-vraisemblance d'un mélange, ce qui n'est pas faisable en pratique. Pour surmonter cette difficulté, nous incorporons des variables ω_j^i's dans le programme d'optimisation sans le modifier grâce au Lemme 1.1.1. Ces variables ω_j^i's peuvent être interprétées comme des moyennes a posteriori de variables latentes Z_j^i's comme pour l'algorithme EM (Dempster et al., 1977). Ensuite, le programme équivalent peut être résolu efficacement en utilisant une descente de coordonnées, voir l'Algorithme 6. Notez que lorsque $\alpha = 1$, notre algorithme est exactement équivalent à l'algorithme populaire d'inférence variationnelle par montée de coordonnées (CAVI) qui est largement utilisé dans la communauté VB (Blei et al., 2017). Nous comparons enfin notre algorithme avec CAVI et EM, qui permet d'obtenir des performances comparables pour l'estimation de mélanges Gaussiens.

3.1.2 Réseaux de neurones

Au cours de la dernière décennie, le Deep Learning (DL) ou apprentissage profond a révolutionné l'intelligence artificielle et le numérique (LeCun et al., 2015; Goodfellow et al., 2016). Malheureusement, les propriétés de généralisation de l'apprentissage profond ne sont pas bien comprises, et une ligne de recherche récente étudie les propriétés des réseaux profonds d'un point de vue théorique. En particulier, certains travaux récents ont porté sur l'estimation des fonctions continues en régression non paramétrique, en utilisant des outils fréquentistes ou Bayésiens (Schmidt-Hieber, 2017; Suzuki, 2018, 2019; Rockova and Polson, 2018). Dans le Chapitre 5, nous fournissons la première analyse théorique de l'inférence variationnelle (tempérée) pour l'apprentissage Bayésien profond en régression non paramétrique. Nous montrons, pour un choix d'une famille variationnelle pertinente, que l'approximation variationnelle conserve les mêmes propriétés que la posterior qu'elle approche. Nous donnons la vitesse de convergence dans un cadre général avec n'importe quelle fonction de régression, et nous montrons qu'en particulier, nous récupérons la

vitesse de convergence minimax (aux termes logarithmiques près) pour le cas des fonctions continues au sens de Hölder pour certaines architectures du réseau.

Nous considérons une régression non-paramétrique avec une collection de variables aléatoires i.i.d. $(X_i, Y_i) \in [-1,1]^d \times \mathbb{R}$ pour $i = 1, ..., n$:

$$\begin{cases} X_i \sim \mathcal{U}([-1,1]^d), \\ Y_i = f_0(X_i) + \zeta_i \end{cases}$$

où $\mathcal{U}([-1,1]^d)$ est la distribution uniforme sur l'intervalle $[-1,1]^d$, $\zeta_1, ..., \zeta_n$ sont des variables aléatoires Gaussiennes i.i.d. centrées et de variance connue σ^2, et $f_0 : [-1,1]^d \to \mathbb{R}$ est la vraie distribution inconnue. Nous estimons f_0 en utilisant des résonnes neuronaux profonds $f_{\boldsymbol{\theta}}$ de sparsité S, de profondeur L et de largeur D.

Là encore, nous adoptons une approche Bayésienne. Nous choisissons une *spike-and-slab* prior induisant de la sparsité, pour lequel un nombre S de neurones sélectionnés du réseau suit une distribution Gaussienne centré réduite, et tous les autres neurones sont mis à 0 avec probabilité 1. L'ensemble variationnel est également composé de Gaussiennes spike-and-slab.

La contribution principale de l'article est une borne sur l'erreur de généralisation pour l'inférence variationnelle en apprentissage profond sparse en régression nonparamétrique:

Theorem 3.1.2 (Informel)**.** *Pour tout $\alpha \in (0,1)$, pour tout $B > 0$,*

$$\mathbb{E}\left[\int \|f_{\boldsymbol{\theta}} - f_0\|_2^2 \widetilde{\Pi}_{n,\alpha}(d\boldsymbol{\theta})\right] \leq \frac{2}{1-\alpha} \inf_{\|\boldsymbol{\theta}^*\|_\infty \leq B} \|f_{\boldsymbol{\theta}^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n,$$

où la vitesse de convergence r_n est d'ordre $\frac{LS}{n}\log(BD)$.

La vitesse de convergence dans le terme de droite de l'inégalité de l'oracle, qui dépend linéairement du nombre de couches et de la sparsité, est déterminé par la condition de prior mass étendue précédente, et on retrouve exactement la vitesse de convergence du minimiseur du risque empirique pour les réseaux de neurones profonds qui est obtenue en utilisant différentes techniques de preuve, en calculant notamment l'entropie de couverture locale i.e. le logarithme du nombre de boules r_n nécessaires pour couvrir un voisinage de la vraie fonction de régression (Schmidt-Hieber, 2017; Suzuki, 2019).

En particulier, lorsque la véritable fonction de régression est continue au sens de Hölder, en supposant que la sparsité, la profondeur et la largeur du réseau sont correctement choisies comme le suggèrent Rockova and Polson (2018), cela implique que l'approximation variationnelle se concentre vers f_0 et bénéficie des mêmes vitesse de convergence quasi-minimax que celles obtenues pour les posteriors exactes tempérées et régulières. Le résultat est établi en calculant l'inégalité oracle PAC-Bayésienne précédente pour une fonction de régression Hölder et en appliquant le résultat d'approximation de Schmidt-Hieber (2017).

Nous considérons en outre l'extension de l'inégalité oracle lorsque l'approximation est calculée via un algorithme d'optimisation dont l'erreur est mesurée par son effet sur l'ELBO. Plus précisément, lorsque nous considérons un algorithme $(\widetilde{\Pi}_{n,\alpha}^{(j)})_j$ pour calculer l'approximation idéale $\widetilde{\Pi}_{n,\alpha}$, il existe un terme supplémentaire dans l'erreur de généralisation :

Theorem 3.1.3 (Informal). *Pour tout $\alpha \in (0,1)$, pour tout $B > 0$ et quelque soit le nombre d'itérations j,*

$$\mathbb{E}\left[\int \|f_{\boldsymbol{\theta}} - f_0\|_2^2 \widetilde{\Pi}_{n,\alpha}^{(j)}(d\boldsymbol{\theta})\right] \leq \frac{2}{1-\alpha} \inf_{\|\boldsymbol{\theta}^*\|_\infty \leq B} \|f_{\boldsymbol{\theta}^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n + \frac{\mathbb{E}[\Delta_{n,j}]}{n},$$

où $\Delta_{n,j}$ est la différence entre la valeur maximale de l'ELBO et la valeur de l'ELBO à la $j^{\text{ème}}$ itération de l'algorithme.

Ainsi, l'algorithme $\widetilde{\Pi}_{n,\alpha}^{(j)}$ converge à la même vitesse r_n que l'approximation variationnelle idéale qu'il calcule $\widetilde{\Pi}_{n,\alpha}$ tant que $\mathbb{E}[\Delta_{n,j}] \lesssim LS \log(BD)$.

3.1.3 Sélection de modèles

La question qui se pose naturellement lorsqu'on traite de mélanges et de réseaux de neurones profonds est respectivement la question de la sélection du nombre de composantes et la question de la sélection de l'architecture du réseau. Par exemple, le choix de l'architecture dans l'apprentissage profond est crucial et peut conduire à une convergence plus rapide et à une meilleure approximation. De même, dans la modélisation Bayésienne des mélanges et dans les situations où le nombre de composantes du mélange est inconnu, on utilise souvent en pratique la méthode Bayésienne hiérarchique complète qui implique de mettre une prior sur le nombre de composantes. Cela nécessite des algorithmes de MCMC à saut réversible pour effectuer les calculs, mais la conception de tels algorithmes est une question délicate dans de nombreuses situations pratiques et ils ne s'adaptent pas bien à la dimension des données. Par conséquent, toute solution rapide à ce problème est d'un grand intérêt.

Plus généralement, cela soulève la question de la sélection de modèles et de la conception d'un critère de sélection adaptatif qui sélectionne un modèle optimal. Le critère de maximisation de l'ELBO est largement utilisé dans la communauté VB et est connu pour bien fonctionner en pratique. Plus précisément, nous considérons plusieurs modèles indexés par K et les approximations variationnelles associées $\widetilde{\Pi}_{n,\alpha}^{K}$, et nous attribuons un poids π_K à chaque modèle. Le critère de maximisation de l'ELBO consiste à choisir le modèle qui fournit l'approximation la plus proche de la log-vraisemblance marginale $\text{ELBO}(K)$, c'est-à-dire la valeur maximale de l'ELBO associée au modèle K. Nous proposons dans Chérief-Abdellatif (2019a) une version pénalisée du critère de maximisation de l'ELBO qui est équivalent à celui qui est habituellement utilisé lorsque l'on considère un nombre fini de modèles et des poids uniformes :

$$\widehat{K} = \arg\max_K \left\{\text{ELBO}(K) - \log\left(\frac{1}{\pi_K}\right)\right\}.$$

Le critère de maximisation de l'ELBO n'a jamais été justifié en théorie. Dans le Chapitre 6, nous montrons que ce critère est adaptatif et sélectionne une approximation variationnelle qui permet d'atteindre la vitesse de convergence optimale entre les modèles concurrents. En particulier, nous montrons respectivement dans les Chapitres 4 et 5 que le critère ELBO sélectionne un certain nombre de composantes et une architecture

de réseau qui donnent la vitesse optimale et n'entraînent pas de surapprentissage. Nous montrons également dans le Chapitre 6 que la vitesse de convergence minimax est atteinte pour l'ACP probabiliste.

Nous savons par Alquier and Ridgway (2017) que dès qu'il existe un vrai modèle qui satisfait la condition de prior mass étendue, alors l'approximation variationnelle correspondant au vrai modèle est consistence à la vitesse définie par la condition de prior mass. Le message principal de Chérief-Abdellatif (2019a) est que même si nous ne savons pas quel modèle est le vrai, le critère ELBO fournit un modèle (pas nécessairement le vrai) tel que l'approximation correspondante est consistente et atteint de manière adaptative la vitesse de convergence associé au vrai modèle. Plus précisément:

Theorem 3.1.4 (Informel)**.** *Supposons que l'hypothèse 1.9 soit satisfaite pour le vrai modèle K_0. Alors pour tout $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{\theta}}, P^0)\widetilde{\Pi}^{\widehat{K}}_{n,\alpha}(d\boldsymbol{\theta})\right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

En fait, la vitesse globale est composé de la vitesse de convergence r_n correspondant au vrai modèle et d'un terme de complexité qui reflète la croyance sur les différents modèles. Par exemple, si nous rangeons un nombre dénombrable de modèles en fonction de notre croyance, alors le terme de complexité sera de l'ordre de K_0/n en choisissant $\pi_K = 2^{-K}$. En pratique, le terme global est de l'ordre de r_n.

L'application de ce résultat aux modèles de mélanges et aux réseaux de neurones profonds se trouve dans les Chapitres 4 et 5. Nous considérons également dans le Chapitre 6 le cas de l'ACP probabiliste. Dans les prochains paragraphes, les matrices seront indiquées en majuscules et en gras. Nous considérons le modèle

$$X_i = \boldsymbol{W}_0 Z_i + \varepsilon_i$$

avec des variables aléatoires Gaussiennes i.i.d. $Z_i \sim \mathcal{N}(0, \boldsymbol{I}_{K_0})$ et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$, où $\boldsymbol{I}_d$ et $\boldsymbol{I}_{K_0}$ sont les matrices d'identité de dimensions respectives d et K_0 ($K_0 < d$), $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K}$ est la matrice de rang K_0 qui contient les axes principaux et σ^2 est un terme de bruit connu. K_0 correspond ici à la "vraie dimension" des données. K_0 est inconnu ici et nous considérons différents modèles correspondant à différentes valeurs de $K = 1, ..., d$ et différentes matrices correspondantes $\boldsymbol{W} \in \mathbb{R}^{d\times K}$.

Nous mettons des poids égaux sur les modèles π_K pour chaque entier $K = 1, ..., d$. Étant donné le rang K, nous plaçons une prior sur la matrice $\boldsymbol{W}$ de rang K afin de définir une distribution sur les axes principaux. Nous choisissons des priors Gaussiennes indépendantes sur les colonnes de $\boldsymbol{W}$. Nous considérons également des approximations variationnelles Gaussiennes indépendantes sur les colonnes de $\boldsymbol{W}$. En utilisant un opérateur de clipping, il est possible d'obtenir la consistence au sens de la norme de Frobenius $\|\cdot\|_F$ de l'approximation variationnelle sélectionnée sous l'hypothèse classique que la norme spectrale $\|\cdot\|_2$ de la matrice vraie $\boldsymbol{W}_0$ est bornée. Le symbole $\mathcal{O}$ dans le résultat suivant cache des constantes universelles qui sont indépendantes de α, d, K_0 et n:

Theorem 3.1.5 (Informel)**.** *Pour tout $\alpha \in (0,1)$, si $\|\boldsymbol{W}_0\|_2 \leq B$, alors:*

$$\mathbb{E}\left[\int \left\| clip_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0\boldsymbol{W}_0^T \right\|_F^2 \widetilde{\Pi}^{\widehat{K}}_{n,\alpha}(d\boldsymbol{W})\right] = \mathcal{O}\left(\frac{1+\alpha}{\alpha(1-\alpha)} \cdot \frac{dK_0 \log(dn)}{n}\right),$$

où $clip_B(\boldsymbol{A})$ *est la matrice dont le coefficient* (i,j) *est égal à* $\begin{cases} \boldsymbol{A}_{i,j} & \text{if } |\boldsymbol{A}_{i,j}| \leq B^2 \\ B^2 & \text{si } \boldsymbol{A}_{i,j} \geq B^2 \\ -B^2 & \text{sinon.} \end{cases}$

Chacun des chapitres suivants est autonome et peut être lu indépendamment. Voici un bref résumé de leurs contributions :

- ***Le Chapitre 5 étudie la consistence des approximations variationnelles de posteriors tempérées pour les mélanges, et donne des vitesse de convergence. Nous abordons également le problème de la sélection du nombre de composantes en maximisant l'ELBO. Ce travail a fait l'objet d'une publication (Chérief-Abdellatif and Alquier, 2018).***

- ***Le Chapitre 6 fournit des bornes de généralisation non asymptotiques assurant la consistence des réseaux de neurones Bayésiens lorsqu'une approximation est utilisée à la place de la distribution a posteriori exacte, ainsi que des vitesse de convergence. Nous montrons qu'elle conduit à une estimation quasi-minimax de fonctions continues pour un choix judicieux de l'architecture. Nous utilisons également le critère ELBO pour la sélection de l'architecture. Ce travail a fait l'objet d'un article accepté à ICML 2020 (Chérief-Abdellatif, 2019b).***

- ***Le Chapitre 7 justifie l'utilisation de la stratégie de maximisation de l'ELBO d'un point de vue théorique. Nous illustrons nos résultats théoriques par une application à la sélection du nombre de composantes principales pour l'ACP probabiliste. Ce travail a fait l'objet d'une publication (Chérief-Abdellatif, 2019a).***

3.2 Inférence variationnelle en ligne

Après les analyses précédentes dans le cas batch, nous étudions l'inférence variationnelle dans le Chapitre 7 en apprentissage séquentiel. Aucune définition consensuelle de l'inférence variationnelle Bayésienne n'a été donnée en apprentissage en ligne. Comme déjà détaillé dans la partie 1.2, conserver la définition standard de l'approximation variationnelle serait une solution simpliste qui retirerait l'aspect séquentiel du problème et conduirait à des problèmes d'ordre computationnel. Dans le Chapitre 7, nous proposons la première analyse théorique de VI dans le cadre de l'apprentissage en ligne avec des données séquentielles. Nous étudions plusieurs classes d'algorithmes en ligne qui sont inspirés de stratégies populaires utilisées en optimisation en ligne telles que la descente en gradient et follow-the-regularized-leader, et nous fournissons des bornes de généralisation dans le cas convexe. Nos techniques de preuve sont basées sur la convexité de la fonction de perte, mais nous soutenons qu'elles devraient pouvoir tenir dans un cadre plus général.

3.2.1 Approximations variationnelles d'EWA

Dans la littérature, deux extensions principales du VB au cadre en ligne ont été formulés. La première définition de VB en ligne étend la définition de Bayes utilisée en statistique batch par le lemme de Donsker-Varadhan (voir la formule 1.5), et restreint l'ensemble de minimisation à une famille de distributions tractables. Lorsque l'on considère les approximations paramétriques $\mathcal{Q} = \{Q_\mu / \mu \in \mathcal{M}\}$, on définit les approximations variationnelles comme $\widetilde{\Pi}_{t,\alpha} = Q_{\mu_t}$ en utilisant une suite de paramètres $(\mu_t)_t$ avec $Q_{\mu_0} = \Pi_0$:

$$\mu_t = \arg\min_{\mu \in \mathcal{M}} \left\{ \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\theta} \sim Q_\mu} \left[\ell_s(\boldsymbol{\theta}) \right] + \frac{\mathrm{KL}(Q \| Q_{\mu_0})}{\alpha} \right\}.$$

Le principal inconvénient de cette méthode est qu'elle perd l'aspect en ligne et exige que l'ensemble des échantillons de données soit chargé en mémoire à chaque étape, ce qui est très coûteux en termes de calcul. C'est pourtant l'approche adoptée dans Guhaniyogi et al. (2013) où les auteurs proposent deux méthodes : la première est une généralisation de l'Équation 1.5, en minimisant sur un ensemble convexe non paramétrique de distributions $\mathcal{Q}$ via une descente de gradient fonctionnel. Le problème est qu'à l'exception du cas trivial où $\mathcal{Q} = \mathcal{M}_1^+(\boldsymbol{\Theta})$, ils ne fournissent aucun exemple de tels ensembles en pratique. Ils utilisent également une famille paramétrique $\mathcal{Q} = \{Q_\mu, \mu \in M\}$. Cependant, ils modifient l'ordre dans le terme en KL, ce qui implique un terme intraitable qu'ils calculent en utilisant du Monte-Carlo. De plus, la quantité apparaissant dans leur borne de regret n'est pas naturelle et ne donne aucune garantie sur l'objectif initial, à savoir la minimisation des ℓ_t.

Nous remarquons que notre approximation variationnelle est formulée comme une stratégie de type FTRL appliquée à la perte espérée $\mu \to \mathbb{E}_{\boldsymbol{\theta} \sim Q_\mu}[\ell_s(\boldsymbol{\theta})]$. Nous proposons donc de linéariser la fonction de perte (espéré) comme proposé dans la partie 1.2:

$$\mu_t = \arg\min_{\mu \in \mathcal{M}} \left\{ \mu^T \sum_{s=1}^{t} \nabla_\mu \mathbb{E}_{\boldsymbol{\theta} \sim Q_{\mu_s}} \left[\ell_s(\boldsymbol{\theta}) \right] + \frac{\mathrm{KL}(Q \| Q_{\mu_0})}{\alpha} \right\}.$$

Nous désignerons cet algorithme *Sequential Variational Approximation (SVA)* dans le reste de ce manuscrit, voir l'Algorithme 10.

Quant à l'algorithme de descente de gradient, il existe une autre formulation de ce programme d'optimisation. Plutôt que de considérer la perte cumulée et de régulariser (en utilisant une norme induite par la divergence KL) par rapport à la première approximation, il est également possible de ne considérer que la dernière perte et de régulariser par rapport à la dernière approximation. Néanmoins, contrairement à la descente de gradient, la deuxième formulation n'est pas équivalente à la première et conduit à un programme d'optimisation différent :

$$\mu_t = \arg\min_{\mu \in \mathcal{M}} \left\{ \mu^T \nabla_\mu \mathbb{E}_{\boldsymbol{\theta} \sim Q_{\mu_t}} \left[\ell_s(\boldsymbol{\theta}) \right] + \frac{\mathrm{KL}(Q \| Q_{\mu_{t-1}})}{\alpha} \right\}.$$

En fait, cet algorithme est conforme à l'autre approche qui étend la définition séquentielle de l'EWA donnée dans l'Algorithme 5 : $\Pi_{t+1,\alpha}(d\boldsymbol{\theta}) \propto \exp(-\alpha \ell_t(\boldsymbol{\theta})) \Pi_{t,\alpha}(d\boldsymbol{\theta})$. Ce point

de vue est particulièrement séduisant car il permet de conserver l'aspect en ligne du problème et il n'est pas nécessaire de stocker les données. La mise à jour est obtenue par une formule de la forme $\widetilde{\Pi}_{t,\alpha} = F(\widetilde{\Pi}_{t-1,\alpha}, \ell_t)$. C'est par exemple le point de vue adopté par Broderick et al. (2013), qui a été approfondi par Nguyen et al. (2017a) et Zeno et al. (2018) où les auteurs ont combiné la définition batch de VB et l'approche de mise à jour en ligne pour obtenir la définition suivante en deux étapes :

- Étape d'approximation : $\bar{\Pi}_{t,\alpha}(d\boldsymbol{\theta}) \propto \exp(-\alpha \ell_t(\boldsymbol{\theta})) Q_{\mu_{t-1}}(d\boldsymbol{\theta})$.
- Étape de projection : $\mu_t = \arg\min_{\mu \in \mathcal{M}} \mathrm{KL}(Q \| \bar{\Pi}_{t,\alpha})$.

En intégrant l'étape d'approximation dans l'étape de projection, on obtient

$$\mu_t = \arg\min_{Q \in \mathcal{Q}} \left\{ \mathbb{E}_{\boldsymbol{\theta} \sim Q}\left[\ell_t(\boldsymbol{\theta})\right] + \frac{\mathrm{KL}(Q \| Q_{\mu_{t-1}})}{\alpha} \right\}.$$

Par conséquent, notre algorithme est une version linéarisée de l'approximation variationnelle précédente d'approximation/projection. Nous désignerons notre algorithme *Streaming Variational Bayes (SVB)* dans la suite, voir l'Algorithme 10. Quant à l'approche utilisant le lemme de Donsker-Varadhan (1.5), aucune analyse rigoureuse des propriétés de généralisation des approximations obtenues par cette approche n'a été fournie et il n'est pas certain que le calcul de $\widetilde{\Pi}_{t,\alpha}$ soit réalisable.

En outre, les algorithmes SVA et SVB sont tous deux calculables et disposent de formules de mises à jour explicites pour les ensembles variationnels utilisés en pratique. Par exemple, en utilisant la classe des approximations Gaussiennes à champ moyen $Q_{(m,\sigma)} = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ de moyenne m et de matrice de covariance diagonale dont la diagonale est le vecteur $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)^T$, SVA et SVB conduisent aux mises à jour de type descente de gradient suivantes:

$$\text{SVA:} \begin{cases} m_{t+1} & \leftarrow m_t - \alpha \sigma_0^2 \bar{g}_{m_t}, \\ g_{t+1} & \leftarrow g_t + \bar{g}_{\sigma_t}, \\ \sigma_{t+1} & \leftarrow h\left(\frac{1}{2}\alpha \sigma_0 g_{t+1}\right) \sigma_0. \end{cases}$$

$$\text{SVB:} \begin{cases} m_{t+1} & \leftarrow m_t - \alpha \sigma_t^2 \bar{g}_{m_t}, \\ \sigma_{t+1} & \leftarrow h\left(\frac{1}{2}\alpha \sigma_t \bar{g}_{\sigma_t}\right) \sigma_t. \end{cases}$$

où $\bar{g}_{m_t}$ et $\bar{g}_{\sigma_t}$ représentent respectivement le gradient de la perte attendue $(m, \sigma) \to \mathbb{E}[\ell_t(\boldsymbol{\theta})]$ par rapport à m et σ évalué en $(m, \sigma) = (m_t, \sigma_t)$, où $h(x) := \sqrt{1 + x^2} - x$ et toutes les opérations vectorielles sont appliquées composantes par composantes.

3.2.2 Bornes de regret

Nous prévoyons dans le Chapitre 7 des bornes de regret pour SVA. Les principales hypothèses requises sont le caractère Lipschitz et la convexité de la perte espéré $\mu \to \mathbb{E}_{\boldsymbol{\theta} \sim Q_\mu}[\ell_t(\boldsymbol{\theta})]$, ainsi que la forte convexité de la régularisation KL par rapport à μ.

Theorem 3.2.1 (Informel). *Pour des pertes espérées L-Lipschitz et convexes et un régularisateur KL fortement convexe, SVA a les regrets suivants:*

$$\sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}_t) \widetilde{\Pi}_{t,\alpha}(d\boldsymbol{\theta}_t) \leq \inf_{\mu \in \mathcal{M}} \left\{ \sum_{t=1}^{T} \int \ell_t(\boldsymbol{\theta}) Q_\mu(d\boldsymbol{\theta}) + \frac{\alpha L^2 T}{\sigma} + \frac{\mathrm{KL}(Q_\mu, \Pi_0)}{\alpha} \right\}.$$

Comme souhaité, ce résultat est presque le même que celui obtenu dans le Théorème 1.2.5 où l'infimum est limité à la famille variationnelle, la borne supérieure B est remplacée par la constante de Lipschitz L, et le facteur 8 par le paramètre de forte convexité σ. Néanmoins, la preuve est très différente et repose sur des arguments provenant de l'optimisation convexe en ligne (Shalev-Shwartz, 2012; Hazan, 2016). Notez que comme dans la partie 1.2, les techniques online-to-batch peuvent conduire à des erreurs de généralisation sous une hypothèse de convexité sur la fonction de perte ℓ_t d'ordre :

$$\mathbb{E}_{\mathcal{D}_T \sim P_0}[R(\bar{\boldsymbol{\theta}}_T)] \leq \inf_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) + \mathcal{O}\left(\frac{L}{\sigma} \sqrt{\frac{\log(dT)}{T}} \right).$$

Nous fournissons également dans le Chapitre 7 un regret pour SVB pour la famille Gaussienne à champs moyens $Q_{(m,\sigma)} = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ avec comme paramétres moyenne/écart type en prenant un pas d'apprentissage dynamique et multidimensionnel $\alpha_t = (\alpha_{t,1}, \cdots, \alpha_{t,j})$. Le regret nécessite un ensemble de projection des paramètres $\mathcal{M} = \mathcal{M}_m \times \mathcal{M}_\sigma$ avec une étape supplémentaire de projection où $\mathcal{M}_m$ et $\mathcal{M}_\sigma$ sont des sous-ensembles fermés, bornés et convexes de $\mathbb{R}^d$ et $\mathbb{R}^d_{\geq 0}$ respectivement, avec $0 \in \mathcal{M}_\sigma$. Nous définissons également le diamètre $D^2 = \sup\{\|m - m'\|_2^2 + \|\sigma\|_2^2 / m, m' \in \mathcal{M}_m, \sigma \in \mathcal{M}_\sigma\}$. Notons que la forte convexité de la régularisation est toujours satisfaite dans ce cas.

Theorem 3.2.2 (Informel). *Pour des pertes convexes ℓ_t et des pertes espérées L-Lipschitz et convexes, et pour $\widehat{\boldsymbol{\theta}}_t \sim Q_{(m_t,\sigma_t)}$, nous avons pour un certain choix du pas d'apprentissage:*

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{\theta}}_t) \leq \inf_{\theta \in \mathcal{M}_m} \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) + DL\sqrt{2T}.$$

De plus, si les pertes attendues sont fortement convexes, nous atteignons même une vitesse logarithmique:

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{\theta}}_t) \leq \inf_{\theta \in \mathcal{M}_m} \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) + \frac{L^2(1 + \log T)}{H}.$$

Une fois de plus, les résultats sont similaires au cas Bayésien classique mais sont directement exprimés en termes de paramètres $\boldsymbol{\theta}$ au lieu de leurs espérances. Il est même possible, en optimisant sur $\mathcal{M}_m = \{m \in \mathbb{R}^d : \|m\|_2 \leq \bar{M}\}$ et $\mathcal{M}_\sigma = \{\sigma \in \mathbb{R}^d_+ : \|\sigma\|_2 \leq \bar{S}\}$, d'obtenir des erreurs de généralisation d'ordre $L(4\bar{M}^2 + \bar{S}^2)^{1/2} T^{-1/2}$ indépendantes de la dimension.

3.2.3 Au-delà de la convexité

Nos preuves sont basées sur un argument de convexité. Cependant, nous nous attendons à ce qu'elles restent valables plus généralement, même pour certaines pertes espérées

non convexes. Cela est particulièrement intéressant car cela nous aiderait à obtenir des garanties de généralisation pour l'algorithme d'inférence variationnelle à gradient naturel (NGVI) (Sato, 2001; Hoffman et al., 2013; Khan and Lin, 2017) en utilisant une famille variationnelle exponentielle.

Cet algorithme est largement utilisé en apprentissage stochastique mais peut être facilement modifié pour le cadre en ligne. La méthode présentée dans Khan and Lin (2017) exploite la dualité entre le paramètre des moments μ et le paramètre naturel λ de la famille exponentielle, et s'écrit:

$$\lambda_t = \lambda_{t-1} - \alpha \nabla_\mu \mathbb{E}_{\boldsymbol{\theta} \sim Q_{\mu_{t-1}}} [\ell_t(\boldsymbol{\theta})],$$

avec une éventuelle étape de projection.

En particulier, cet algorithme de gradient naturel peut être s'écrire simplement comme l'Algorithme SVB appliqué à une famille variationnelle exponentielle avec un paramètre des moments. Par exemple, en utilisant une famille Gaussienne à champ moyen, le paramètre des moments est simplement composé des deux premiers moments, c'est-à-dire la moyenne et la matrice de corrélation. On sait qu'il fonctionne très bien en pratique, et il est démontré dans Chérief-Abdellatif et al. (2019) qu'il surpasse à la fois SVA et SVB appliqués à la paramétrisation usuelle moyenne/écart type (m, σ) dans plusieurs cadres et pour différentes pertes. Malheureusement, la perte espérée n'est pas convexe par rapport au paramètre des moments pour la plupart des pertes bien qu'elle soit convexe par rapport à (m, σ), et nous ne sommes donc pas en mesure d'analyser les performances de l'Algorithme NGVI.

L'une des principales conclusions et suggestions empiriques du Chapitre 7 est que les propriétés de généralisation de l'inférence variationnelle en ligne semblent aller au-delà de l'hypothèse convexe requise pour obtenir les résultats théoriques, et que la convexité de la perte espérée n'est pas la pierre angulaire de la généralisation de VB en ligne.

Dans le Chapitre 7, nous calculons les premières limites de généralisation de l'inférence variationnelle en ligne. En utilisant les méthodes existantes, nous proposons quelques méthodes en ligne pour l'inférence variationnelle, à savoir les algorithmes SVA et SVB, et nous fournissons des bornes de généralisation. Nous étayons nos conclusions théoriques avec des expériences numériques sur des données simulées et réelles. Nous observons que l'Algorithme NGVI surpasse toutes les autres méthodes, nous conjecturons que l'hypothèse théorique de convexité sur la perte espérée peut être relâchée en pratique, et nous pensons que notre analyse théorique peut être étendue à l'algorithme NGVI. Ce travail a fait l'objet d'une publication (Chérief-Abdellatif et al., 2019).

3.3 Robustesse via Maximum Mean Discrepancy

Comme indiqué dans le Chapitre 1.3, la plupart des estimateurs qui sont à la fois consistents et statistiquement optimaux lorsque la distribution génératrice des données appartient au modèle et robuste aux petits écarts par rapport aux hypothèses du modèle ne

sont pas tractables en pratique. L'estimateur par minimisation de distance récemment introduit par Briol et al. (2019) et basé sur la Maximum Mean Discrepancy (DMM) est consistent avec des vitesses optimales lorsque le modèle est bien spécifié, robuste aux erreurs de spécification, et peut être facilement calculé en utilisant une descente de gradient stochastique. Dans le Chapitre 8, nous montrons que cet estimateur est robuste à la fois à la dépendance et à la présence de valeurs aberrantes dans le jeu de données. Nous relions également cet estimateur basé sur le MMD à l'estimation L_2, et nous proposons une analyse théorique de l'algorithme de gradient utilisé pour calculer l'estimateur. Nous appuyons notre analyse théorique par des simulations numériques. Nous proposons également une version Bayésienne dans le Chapitre 9, où nous étudions ses propriétés de concentration sous une condition de prior mass et fournissons un algorithme explicite pour calculer la pseudo postérieure basée sur la MMD en utilisant l'inférence variationnelle.

3.3.1 L'estimateur MMD

Dans la suite, nous considérons un noyau défini positif k, c'est-à-dire une fonction symétrique $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que pour tout entier $n \geq 1$, pour tout $x_1, ..., x_n \in \mathcal{X}$ et pour tout $c_1, ..., c_n \in \mathbb{R}$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0,$$

et l'espace de Hilbert à noyau reproduisant correspondant (RKHS) $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ qui satisfait la propriété de reproduction $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$ pour toute fonction $f \in \mathcal{H}_k$ et tout $x \in \mathcal{X}$. Nous supposons que le noyau est borné par une constante positive, disons 1 sans perte de généralité.

Nous définissons la norme induite $\|m\|_{\mathcal{H}_k} = \langle m, m \rangle_{\mathcal{H}_k}$ et la boule unité $\mathbb{B}_k = \{m \in \mathcal{H}_k : \|m\|_{\mathcal{H}_k} \leq 1\}$.

Le *kernel mean embedding* d'une mesure de probabilité P est l'application $\mu_P \in \mathcal{H}_k$ telle que :

$$\mu_P(\cdot) := \mathbb{E}_{X \sim P}[k(X, \cdot)] \in \mathcal{H}_k.$$

Nous supposons également que le noyau est *caractéristique*, ce qui signifie que $P \mapsto \mu_P$ est injectif. Nous pouvons alors définir la distance :

$$\mathbb{D}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$$

qui est la distance MMD entre P et Q.

Étant donné un ensemble de paramètres $\boldsymbol{\Theta}$, nous définissons l'estimateur MMD $\widehat{\boldsymbol{\theta}}_n$ comme dans Briol et al. (2019):

$$\mathbb{D}_k(P_{\widehat{\boldsymbol{\theta}}_n}, \widehat{P}_n) = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{D}_k(P_{\boldsymbol{\theta}}, \widehat{P}_n).$$

Il est également possible d'envisager un minimiseur approché lorsque le minimum exact n'existe pas.

3.3.2 Robustesse à la spécification et à la dépendance

Le Chapitre 8 examine les propriétés d'universalité de l'estimation MMD, notamment en ce qui concerne la dépendance à une mauvaise spécification et à la dépendance.

Une borne de généralisation:

Le premier résultat est une inégalité oracle qui est valable même en case de dépendance entre les variables et qui induit une décroissance en $n^{-1/2}$ de l'erreur de généralisation mesurée en la distance MMD, vitesse qui est optimale(Tolstikhin et al., 2017):

Theorem 3.3.1 (Informal). *Sous certaines hypothèses de dépendance faible impliquant deux constantes Σ et Γ, nous avons pour tout $\delta \in (0,1)$, avec probabilité supérieure ou égale à $1-\delta$,*

$$\mathbb{D}_k\left(P_{\widehat{\boldsymbol{\theta}}_n}, P^0\right) \leq \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{D}_k\left(P_{\boldsymbol{\theta}}, P^0\right) + 2\frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}.$$

En particulier, sous l'hypothèse i.i.d. $\Sigma = \Gamma = 0$.

Le résultat ci-dessus est basé sur deux principales hypothèses de dépendance. La première est suffisante pour obtenir un résultat en espérance. Nous introduisons dans Chérief-Abdellatif and Alquier (2019) un nouveau coefficient de dépendance faible dans le sillage des travaux de Doukhan and Louhichi (1999) qui est exprimé sous la forme d'une covariance dans le RKHS associé au noyau. Un tel coefficient est très simple à utiliser en pratique, est beaucoup plus facile à calculer que les coefficients de mélange et la sommabilité de la série des coefficients inclut une classe plus large de processus. Lorsque la somme des séries existe, on dit que le processus stochastique est *faiblement dépendant*. Σ désigne la somme des séries dans le théorème précédent. Nous étudions plusieurs exemples de processus qui ne sont pas mélangeant pas mais qui sont malgré tout faiblement dépendants dans le Chapitre 8.

La deuxième hypothèse est fondamentale pour obtenir un résultat en probabilité. En effet, une inégalité de concentration majorant la distance MMD entre la distribution empirique et la vraie distribution est nécessaire pour obtenir une telle inégalité oracle. Par exemple, le résultat de Briol et al. (2019) est basé sur l'inégalité de McDiarmid (McDiarmid, 1989), mais cette inégalité de type Hoeffding n'est valable que dans le cadre indépendant. Nous exploitons donc ici une version de l'inégalité de McDiarmid conçue pour les séries temporelles, qui est disponible sous une hypothèse de décroissance polynomiale sur certains coefficients de mélange $(\gamma_{i,j})_{1\leq i<j}$ (Rio, 2013). Ici encore, nous adaptons l'hypothèse de décroissance au RKHS $\mathcal{H}_k$, et la constante Γ dans le théorème précédent dsigne la somme des séries de ces coefficients de mélange.

Bornes en distance L_2:

Nous relions également nos bornes en distance MMD à d'autres métriques, en particulier à la distance L_2. Certaines tentatives passées de conception d'estimateurs robustes

aux erreurs de mauvaise spécification donnent des bornes en TV ou en distances de Hellinger (Baraud and Birgé, 2016; Devroye and Lugosi, 2001), et l'utilisation de cette perte quadratique comme estimateur par minimisation de distance est déconseillée. La première raison principale est que la métrique L_2 n'est pas universelle. En effet, toutes les mesures de probabilité n'ont pas nécessairement une densité par rapport à une mesure de référence (par exemple, la mesure de Lebesgue), et certaines d'entre elles peuvent en avoir une mais elle n'est pas L_2-intégrée. En outre, une telle densité dépend fortement du choix de la mesure de référence, comme expliqué dans la partie 1.3.

Nous soutenons dans le Chapitre 8 que pour les noyaux de la forme $k(x,y) = F(\|x-y\|/\gamma)$ pour une fonction $F : [0,+\infty) \to [0,1]$, la distance MMD peut être exprimé comme une approximation de la distance quadratique qui est bien définie pour toute distribution de probabilité, et qu'il est possible d'obtenir des inégalités oracle dans la distance L_2 lorsque les densités existent et sont intégrables. Par exemple, pour le noyau Gaussien $k_\gamma(x,y) = \exp(-\|x-y\|^2/\gamma^2)$, $\mathbb{D}_{k_\gamma}(P,Q) \sim \pi^{\frac{d}{4}}\gamma^{\frac{d}{2}}\|p-q\|_{L_2}$ lorsque $\gamma \to 0$ et où p et q sont des densités de P et Q respectivement par rapport à la mesure Lebesgue. Ainsi, pour γ assez petit, nous donnons un sens à l'estimation L_2 même pour des densités qui ne sont pas L_2-intégrables. En outre, la distance MMD ne dépend d'aucune mesure de référence, bien qu'il dépende du choix du noyau. Cette dépendance en k est en fait une caractéristique intéressante car elle donne la possibilité de prendre en compte la géométrie sous-jacente de $\mathcal{X}$ via le choix d'une distance sur cet espace, une propriété qui explique la popularité de la distance de Wasserstein en statistique. Par exemple, $\mathbb{D}_k(\delta_x,\delta_y) \to 0$ quand $x \to y$, une propriété qui est partagée avec la distance de Wasserstein mais qui n'est valable ni pour la distance de Hellinger ni pour la distance TV.

Robustesse à la contamination adversariale:

L'estimation basée sur la distance MMD peut également être utilisée dans le cas particulier de l'estimation paramétrique robuste avec contamination adversariale, où la distribution cible $P_{\boldsymbol{\theta}_0}$ appartient au modèle mais une fraction ε des données est contaminée par un adversaire. Dans ce cadre, nous obtenons le résultat suivant :

Theorem 3.3.2 (Informel)**.** *Sous les mêmes hypothèses que pour le Théorème 2.3.1, nous avons pour tout $\delta \in (0,1)$, avec une probabilité d'au moins $1-\delta$:*

$$\mathbb{D}_k\left(P_{\widehat{\boldsymbol{\theta}}_n}, P_{\boldsymbol{\theta}_0}\right) \leq 4\left(\varepsilon + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right).$$

Là encore, il est possible d'obtenir un résultat plus faible en espérance en assouplissant l'hypothèse impliquant Γ. Notons que ce résultat est toujours valable pour le paramètre de contamination de Hüber, mais la constante 4 est remplacée par la constance plus petite 2.

La vitesse obtenu par l'estimateur MMD (en MMD) est $\max(1/\sqrt{n}, \varepsilon)$, et a la même forme que la vitesse minimax pour l'estimation de la moyenne d'une distribution Gaussienne. Nous retrouvons la vitesse de convergence optimale par rapport à n sans contam-

ination lorsque $\varepsilon \lesssim 1/\sqrt{n}$, alors que la vitesse est dominé par le ratio de contamination ε sinon. Ainsi, le nombre maximum de valeurs aberrantes tolérées sans affecter la vitesse sans outliers est de $\sqrt{n}$, et est indépendant de la dimension. Dans le cas particulier de l'estimation d'une moyenne Gaussienne avec la matrice de covariance $\sigma^2 I_d$, nous obtenons via le noyau Gaussien la même vitesse que la médiane coordonnées par coordonnées, i. e. $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = \mathcal{O}(\max(d^{1/2}n^{-1/2}, d^{1/2}\varepsilon))$, tandis que la vitesse minimax est $\max(d^{1/2}n^{-1/2}, \varepsilon)$. À titre de comparaison, la méthodologie Median-of-Means conduit à une estimation en $\mathcal{O}(\max(d^{1/2}n^{-1/2}, \varepsilon^{1/2}))$, c'est-à-dire un nombre maximum de valeurs aberrantes qui est toléré sans affecter la vitesse sans outliers est d'ordre d. Nous pensons que nos vitesse obtenues à l'aide de l'estimateur MMD peuvent être améliorés par un choix approprié du noyau.

Considérations computationnelles:

L'estimateur $\widehat{\boldsymbol{\theta}}_n$ peut être calculé à l'aide d'un algorithme de gradient lorsque $\boldsymbol{\Theta} \subset \mathbb{R}^d$ pour un modèle génératif, c'est-à-dire lorsqu'il est possible de simuler à partir de n'importe quel $P_{\boldsymbol{\theta}}$. L'idée d'exploiter la descente de gradient stochastique pour calculer $\widehat{\boldsymbol{\theta}}_n$ remonte à Dziugaite et al. (2015) qui a utilisé la SGD pour un réseau de neurones génératif, et a été à nouveau discutée dans Briol et al. (2019). L'algorithme est basé sur une approximation en U-statistique du critère MMD et est détaillé dans l'Algorithme 12. Nous fournissons également une analyse théorique de l'algorithme et des simulations numériques, où nous testons l'estimation robuste d'une uni- et multidimensionnelle Gaussienne univariée, une uniforme, une Cauchy, et un mélange Gaussien.

3.3.3 Un estimateur Bayésien

Comme déjà expliqué dans Section 1.3, l'approche Bayésienne n'est pas robuste à la mauvaise spécification, et la posterior n'est pas consistente dans de nombreuses situations (Barron et al., 1999; Grünwald et al., 2017). Nous proposons à nouveau dans Chérief-Abdellatif and Alquier (2020) d'utiliser la distance MMD pour concevoir un estimateur Bayésien robuste.

Nous soutenons dans le Chapitre 9 que le choix de la distance MMD est plus adapté pour effectuer une estimation robuste que la divergence KL. Pour justifier notre affirmation, nous montrons que dans le modèle de contamination de Hüber pour l'estimation d'une moyenne Gaussienne corrompue par une autre distribution Gaussienne, nous retrouvons exactement la vraie moyenne lorsque nous utilisons le minimiseur en MMD par rapport au vrai mélange, alors que nous n'y parvenons pas avec la divergence KL. Ainsi, suivant l'idée centrale de la théorie PAC-Bayésienne, nous remplaçons la log-vraisemblance ℓ_n par la MMD dans la formule de Bayes, et nous appelons cette pseudo-posterior MMD-Bayes:

$$\Pi_{n,\alpha}(d\boldsymbol{\theta}) \propto \exp\left(-\alpha \cdot \mathbb{D}_k^2(P_{\boldsymbol{\theta}}, \widehat{P}_n)\right)\Pi_0(d\boldsymbol{\theta}).$$

Dans le Chapitre 9, nous montrons que la MMD-Bayes se concentre en la vraie distribution lorsque le modèle est bien spécifié sous une version de la condition de prior

mass adaptée à l'estimation MMD. Lorsque la métrique et le rayon des voisinages MMD dans la condition de prior mass sont respectivement la métrique MMD et $n^{-1/2}$, alors la MMD-Bayes se concentre à vitesse optimale $n^{-1/2}$ en MMD (Tolstikhin et al., 2017).

En outre, lorsque le modèle est mal spécifié, il est encore possible d'obtenir une inégalité oracle pour la pseudo-posterior lorsque la condition de prior mass est basée sur un voisinage d'une approximation $P_{\boldsymbol{\theta}^*} = \arg\min_{P_{\boldsymbol{\theta}}} \mathbb{D}_k\left(P_{\boldsymbol{\theta}}, P_0\right)$ de la vraie distribution au lieu de la vraie distribution elle-même :

Theorem 3.3.3 (Informel)**.** *Sous la condition de prior mass appliquée aux voisinage de la meilleure approximation, nous avons pour tout $\alpha \in (0,1)$:*

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_{\boldsymbol{\theta}}, P_0\right) \Pi_{n,\alpha}(\mathrm{d}\boldsymbol{\theta})\right] \leq 8 \inf_{\boldsymbol{\Theta}} \mathbb{D}_k^2\left(P_{\boldsymbol{\theta}}, P_0\right) + \frac{16}{n}.$$

Le paramètre α n'apparaît pas dans le terme de droite de l'inégalité, ce phénomène sera expliqué en détails au Chapitre 9 et a déjà été rencontré par exemple dans Dalalyan et al. (2018). Nous fournissons également un exemple de calcul d'une telle masse de la prior dans ce chapitre.

Pour surmonter l'intractibilité de la MMD-Bayes dans les modèles complexes, nous utilisons l'inférence variationnelle dans Chérief-Abdellatif and Alquier (2020). Nous montrons que son approximation variationnelle conserve les mêmes propriétés théoriques dans une condition de prior mass étendue, et nous étayons nos conclusions théoriques avec des simulations numériques en utilisant un algorithme SGD comme pour l'estimation par minimisation de distance basée sur la MMD.

Dans le Chapitre 8, nous donnons un moyen simple de définir des procédures d'estimation universelles via la métrique MMD. En particulier:

- ***Nous donnons des inégalités oracle qui impliquent une estimation robuste sous l'hypothèse i.i.d dans les modèles de Hüber et de contamination adversariale.***
- ***Nous allons au-delà de l'hypothèse i.i.d. classique en introduisant un nouveau coefficient de dépendance faible simple et exprimé sous forme de covariance dans le RKHS, nous montrons que l'estimateur MMD est robuste à la dépendance entre les observations.***
- ***Nous relions également notre estimateur MMD à l'estimation par minimisation de distance en utilisant métrique L_2.***
- ***Nous donnons une analyse théorique d'un algorithme de descente de gradient stochastique utilisé pour calculer cet estimateur pour les modèles fini-dimensionnels, que nous justifions empiriquement.***

Ce travail a fait l'objet d'un article actuellement soumis à Bernoulli (Chérief-Abdellatif and Alquier, 2019).

Dans le Chapitre 9, nous donnons une version Bayésienne de l'estimateur MMD qui est consistente avec des propriétés optimales dans le cas bien spécifié, et qui est robuste dans le cas contraire. Ce travail dans ce chapitre a fait l'objet d'une publication (Chérief-Abdellatif and Alquier, 2020).

Part II

Consistency of variational inference for estimation and model selection

Chapter 4

Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures

Mixture models are widely used in Bayesian statistics and machine learning and proved their efficiency in many fields such as computational biology or natural language processing... Variational inference, a technique for approximating intractable posteriors thanks to optimization algorithms, is extremely popular in practice when dealing with complex models such as mixtures. The contribution of this chapter is two-fold. First, we study the concentration of variational approximations of posteriors, which is still an open problem for general mixtures, and we derive consistency and rates of convergence. We also tackle the problem of model selection for the number of components: we study the approach already used in practice, which consists in maximizing a numerical criterion (ELBO). We prove that this strategy indeed leads to strong oracle inequalities. We illustrate our theoretical results by applications to Gaussian and multinomial mixtures.

4.1 Introduction

This chapter studies the statistical properties of variational inference as a tool to tackle two problems of interest: estimation and model selection in mixture models. Mixtures are often used for modelling population heterogeneity, leading to practical clustering methods (Bouveyron and Brunet-Saumard, 2014; McNicholas, 2016). Moreover they have enough flexibility to approximate accurately almost every density (Bacharoglou, 2010; Kruijer et al., 2010). Mixtures are used in many various areas such as computer vision (Ayer and Sawhney, 1995), genetics (Pan et al., 2003), economics (Deb et al., 2011), transport data analysis (Carel and Alquier, 2017)... We refer the reader to Celeux et al. (2018) for an account of the recent advances on mixtures. The most famous procedure for mixture density estimation in the frequentist literature is probably Expectation-Maximization (Dempster et al., 1977), a maximum-likelihood algorithm that yields increasingly higher likelihood. At the same time, the Bayesian paradigm has raised great interest among researchers and practitioners, especially through the Variational Bayes (VB) framework which aims at

maximizing a quantity referred to as Evidence Lower Bound on the marginal likelihood (ELBO). Variational Bayes inference is a privileged tool for approximating intractable posteriors. It is known to work well in practice for mixture models: one of the most recent survey on VB (Blei et al., 2017) chooses mixtures as an example of choice to illustrate the power of the method. Moreover, Blei et al. (2017) states: "the [evidence lower] bound is a good approximation of the marginal likelihood, which provides a basis for selecting a model. Though this sometimes works in practice, selecting based on a bound is not justified in theory". The main contribution of this chapter is to prove that VB is consistent for estimation in mixture models, and that the ELBO maximization strategy used in practice is consistent for model selection. Thus we solve the question raised by Blei et al. (2017).

Variational Bayes is a method for computing intractable posteriors in Bayesian statistics and machine learning. Markov Chain Monte Carlo (MCMC) algorithms remain the most widely used methods in computational Bayesian statistics. Nevertheless, they are often too slow for practical uses when the dataset is very large. A more and more popular alternative consists in finding a deterministic approximation of the target distribution called Variational Bayes approximation. The idea is to minimize the Kullback-Leibler divergence of a tractable distribution ρ with respect to the posterior, which is also equivalent to maximizing the ELBO. This optimization procedure is much faster than MCMC sampling and proved its efficiency in many different fields: matrix completion for collaborative filtering (Cottet and Alquier, 2018), computer vision (Sudderth and Jordan, 2009), computational biology (Carbonetto and Stephens, 2012) and natural language processing (Hoffman et al., 2013), to name a few prominent examples.

However, variational inference is mainly used for its practical efficiency and only little attention has been put in the literature towards theoretical properties of the VB approximation until very recently. In Alquier et al. (2016) the properties of variational approximations of Gibbs distributions used in machine learning are derived. The results are essentially valid for bounded loss functions, which makes them difficult to use beyond the problem of supervised classification. Based on some technical advances from Bhattacharya et al. (2016), Alquier and Ridgway (2017) removed the boundedness assumption in Alquier et al. (2016), allowing to study more general statistical models. In Bhattacharya et al. (2018), the authors extended the range of models covered by Alquier et al. (2016). They even studied mixture of Gaussian distributions as a short example. Many questions are still left unanswered: model selection, and the estimation of mixture of non-Gaussian distributions. For example mixture of multinomials are widely used in practice (Carel and Alquier, 2017), as well as more intricated examples such as nonparametric mixtures (Gassiat et al., 2018). Note that all the results in Bhattacharya et al. (2016); Alquier and Ridgway (2017); Bhattacharya et al. (2018) are limited to so-called tempered posteriors, that is, where the likelihood is taken to some power α. Still, the use of tempered posteriors is highly recommended by many authors as a way to overcome model misspecification, see Grünwald et al. (2017) and the references therein. Indeed some results in Alquier and Ridgway (2017) are valid in a misspecified setting. Note that alternative approaches were developed to study VB: Wang and Blei (2018) established Bernstein-von-Mises type theorems on the variational approximation of the posterior. They provide very interesting results for parametric models but it is unclear whether

these results can be extended to model selection or misspecified case. More recently, Zhang and Gao (2017) succeeded in adapting the now classical results of Ghosal et al. (2000) to Variational Bayes and showed that a slight modification in the three classical "prior mass and testing conditions" leads to the convergence of their variational approximations, again under the assumption that the model is true. The first contribution of this chapter is to study the statistical properties of VB for general mixture models, both in the well-specified and misspecified setting.

The other point addressed in this chapter is model selection. This is a natural question which can be interpreted in this context as the determination of the number of components of the approximating mixture. This point is crucial: indeed, too many components can lead to estimates with too large variances whereas with too few components, we may obtain mixtures which are not able to fit the data properly. This is a common issue and a lot of statisticians worked on this question. In the literature, criteria such as AIC (Akaike, 1974) and BIC (Schwarz, 1978) are popular. It is well known that in some collections of models, AIC optimizes the prediction ability while BIC recovers with high probability the true model (when there is one). These two objectives are not compatible in general (Yang, 2005). Anyway, these results depend on asssumptions that are not satisfied by mixtures. It seems thus more natural to develop criteria suited to a given objective. For example, Biernacki et al. (1999) proposed a procedure to select a number of components that is the most relevant for clustering. A non-asymptotic theory of penalization has been developed during the last two decades using oracle inequalities (Massart, 2007). In the wake of those works, this chapter studies mixture model selection based on the ELBO criterion. We prove a general oracle inequality. This result establishes the consistency of ELBO maximization when the primary objective is the estimation of the distribution of the data.

The rest of this chapter is organized as follows. In Section 4.2 we introduce the background and the notations that will be adopted. Consistency of the Variational Bayes for estimation in a mixture model is studied in Section 4.3. First, we give the general results under a "prior mass" assumption, as well as a general form for the algorithm to compute the VB approximation (Subsection 4.3.1). We then apply these results to mixtures of multinomials (Subsection 4.3.2) and Gaussian mixtures (Subsection 4.3.3). In each case, we provide a rate of convergence of VB and discuss its numerical implementation. We extend the setting to the misspecified case in Subsection 4.3.4. Finally, we address the issue of selecting based on the ELBO in Section 4.4. Section 4.6 is dedicated to the proofs.

4.2 Background and notations

Let us precise the notations and the framework we adopt in this chapter . We observe in a measurable space $\left(\mathbb{X}, \mathcal{X}\right)$ a collection of n i.i.d. random variables X_1,...,X_n sampled from a probability distribution which density with respect to some dominating measure μ is denoted by P^0. We put $(X_1, ..., X_n) = X_1^n$. The goal is to estimate the generating distribution P^0 of the X_i's by a K-components mixture model. We will study the (frequentist) properties of variational approximations of the posterior. The extension to

selection of the number of components is also tackled in this chapter , but in a first time we deal with a fixed K. We introduce a collection of distributions $\{Q_\theta/\theta \in \Theta\}$ indexed by a parameter space Θ from which we will take the different components of our mixture model. We assume that for each $\theta \in \Theta$, the probability distribution Q_θ is dominated by the reference measure μ and that the density $q_\theta = \frac{dQ_\theta}{d\mu}$ is such that the map $(x, \theta) \to q_\theta(x)$ is $\mathcal{X} \times \mathcal{T}$-measurable, $\mathcal{T}$ being some sigma-algebra on Θ. Unless explicitly stated otherwise, all the distributions that will be considered in this chapter will be characterized by their density with respect to the dominating measure μ. We can now consider the statistical mixture model of $K \geq 1$ components defined as:

$$\left\{ \sum_{j=1}^{K} p_j q_{\theta_j} \ \Big/ \ \theta_j \in \Theta \text{ for } j = 1, ..., K, \ p = (p_1, ..., p_K) \in \mathcal{S}_K \right\}$$

where $\mathcal{S}_K = \left\{ p = (p_1, ..., p_K) \in \mathbb{R}^K \Big/ \ p_j \geq 0 \text{ for } j = 1, ..., K \text{ and } \sum_{j=1}^{K} p_j = 1 \right\}$ is the $K-1$ dimensional simplex. We will write $\theta = (p_1, ..., p_K, \theta_1, ...\theta_K) \in \Theta_K$ for short, where $p \in \mathcal{S}_K$, $\theta_j \in \Theta$ for $j = 1, ..., K$ and $\Theta_K = \mathcal{S}_K \times \Theta^K$. The mixture corresponding to parameter $\theta = (p_1, ..., p_K, \theta_1, ..., \theta_K)$ will be denoted $P_\theta = \sum_{j=1}^{K} p_j q_{\theta_j}$.

First, we consider the well-specified case, assuming that the true distribution belongs to the K-components mixture model. Thus, we define the true distribution P^0 from which data are sampled:

$$X_1, ..., X_n \sim \sum_{j=1}^{K} p_j^0 q_{\theta_j^0} \text{ with } \theta_j^0 \in \Theta \text{ for } j \in 1, ..., K \text{ and } p^0 \in \mathcal{S}_K.$$

Hence, we want to estimate the true distribution P_{θ^0} using a Bayesian approach. Therefore, we define a prior $\pi = \pi_p \bigotimes_{j=1}^{K} \pi_j$ on θ, $\pi_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ being a probability distribution on some measurable space $(\mathcal{S}_K, \mathcal{A})$, and each $\pi_j \in \mathcal{M}_1^+(\Theta)$ a probability distribution on the measurable space $(\Theta, \mathcal{T})$. We will also consider in this chapter the misspecified case where the true distribution does not belong to our statistical model i.e. is not necessarily a mixture, but the specific notations and framework will be precised later.

Let us remind some notations. The likelihood will be denoted by L_n and the log-likelihood by ℓ_n, that is, for any $\theta = (p_1, ..., p_K, \theta_1, ...\theta_K)$,

$$L_n(\theta) = \prod_{i=1}^{n} \sum_{j=1}^{K} p_j q_{\theta_j} \ , \ \ \ell_n(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} p_j q_{\theta_j} \right).$$

The negative log-likelihood ratio r_n between two distributions P and R is given by

$$r_n(P, R) = \sum_{i=1}^{n} \log \left(\frac{R(X_i)}{P(X_i)} \right)$$

(note that $r_n(\theta, \theta')$ is used by many authors instead of $r_n(P_\theta, P_{\theta'})$ but our notation is more convenient for the extension to the misspecified case). The Kullback-Leibler (KL) divergence between two probability distributions P and R is given by

$$\mathcal{K}(P, R) = \begin{cases} \int \log \left(\frac{dP}{dR} \right) dP \text{ if } R \text{ dominates } P, \\ +\infty \text{ otherwise.} \end{cases}$$

If some measure λ dominates both P and R distributions represented here by their densities f and g with respect to this measure, we have

$$\mathcal{K}(P,R) = \int f \log\left(\frac{f}{g}\right) d\lambda$$

and we will use $K(P,R)$ or $\mathcal{K}(f,g)$ to denote this quantity, depending on the context.

We also remind that the α-Renyi divergence between P and R,

$$D_\alpha(P,R) = \begin{cases} \frac{1}{\alpha-1}\log\int\left(\frac{dP}{dR}\right)^{\alpha-1} dP \text{ if } R \text{ dominates } P, \\ +\infty \text{ otherwise.} \end{cases}$$

When for some λ we have $P = f.\lambda$ and $R = g.\lambda$,

$$D_\alpha(P,R) = D_\alpha(f,g) = \frac{1}{\alpha-1}\log\int f^\alpha g^{1-\alpha} d\lambda.$$

Some useful properties of Renyi divergences can be found in Van Erven and Harremos (2014). In particular, the Renyi divergence between two probability distributions P and R can be related to the classical total variation TV and Hellinger H distances respectively defined as $TV(P,R) = \frac{1}{2}\int|dP - dR|$ and $H(P,R)^2 = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dR})^2 = 1 - e^{-\frac{1}{2}D_{1/2}(P,R)}$ through:

$$TV(P,R)^2 \leq 2H(P,R)^2 \leq D_{1/2}(P,R) \text{ and } D_\alpha(P,R) \xrightarrow[\alpha\to 1]{\nearrow} \mathcal{K}(P,R).$$

The tempered Bayesian posterior $\pi_{n,\alpha}(.|X_1^n)$, which is our target here, is defined for $0 < \alpha \leq 1$ by

$$\pi_{n,\alpha}(d\theta|X_1^n) = \frac{e^{\alpha r_n(P_\theta,P^0)}\pi(d\theta)}{\int e^{\alpha r_n(P_\phi,P^0)}\pi(d\phi)} \propto L_n(\theta)^\alpha \pi(d\theta)$$

(it is also referred to as fractional posterior, for example in Bhattacharya et al. (2016)). Note that when $\alpha = 1$, then we recover the "true" Bayesian posterior, but the case $\alpha < 1$ has many advantages: it is often more tractable from a computational perspective (Neal, 1996; Behrens et al., 2012), it is consistent under less stringent assumptions than required for $\alpha = 1$ (Bhattacharya et al., 2016) and it is more robust to misspecification (Grünwald et al., 2017).

We are now in position to define the VB approximation $\widetilde{\pi}_{n,\alpha}(.|X_1^n)$ of the tempered posterior with respect to some set of distributions $\mathcal{F}$: it is the projection, with respect to the Kullback-Leibler divergence, of the tempered posterior onto the mean-field variational set $\mathcal{F}$,

$$\widetilde{\pi}_{n,\alpha}(.|X_1^n) = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}\Big(\rho, \pi_{n,\alpha}(.|X_1^n)\Big).$$

The mean-field approximation is very popular in the Variational Bayes literature. It is based on a decomposition of the space of parameters Θ_K as a product. Then $\mathcal{F}$ consists in compatible product distributions. Here, a natural choice (Blei et al., 2017) is $\Theta_K = \mathcal{S}_K \times \Theta \times \cdots \times \Theta$ and

$$\mathcal{F} = \left\{\rho_p \bigotimes_{j=1}^K \rho_j/\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K),\ \rho_j \in \mathcal{M}_1^+(\Theta)\ \forall j = 1,...,K\right\}.$$

We will work on this particular set in the following and we will often use ρ instead of $\rho_p \bigotimes_{j=1}^K \rho_j$ or $(\rho_p, \rho_1, ..., \rho_K)$ to ease notation.

We end this section by recalling Donsker and Varadhan's variational formula. Refer for example to Catoni (2007) for a proof (Lemma 1.1.3).

Lemma 4.2.1. *For any probability λ on some measurable space $(\boldsymbol{E}, \mathcal{E})$ and any measurable function $h : \boldsymbol{E} \to \mathbb{R}$ such that $\int \mathrm{e}^h \mathrm{d}\lambda < \infty$,*

$$\log \int \mathrm{e}^h \mathrm{d}\lambda = \sup_{\rho \in \mathcal{M}_1^+(\boldsymbol{E})} \left\{ \int h \mathrm{d}\rho - \mathcal{K}(\rho, \lambda) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, if h is upper-bounded on the support of λ, then the supremum on the right-hand side is reached by the distribution of the form:

$$\lambda_h(d\beta) = \frac{\mathrm{e}^{h(\beta)}}{\int \mathrm{e}^h \mathrm{d}\beta} \lambda(d\beta).$$

This technical lemma is one of the main ingredients for the proof of our results, but it is also very helpful to understand variational approximations. Indeed, for $\mathbf{E} = \Theta_K$ and using the definition of $\pi_{n,\alpha}(.|X_1^n)$, we get:

$$\pi_{n,\alpha}(\cdot|X_1^n) = \underset{\rho \in \mathcal{M}_+^1(\Theta_K)}{\arg\min} \left\{ \alpha \int r_n(P_\theta, P^0) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}$$

and simple calculations give

$$\begin{aligned}
\widetilde{\pi}_{n,\alpha}(\cdot|X_1^n) &= \underset{\rho \in \mathcal{F}}{\arg\min} \left\{ \alpha \int r_n(P_\theta, P^0) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\} \\
&= \arg\max_{\rho \in \mathcal{F}} \left\{ \alpha \int \ell_n(\theta) \rho(d\theta) - \mathcal{K}\Big(\rho, \pi\Big) \right\} \qquad (4.1) \\
&= \arg\min_{\rho \in \mathcal{F}} \left\{ -\alpha \sum_{i=1}^n \int \log \left(\sum_{j=1}^K p_j q_{\theta_j}(X_i) \right) \rho(d\theta) + \mathcal{K}\Big(\rho_p, \pi_p\Big) + \sum_{j=1}^K \mathcal{K}\Big(\rho_j, \pi_j\Big) \right\}. \qquad (4.2)
\end{aligned}$$

The quantity maximized in (4.1) is called the ELBO in the litterature (ELBO stands for Evidence Lower Bound), and many authors actually take this as the definition of VB (Blei et al., 2017).

4.3 Variational Bayes estimation of a mixture

4.3.1 A PAC-Bayesian inequality

We start with a result for general mixtures. Later in this section we provide corollaries obtained by applying this theorem to special cases: mixture of multinomials and Gaussian mixtures.

Theorem 4.3.1. *For any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0}\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right]$$
$$\leq \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\left[\int \mathcal{K}(p^0,p)\rho_p(dp) + \sum_{j=1}^K \int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})\rho_j(d\theta_j)\right] + \frac{\mathcal{K}(\rho_p,\pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_j,\pi_j)}{n(1-\alpha)}\right\}.$$

As a special case, when there exists $r_{n,K}$ such that there are distributions $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\rho_{j,n} \in \mathcal{M}_1^+(\Theta)$ $(j = 1,...,K)$ such that for $j = 1,...,K$

$$\int \mathcal{K}(p^0,p)\rho_{p,n}(dp) \leq Kr_{n,K}, \quad \int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})\rho_{j,n}(d\theta_j) \leq r_{n,K} \tag{4.3}$$

and

$$\mathcal{K}(\rho_{p,n},\pi_p) \leq Knr_{n,K}, \quad \mathcal{K}(\rho_{j,n},\pi_j) \leq nr_{n,K}, \tag{4.4}$$

then for any $\alpha \in (0,1)$

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0}\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha} 2Kr_{n,K}.$$

The proof is given in Section 4.6. This theorem provides the consistency of the Variational Bayes for mixture models as soon as (4.3) and (4.4) are satisfied. In Alquier and Ridgway (2017), the authors use similar conditions ((3) and (4) in their Theorem 2.6), and show that they are strongly linked to the assumptions on the prior used by Ghosal et al. (2000); Bhattacharya et al. (2016) to derive concentration of the posterior. Thus they cannot be removed in general. Theorem 4.3.1 states that finding $r_{n,K}$ filling (4.3) and (4.4) independently for the weights and for each component is sufficient to obtain the rate of convergence $Kr_{n,K}$ of the VB estimator towards the true distribution.

Note that there always exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that the two quantities corresponding to the weights $\int \mathcal{K}(p^0,p)\rho_{p,n}(dp)$ and $\mathcal{K}(\rho_{p,n},\pi_p)$ are bounded as required in Theorem 4.3.1 for $r_{n,K} = \frac{4\log(nK)}{n}$ when the chosen prior is a Dirichlet distribution $\pi_p = \mathcal{D}_K(\alpha_1,...,\alpha_K)$ under some minor restriction on $\alpha_1,...,\alpha_K$. This result summarized below for any $K \geq 2$ help find explicit rates of convergence for the VB approximation.

Lemma 4.3.2. *For $r_{n,K} = \frac{4\log(nK)}{n}$ and a prior $\pi_p = \mathcal{D}_K(\alpha_1,...,\alpha_K) \in \mathcal{M}_1^+(\mathcal{S}_K)$ with $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1,...,K$, we can find a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that*

$$\int \mathcal{K}(p^0,p)\rho_{p,n}(dp) \leq Kr_{n,K}$$

and

$$\mathcal{K}(\rho_{p,n},\pi_p) \leq Knr_{n,K}.$$

Thus, conditions among (4.3) and (4.4) applying on the components of the mixtures are always sufficient for guarantying consistency of the Variational Bayes in mixtures and obtaining its rate of convergence.

Remark 4.3.1. *When $K = 1$, Lemma 4.3.2 does not apply as $\frac{2}{K} > 1$. Nevertheless, as there is only one component, then any $p \in \mathcal{S}_K$ is equal to 1 and the two conditions are immediately satisfied for any prior π_p and any rate $r_{n,K}$ with $\rho_{p,n} = \pi_p$.*

The central idea of the proof of Lemma 4.3.2 (given in details in Section 4.6) is to consider the ball $\mathcal{B}$ centered at p^0 of radius $Kr_{n,K}$ defined as:

$$\mathcal{B} = \left\{ p \in \mathcal{S}_K / \ \mathcal{K}(p^0, p) \leq Kr_{n,K} \right\}.$$

Hence, when considering the restriction $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ of π to $\mathcal{B}$, condition (4.3) is trivially satisfied and condition (4.4) is restricted to

$$\pi_p(\mathcal{B}) \geq e^{-nKr_{n,K}}.$$

This is a very classical assumption stated in many papers to study the concentration of the posterior (Ghosal et al., 2000; Alquier and Ridgway, 2017; Zhang and Gao, 2017). However, the computation of such a prior mass $\pi_p(\mathcal{B})$ is a major difficulty. Lemma 6.1 in Ghosal et al. (2000) treated the case of L_1-balls for Dirichlet priors. Since then, only a few papers in the literature addressed this issue. Our result extends the work in Ghosal et al. (2000) to KL-balls, which is of great interest in our study. Moreover, the range of Dirichlet priors for which Lemma 4.3.2 is available is the same as the one in Ghosal et al. (2000).

We conclude Subsection 4.3.1 by a short discussion on the implementation of the VB approximation. Indeed, VB methods are meant to be practical objects, so there would be no point in proving the consistency of a VB approximation that would not be computable in practice. Many algorithms have been studied in the literature, with good performances – see Blei et al. (2017) and the references therein. In the case of mean-field approximation, the most popular method is to optimize iteratively with respect to all the independent components. Here this might seem difficult: it is indeed as difficult as maximizing the likelihood of a mixture. But a trick widely used in practice (see for example Section 7 in Hershey and Olsen (2007)) is to use the equality

$$\text{for any } i = 1, ..., K \qquad -\log\left(\sum_{j=1}^{K} p_j q_{\theta_j}(X_i)\right) = \min_{\omega^i \in \mathcal{S}_K} \left\{ -\sum_{j=1}^{K} \omega_j^i \log(p_j q_{\theta_j}(X_i)) + \sum_{j=1}^{K} \omega_j^i \log(\omega_j^i) \right\}.$$

This equality is once again a consequence of Lemma 4.2.1 (take $\mathbf{E} = \{1, ..., K\}$, $\lambda = (1/K, ..., 1/K)$ and $h(j) = \log(p_j q_{\theta_j}(X_i))$). This leads to the program:

$$\min_{\rho \in \mathcal{F},\ w \in \mathcal{S}_K^n} \left\{ -\alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \left(\int \log(p_j) \rho_p(dp) + \int \log(q_{\theta_j}(X_i)) \rho_j(d\theta_j) \right) \right.$$
$$\left. + \alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \log(\omega_j^i) + \mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^{K} \mathcal{K}(\rho_j, \pi_j) \right\}.$$

This version can be solved by coordinate descent, see Algorithm 1. Update formulas once again follow from Lemma 4.2.1 (for instance, line 7 can be obtained with $\mathbf{E} = \{1, ..., K\}$,

$\lambda = (1/K, ..., 1/K)$ and $h(j) = \int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j)$, more details are provided in Section 4.6). This algorithm is, in the case $\alpha = 1$, exactly equivalent to the popular CAVI algorithm (Bhattacharya et al., 2018; Blei et al., 2017; Hoffman et al., 2013), where the ω_j^i's are interpreted as the posterior means of the latent variables Z_j^i's. A very short numerical study is provided in the Supplementary Material but note that CAVI has already been extensively tested in practice (Blei et al., 2017).

Algorithm 6 Coordinate Descent Variational Bayes for mixtures

Require: A dataset $(X_1, ..., X_n)$, priors π_p,$\{\pi_j\}_{j=1}^K$ and a family $\{q_\theta/\theta \in \Theta\}$, initial variational factors ρ_p, $\{\rho_j\}_{j=1}^K$.

repeat

for $i = 1, ..., n$ **do**

for $j = 1, ..., K$ **do**

set $w_j^i = \exp\left(\int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j)\right)$

end for

normalize $(w_j^i)_{1\leq j\leq K}$

end for

set $\rho_p(dp) \propto \exp\left(\alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \log(p_j)\right)\pi_p(dp)$

for $j = 1, ..., K$ **do**

set $\rho_j(d\theta_j) \propto \exp\left(\alpha \sum_{i=1}^n \omega_j^i \log(q_{\theta_j}(X_i))\right)\pi_j(d\theta_j)$

end for

until convergence of the objective function

4.3.2 Application to multinomial mixture models

We present in this section an application to the multinomial mixture model frequently used for text clustering (Rigouste et al., 2007), transport schedule analysis (Carel and Alquier, 2017)... The parameter space is the $V-1$ dimensional simplex $\Theta = \mathcal{S}_V$ with $V \in \mathbb{N}^*$. We choose conjugate Dirichlet priors as in Rigouste et al. (2007) $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ and $\pi_j = \mathcal{D}_V(\beta_1, ..., \beta_V)$ with $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$ and $\frac{2}{V} \leq \beta_\ell \leq 1$ for $\ell = 1, ..., V$.

The following corollary of Theorem 4.3.1 states that convergence of the VB approximation for the multinomial mixture model is achieved at rate $\frac{KV\log(nV)}{n}$ as soon as $V^V \geq K$, which is the case in many text mining models such as Latent Dirichlet Allocation (Blei et al., 2003) for which the size of the vocabulary is very large:

Corollary 4.3.3. *For any* $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0}\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}\left[\frac{8KV\log(nV)}{n} \bigvee \frac{8K\log(nK)}{n}\right].$$

The proof is in Section 4.6. We also explicit Algorithm 1 in this setting: see Algorithm 2. There ψ denotes the Digamma function, $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log[\Gamma(x)]$ where Γ stands for the Gamma function $\Gamma(x) = \int_0^\infty \exp(-t)t^{x-1}\mathrm{d}t$.

Algorithm 7 Coordinate Descent Variational Bayes for multinomial mixtures

Require: Initial variational parameters $(\phi_1, ..., \phi_K) \in \mathbb{R}_+^K$, $(\gamma_{1j}, ..., \gamma_{Vj}) \in \mathbb{R}_+^V$ and corresponding variational distributions $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$, $\rho_j = \mathcal{D}_V(\gamma_{1j}, ..., \gamma_{Vj})$ for $j = 1, ..., K$.

repeat

  for $i = 1, ..., n$ **do**

    for $j = 1, ..., K$ **do**

      set $w_j^i = \exp\left(\psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell) + \psi(\gamma_{X_i,j}) - \psi\left(\sum_{v=1}^{V} \gamma_{vj}\right)\right)$

    end for

    normalize $(w_j^i)_{1 \leq j \leq K}$

  end for

  set $\phi_j = \alpha_j + \alpha \sum_{i=1}^{n} \omega_j^i$ for $j = 1, ..., K$

  set $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$

  for $j = 1, ..., K$ **do**

    set $\gamma_{vj} = \beta_v + \alpha \sum_{i=1}^{n} \omega_j^i \mathbb{1}(X_i = v)$ for $v = 1, ..., V$

    set $\rho_j = \mathcal{D}_V(\gamma_{1j}, ..., \gamma_{Vj})$

  end for

until convergence of the objective function

4.3.3 Application to Gaussian mixture models

Let us now address the case of the Gaussian mixture model. This is one of the most popular mixture models for many applications including model based clustering (Bouveyron and Brunet-Saumard, 2014; McNicholas, 2016) and VB approximations have been studied in depth for this model (Nasios and Bors, 2006). First, we will explicit rates of convergence of the VB approximation of the tempered posterior when the variance is known, and then when the variance is unknown.

First, we consider mixtures of V^2-variance Gaussians. The mean parameter space is $\Theta = \mathbb{R}$. We select priors $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ and $\pi_j = \mathcal{N}(0, \mathcal{V}^2)$ with $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$ and $\mathcal{V}^2 > 0$. The following result gives a rate of convergence $Kr_{n,K}$ of the VB approximation:

Corollary 4.3.4. *Let us define* $r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2}\right]$. *Then, for any* $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^{K} p_j q_{\theta_j}, \sum_{j=1}^{K} p_j^0 q_{\theta_j^0}\right) \widetilde{\pi}_{n,\alpha}(d\theta | X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha} 2Kr_{n,K}.$$

One can see that for n large enough, the consistency rate is $\frac{K\log(nK)}{n}$, which comes from the estimation of the weights of the mixture.

We can also provide a similar result when the variance of each component is unknown.

The consistency rate remains the same, and is entirely characterized by the weights consistency rate. The parameter space is now $\Theta = \mathbb{R} \times \mathbb{R}_+^*$. We consider again a Dirichlet prior $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ with $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$ on $p \in \mathcal{S}_K$, and we will provide our results for two different priors π_j frequently used in the literature: a Normal-Inverse-Gamma prior (Stoneking, 2014) and a factorized prior (Watier et al., 1999).

Corollary 4.3.5. *Let us fix $\alpha \in (0, 1)$.*

- *For a Normal-Inverse-Gamma prior $\pi_j = \mathcal{NIG}(0, \mathcal{V}^2, 1, \gamma^2)$ for each $j = 1, ..., K$. With*
$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^K \frac{1}{n} \left[\frac{1}{2} \log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - (\sigma_j^0)^2}{(\sigma_j^0)^2} \right]$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0} \right) \widetilde{\pi}_{n,\alpha}(d\theta | X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

- *For the factorized prior $\pi_j = \mathcal{N}(0, \mathcal{V}^2) \bigotimes \mathcal{IG}(1, \gamma^2)$ for each $j = 1, ..., K$. With*
$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^K \frac{1}{2(\sigma_j^0)^2 n} \bigvee_{j=1}^K \frac{1}{n} \left[\frac{1}{2} \log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - (\sigma_j^0)^2}{(\sigma_j^0)^2} \right]$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0} \right) \widetilde{\pi}_{n,\alpha}(d\theta | X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

One can see that even when the variance has to be estimated, the consistency rate still achieves $\frac{K\log(nK)}{n}$ for n large enough, whatever the form of the prior - factorized or not.

We give in Algorithm 3 a version of Algorithm 1 for unit-variance Gaussian mixtures with priors $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ and $\pi_j = \mathcal{N}(0, \mathcal{V}^2)$ where $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$ and $\mathcal{V}^2 > 0$.

4.3.4 Extension to the misspecified case

From now we do not assume any longer that the true distribution P^0 belongs to the K-mixtures model. We still consider a prior $\pi = \pi_p \bigotimes_{j=1}^K \pi_j$ on $\theta \in \Theta_K$ for which $\pi_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\pi_j \in \mathcal{M}_1^+(\Theta)$ for $j = 1, ..., K$.

For some value $r_{n,K}$, we introduce the set $\Theta_K(r_{n,K})$ of parameters $\theta^* \in \Theta_K$ such that:

- there exists a set $\mathcal{A}_{n,K} \subset \mathcal{S}_K$ satisfying:
  - for each $p \in \mathcal{A}_{n,K}$, for each $j = 1, ..., K$, $\quad \log\left(\frac{p_j^*}{p_j}\right) \leq K r_{n,K}$,
  - $\pi_p(\mathcal{A}_{n,K}) \geq e^{-nKr_{n,K}}$.

Algorithm 8 Coordinate Descent Variational Bayes for Gaussian mixtures

Require: Initial variational parameters $(\phi_1, ..., \phi_K) \in (\mathbb{R}_+^*)^K$, $(n_j, s_j^2) \in \mathbb{R} \times \mathbb{R}_+^*$ and corresponding variational distributions $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$, $\rho_j = \mathcal{N}(n_j, s_j^2)$ for $j = 1, ..., K$.

repeat

  for $i = 1, ..., n$ **do**

    for $j = 1, ..., K$ **do**

      set $w_j^i = \exp\left(\psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell) - \frac{1}{2}\left\{s_j^2 + (n_j - X_i)^2\right\}\right)$

    end for

    normalize $(w_j^i)_{1 \leq j \leq K}$

  end for

  set $\phi_j = \alpha_j + \alpha \sum_{i=1}^{n} \omega_j^i$ for $j = 1, ..., K$

  set $\rho_p = \mathcal{D}_K(\phi_1, ..., \phi_K)$

  for $j = 1, ..., K$ **do**

    set $n_j = \frac{\alpha \sum_{i=1}^n \omega_j^i X_i}{1/\mathcal{V}^2 + \alpha \sum_{i=1}^n \omega_j^i}$ and $s_j^2 = \frac{1}{1/\mathcal{V}^2 + \alpha \sum_{i=1}^n \omega_j^i}$

    set $\rho_{\mu,j} = \mathcal{N}(n_j, s_j^2)$

  end for

until convergence of the objective function

- there are distributions $\rho_{j,n} \in \mathcal{M}_1^+(\Theta)$ $(j = 1, ..., K)$ such that for $j = 1, ..., K$:

$$\int \mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right] \rho_{j,n}(d\theta_j) \leq r_{n,K} \ , \quad \mathcal{K}(\rho_{j,n}, \pi_j) \leq n r_{n,K}. \tag{4.5}$$

Let us discuss this definition. To begin with, the first item of the definition of $\Theta_K(r_{n,K})$ can seem quite restrictive. It is even a much more stronger assumption than (4.3) and (4.4). Nevertheless, the way to find the required measures $\rho_{p,n}$ in Lemma 4.3.2 in the well-specified case implies constructing in the proof such sets $\mathcal{A}_{n,K}$ for the true parameter weight p^0. As a consequence, it might seem reasonable to replace conditions (4.3) and (4.4) by the first part of the definition of $\Theta_K(r_{n,K})$. On the other hand, the condition given by (4.5) looks like those of Theorem 2.7 in Alquier and Ridgway (2017). Once again, the difference is that inequalities must be satisfied here for each component. A condition on both the true distribution P^0 and the considered parameter θ^* is required through the expectation term. Besides, condition (4.5) is equivalent to (4.3) and (4.4) when the model is well-specified.

Theorem 4.3.6. *For any $\alpha \in (0, 1)$,*

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^{K} p_j q_{\theta_j}, P^0\right) \widetilde{\pi}_{n,\alpha}(d\theta | X_1^n)\right] \leq \frac{\alpha}{1-\alpha} \inf_{\theta^* \in \Theta_K(r_{n,K})} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

Remark 4.3.2. *If there is no $r_{n,K}$ such that $\Theta_K(r_{n,K})$ is not empty, then the right-hand side is equal to infinity (by convention) for any value of $r_{n,K}$ and the inequality is useless. Nevertheless, this is not the case in models used in practice. We show an example below.*

It is worth mentioning that even if this is not exactly an oracle inequality as the risk function in the left-hand side (α-Renyi divergence) is lower than the right-hand side one (Kullback-Leibler divergence), but the theorem still remains of great interest. Indeed, when the minimizer of $\mathcal{K}(P^0, P_\theta)$ with respect to $\theta \in \Theta_K(r_{n,K})$ exists and is such that the corresponding Kullback-Leibler divergence is small, then our oracle inequality is informative as it gives a small bound on the expected risk of the Variational Bayes.

To illustrate the relevance of Theorem 4.3.6, we provide the following result available for a wide range of generating distributions when considering the family of unit-variance Gaussian mixtures with priors $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K) \in \mathcal{M}_1^+(\mathcal{S}_K)$ with $\frac{2}{K} \leq \alpha_j \leq 1$ for $j = 1, ..., K$ $(K \geq 2)$ and $\pi_j = \mathcal{N}(0, \mathcal{V}^2) \in \mathcal{M}_1^+(\mathbb{R})$ for $j = 1, ..., K$ with $\mathcal{V}^2 > 0$:

Corollary 4.3.7. *Assume that the true distribution P^0 is such that $\mathbb{E}|X| < +\infty$. Let $L > 0$.*
For $r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^K \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log\left(\mathcal{V}\right) + \frac{L^2}{2\mathcal{V}^2} - \frac{1}{2}\right]$, we get $\mathcal{S}_K \times [-L, L]^K \subset \Theta_K(r_{n,K})$ and for any $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, P^0\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{\alpha}{1-\alpha}\inf_{\theta^* \in \mathcal{S}_K \times [-L,L]^K} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

Remark 4.3.3. *If the true distribution is a mixture of unit-variance Gaussians with components means between $-L$ and L, then $\mathbb{E}|X| < +\infty$ and the first term of the right-hand side of the inequality is equal to zero, which gives directly for any $\alpha \in (0, 1)$,*

$$\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, P^0\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha} 2K r_{n,K}.$$

4.4 Variational Bayes model selection

In this section, we extend the problem to a larger family of distributions. We want to model the generating distribution P^0 using mixtures with an unknown number of components in a possibly misspecified setting. Thus, we consider a countable collection $\{\mathcal{M}_K / K \in \mathbb{N}^*\}$ of statistical mixture models

$$\mathcal{M}_K = \left\{ P_{\theta_K} = \sum_{j=1}^K p_{j,K} q_{\theta_{j,K}} \ / \ \theta_K \in \Theta_K \right\}$$

with $\Theta_K = \mathcal{S}_K \times \Theta^K$, $\mathcal{S}_K = \{p_K = (p_{1,K}, ..., p_{K,K}) \in [0,1]^K / \sum_{j=1}^K p_{j,K} = 1\}$ and the general notation $\theta_K = (p_K, \theta_{1,K}, ..., \theta_{K,K})$. We precise that the notations are slightly different as the size of each component parameter depends on the model complexity K. The entire parameter space Ω is the union of all parameter spaces Θ_K associated with each model index K: $\Omega = \cup_{K=1}^\infty \Theta_K$, and we can think of a whole statistical model $\mathcal{M} = \cup_{K=1}^\infty \mathcal{M}_K$ as the union of all collections $\mathcal{M}_K$. First, we can notice that different models $\mathcal{M}_K$ never overlap as parameters in each one do not have the same length. Nonetheless, parameters in complex models (models $\mathcal{M}_K$ with large K) can be sparse and therefore

contain the "same information" as parameters in less complex ones, i.e. can lead to the same density P_θ.

The prior specification is a crucial point. As mentioned above, each parameter depends on the number of components. Then, we specify a prior weight π_K assigned to the model $\mathcal{M}_K$ and a conditional prior $\Pi_K(.)$ on $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$. More precisely, we define our conditional prior on $\theta_K = (p_K, \theta_{1,K}, ..., \theta_{K,K})$ as follows: given K, the weight parameter $p_K = (p_{1,K}, ..., p_{K,K})$ is supposed to follow a distribution $\pi_{p,K}$ on $\mathcal{M}_1^+(\mathcal{S}_K)$; finally, given K, we set independent priors $\pi_{j,K}$ for the component parameters $\theta_{j,K}$ where each $\pi_{j,K}$ is a probability distribution on $\mathcal{M}_1^+(\Theta)$. In a nutshell:

$$\pi = \sum_{K=1}^{+\infty} \pi_K \Pi_K$$

with

$$\Pi_K(\theta_K) = \pi_{p,K}(p_K) \prod_{j=1}^{K} \pi_{j,K}(\theta_{j,K}).$$

We have to adapt the notations for the VB approximations. The tempered posteriors $\pi_{n,\alpha}^K(.|X_1^n)$ on parameter $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$, is defined again as

$$\pi_{n,\alpha}^K(d\theta_K|X_1^n) \propto L_n(\theta_K)^\alpha \Pi_K(d\theta_K).$$

The Variational Bayes $\widetilde{\pi}_{n,\alpha}^K(.|X_1^n)$ is the projection of the tempered posterior onto some set $\mathcal{F}_K$ following the mean-field assumption: the variational factor corresponding to the weight parameter $p_K = (p_{1,K}, ..., p_{K,K})$ is any distribution ρ_p on $\mathcal{M}_1^+(\mathcal{S}_K)$; besides, we consider independent variational distributions $\rho_j(\theta_{j,K})$ for the component parameters $\theta_{j,K}$ where each ρ_j is a probability distribution on $\mathcal{M}_1^+(\Theta)$. Then, $\mathcal{F}_K = \{\rho_p \bigotimes_{j=1}^K \rho_j / \rho_p \in \mathcal{M}_1^+(\mathcal{S}_K),\ \rho_j \in \mathcal{M}_1^+(\Theta)\ \forall j = 1, ..., K\}$, and

$$\widetilde{\pi}_{n,\alpha}^K(.|X_1^n) = \arg\min_{\rho_K \in \mathcal{F}_K} \mathcal{K}\Big(\rho_K, \pi_{n,\alpha}^K(.|X_1^n)\Big).$$

We recall that an alternative way to define the variational estimate is to use the Evidence Lower Bound via the optimization program (4.1):

$$\widetilde{\pi}_{n,\alpha}^K(.|X_1^n) = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K) \rho_K(d\theta_K) - \mathcal{K}\Big(\rho_K, \Pi_K\Big) \right\}$$

where the function inside the argmax operator is the ELBO $\mathcal{L}(\rho_K)$. For simplicity, we will just call ELBO $\mathcal{L}(K)$ the closest approximation to the log-evidence, i.e. the value of the lower bound evaluated in its maximum:

$$\mathcal{L}(K) = \alpha \int \ell_n(\theta_K) \widetilde{\pi}_{n,\alpha}^K(d\theta_K|X_1^n) - \mathcal{K}(\widetilde{\pi}_{n,\alpha}^K(.|X_1^n), \Pi_K).$$

The objective is to propose a data-driven estimate $\widehat{K}$ of the number of components from which we will pick up our final VB estimate $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n)$ and derive an oracle inequality

in the spirit of Massart (2007). It is stated in Blei et al. (2017) that $\arg\max_{K\geq 1} \mathcal{L}(K)$ is widely used in practice, without any theoretical justification. We propose

$$\widehat{K} = \arg\max_{K\geq 1} \left\{ \mathcal{L}(K) - \log\left(\frac{1}{\pi_K}\right) \right\}$$

which is a penalized version of the ELBO. Note that taking (π_K) as uniform on a finite set $\{1, 2, \ldots, K_{\max}\}$ leads to the procedure described in Blei et al. (2017). We discuss below the choice $\pi_K = 2^{-K}$.

We can now state the following result which provides an oracle-type inequality for $\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(.|X_1^n)$:

Theorem 4.4.1. *For any* $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \leq \inf_{K\geq 1}\left\{\frac{\alpha}{1-\alpha}\inf_{\theta^*\in\Theta_K(r_{n,K})}\mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha}2Kr_{n,K} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

This oracle inequality shows that our variational distribution adaptively satisfies the best possible balance between bias (misspecification error) and variance (estimation error). If we assume that there is actually a K_0 and $\theta^* \in \Theta_{K_0}$ such that $P^0 = P_{\theta^*}$ then the theorem will imply

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}2K_0 r_{n,K_0} + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

Note that this does not mean that $\widehat{K} = K_0$, but this means that the convergence rate of P_θ to P^0 $\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(.|X_1^n)$ is as good as is we actually knew P_0. The objective of estimating K_0 is a completely different task (Yang, 2005). Estimating K_0 would also require identifiability conditions that are not necessary for our results.

The variance term is composed of two parts. The first one, $Kr_{n,K}$ up to a multiplicative constant, corresponds to the rate obtained when approximating the true distribution with mixtures of model $\mathcal{M}_K$. The second part of the overall rate can be interpreted as a complexity term over the different models reflecting our prior belief. For instance, if we want to penalize more heavily complex models, we can take $\pi_K = 2^{-K}$ and the corresponding term will be of order K/n. In practice, as soon as $\frac{1}{n} \lesssim r_{n,K}$, then this penalty term is negligible when compared to the approximating rate $Kr_{n,K}$: this means that this choice can be considered as safe, it does not interfere with the estimation rate.

4.5 Conclusion

Using variational inference, we studied consistency of variational approximations for estimation and model selection in mixtures. When considering tempered posteriors, we showed that the Variational Bayes is consistent and we gave statistical guaranties to selecting based on the ELBO. For further investigation, it would be interesting to explore the case of Bayesian posteriors when $\alpha = 1$. The recent work of Zhang and Gao (2017)

gives the tools for tackling such an issue, and allows to consider risk functions different than α-Renyi divergence. But the conditions would be more stringent, and misspecification would be more problematic in this case.

Another point of interest is the study of the non-convex optimization program (4.2). Indeed, the proposed coordinate optimization can lead to a local extremum, which implies paying attention to the initialization. The same problem actually arises with EM. In practice, users often run EM or CAVI several times with different initial distributions. Many practical ideas were proposed to target the global extrema more efficiently with EM (O'Hagan et al., 2012) and could be extended to CAVI. But the question of the convergence remains open in theory.

Finally, note that our results are remarkable as there is almost no conditions on the mixtures considered. The counterpart are about the estimation of the true probability distribution P^0, even in the well-specified result. We have no results on the estimation of the parameters. In the case of mixtures, these results are extremely difficult to obtain even for Gaussian mixtures (Wu and Yang, 2018). They require restrictions on the parameters set and lead to different rates of convergence. The consistency of VB for the estimation of the parameters remains open.

4.6 Proofs

4.6.1 Some useful lemmas

We provide in this section two useful lemmas required in many proofs below.

An upper bound on the Kullback-Leibler divergence between two mixtures

The lemma below was first stated by Singer and Warmuth (1999) for mixtures of Gaussians, Do (2003) checked that the proof remains valid for general mixtures. It is a tool widely used in signal processing (Hershey and Olsen, 2007). We provide the proof for the sake of comprehension.

Lemma 4.6.1. *Let $p, p^0 \in \mathcal{S}_K$ and $\theta_j, \theta_j^0 \in \Theta$ for $j = 1, ..., K$. Then,*

$$\mathcal{K}\left(\sum_{j=1}^{K} p_j^0 q_{\theta_j^0}, \sum_{j=1}^{K} p_j q_{\theta_j}\right) \leq \mathcal{K}(p^0, p) + \sum_{j=1}^{K} p_j^0 \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})$$

Proof. For any nonnegative numbers $\alpha_1, ..., \alpha_K$ and positive $\beta_1, ..., \beta_K$, we have:

$$\left(\sum_{j=1}^{K}\alpha_j\right)\log\left(\frac{\sum_{j=1}^{K}\alpha_j}{\sum_{j=1}^{K}\beta_j}\right) = \left(\sum_{j=1}^{K}\beta_j\right)\left(\frac{\sum_{j=1}^{K}\alpha_j}{\sum_{j=1}^{K}\beta_j}\right)\log\left(\frac{\sum_{j=1}^{K}\alpha_j}{\sum_{j=1}^{K}\beta_j}\right)$$
$$= \left(\sum_{j=1}^{K}\beta_j\right)\left(\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l}\frac{\alpha_j}{\beta_j}\right)\log\left(\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l}\frac{\alpha_j}{\beta_j}\right)$$
$$= \left(\sum_{j=1}^{K}\beta_j\right) f\left(\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l}\frac{\alpha_j}{\beta_j}\right)$$

where f is the convex function $x \longmapsto x\log(x)$. As $\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l} = 1$, then using Jensen's inequality:

$$\left(\sum_{j=1}^{K}\alpha_j\right)\log\left(\frac{\sum_{j=1}^{K}\alpha_j}{\sum_{j=1}^{K}\beta_j}\right) = \left(\sum_{j=1}^{K}\beta_j\right) f\left(\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l}\frac{\alpha_j}{\beta_j}\right)$$
$$\leq \left(\sum_{j=1}^{K}\beta_j\right)\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l} f\left(\frac{\alpha_j}{\beta_j}\right)$$
$$= \left(\sum_{j=1}^{K}\beta_j\right)\sum_{j=1}^{K}\frac{\beta_j}{\sum_{l=1}^{K}\beta_l}\frac{\alpha_j}{\beta_j}\log\left(\frac{\alpha_j}{\beta_j}\right)$$
$$= \sum_{j=1}^{K}\alpha_j\log\left(\frac{\alpha_j}{\beta_j}\right).$$

The inequality remains available when some or all β_j's are zero. Indeed, assume that $\beta_j = 0$. If $\alpha_j \neq 0$, then the j^{th} term of the sum in the right-hand side is $\alpha_j\log(\alpha_j/\beta_j) = +\infty$, and the result is obvious. Otherwise, $\alpha_j = 0$, hence the j^{th} term of each sum in the inequality is zero as $\alpha_j\log(\alpha_j/\beta_j) = 0$, and the inequality can be obtained considering only the other numbers.

Thus, for $p, p^0 \in \mathcal{S}_K$ and $\theta_j, \theta_j^0 \in \Theta$ for $j = 1, ..., K$:

$$\mathcal{K}\left(\sum_{j=1}^{K}p_j^0 q_{\theta_j^0}, \sum_{j=1}^{K}p_j q_{\theta_j}\right) = \int\left(\sum_{j=1}^{K}p_j^0 q_{\theta_j^0}\right)\log\left(\frac{\sum_{j=1}^{K}p_j^0 q_{\theta_j^0}}{\sum_{j=1}^{K}p_j q_{\theta_j}}\right)$$
$$\leq \int\sum_{j=1}^{K}p_j^0 q_{\theta_j^0}\log\left(\frac{p_j^0 q_{\theta_j^0}}{p_j q_{\theta_j}}\right)$$
$$= \int\sum_{j=1}^{K}p_j^0 q_{\theta_j^0}\log\left(\frac{p_j^0}{p_j}\right) + \int\sum_{j=1}^{K}p_j^0 q_{\theta_j^0}\log\left(\frac{q_{\theta_j^0}}{q_{\theta_j}}\right)$$
$$= \sum_{j=1}^{K}p_j^0\log\left(\frac{p_j^0}{p_j}\right)\left(\int q_{\theta_j^0}\right) + \sum_{j=1}^{K}p_j^0\int q_{\theta_j^0}\log\left(\frac{q_{\theta_j^0}}{q_{\theta_j}}\right)$$
$$= \mathcal{K}(p^0, p) + \sum_{j=1}^{K}p_j^0\mathcal{K}(q_{\theta_j^0}, q_{\theta_j}).$$

which ends the proof. □

KL-divergence between Gaussian distributions and between Normal-Inverse-Gamma distributions

We give in this section the Kullback-Leibler divergence between 1-dimensional Gaussian distributions and between Normal-Inverse-Gamma distributions. To begin with, one definition:

Definition 4.6.1. *The Normal-Inverse-Gamma $\mathcal{NIG}(\mu, \theta^2, a, b)$ is the distribution which density f with respect to Lebesgue measure is defined by $f(x,y) = g(x|\mu, \frac{y}{\theta^2})h(y|a,b)$, where $g(.|\mu,\sigma^2)$ is the density function of a Gaussian distribution of mean μ and variance σ^2, and $h(.|a,b)$ is the density distribution of an Inverse-Gamma of parameters a and b.*

Lemma 4.6.2. *We denote u and v the density functions of the respective Gaussian distributions $\mathcal{N}(\mu_u, \sigma_u^2)$ and $\mathcal{N}(\mu_v, \sigma_v^2)$. Similarly, we denote p and q the two densities of $\mathcal{NIG}(\mu_1, \theta_1^2, a_1, b_1)$ and $\mathcal{NIG}(\mu_2, \theta_2^2, a_2, b_2)$. Then:*

$$\mathcal{K}(u,v) = \frac{1}{2}\log\left(\frac{\sigma_v^2}{\sigma_u^2}\right) + \frac{\sigma_u^2}{2\sigma_v^2} + \frac{(\mu_v - \mu_u)^2}{2\sigma_v^2} - \frac{1}{2}$$

and

$$\begin{aligned}\mathcal{K}(p,q) = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2-\mu_1)^2}{2}\frac{a_1}{b_1} - \frac{1}{2} \\ + (a_1 - a_2)\psi(a_1) + \log\left(\frac{\Gamma(a_2)}{\Gamma(a_1)}\right) + a_2\log\left(\frac{b_1}{b_2}\right) + a_1\frac{b_2-b_1}{b_1}.\end{aligned}$$

Proof. The first equality is extremely classical so we don't provide the proof. For the second one,

$$\begin{aligned}\mathcal{K}(p,q) &= \int_{\mathbb{R}_+^*}\int_{\mathbb{R}} p(x,y)\log\left(\frac{p(x,y)}{q(x,y)}\right)dxdy \\ &= \int_{\mathbb{R}_+^*}\int_{\mathbb{R}} p(x|Y=y)p_Y(y)\log\left(\frac{p(x|Y=y)}{q(x|Y=y)}\frac{p_Y(y)}{q_Y(y)}\right)dxdy \\ &= \int_{\mathbb{R}_+^*} p_Y(y)\left(\int_{\mathbb{R}} p(x|Y=y)\log\left(\frac{p(x|Y=y)}{q(x|Y=y)}\right)dx\right)dy + \int_{\mathbb{R}_+^*} p_Y(y)\log\left(\frac{p_Y(y)}{q_Y(y)}\right)dy \\ &= \mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\Big[\mathcal{K}(p(.|Y), q(.|Y))\Big] + \mathcal{K}(p_Y, q_Y).\end{aligned}$$

Using the KL-divergence between Gaussians:

$$\mathcal{K}(p(.|Y), q(.|Y)) = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2-\mu_1)^2}{2Y} - \frac{1}{2}$$

hence

$$\mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\Big[\mathcal{K}(p(.|Y), q(.|Y))\Big] = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2-\mu_1)^2}{2}\mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\left[\frac{1}{Y}\right] - \frac{1}{2}$$

i.e.

$$\mathbb{E}_{Y\sim\mathcal{IG}(a_1,b_1)}\Big[\mathcal{K}(p(.|Y), q(.|Y))\Big] = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2-\mu_1)^2}{2}\frac{a_1}{b_1} - \frac{1}{2},$$

and using the KL-divergence between Inverse-Gamma distributions

$$\mathcal{K}(p_Y, q_Y) = (a_1 - a_2)\psi(a_1) + \log\left(\frac{\Gamma(a_2)}{\Gamma(a_1)}\right) + a_2 \log\left(\frac{b_1}{b_2}\right) + a_1 \frac{b_2 - b_1}{b_1}$$

where Γ and ψ are respectively the Gamma and Digamma functions, we have:

$$\begin{aligned}\mathcal{K}(p, q) = \frac{1}{2}\log\left(\frac{\theta_1^2}{\theta_2^2}\right) + \frac{\theta_2^2}{2\theta_1^2} + \frac{\theta_2^2(\mu_2 - \mu_1)^2}{2}\frac{a_1}{b_1} - \frac{1}{2} \\ + (a_1 - a_2)\psi(a_1) + \log\left(\frac{\Gamma(a_2)}{\Gamma(a_1)}\right) + a_2 \log\left(\frac{b_1}{b_2}\right) + a_1 \frac{b_2 - b_1}{b_1}.\end{aligned}$$

□

4.6.2 Proof of Theorem 4.3.1

This result relies on an application of Theorem 2.6 in Alquier and Ridgway (2017) to mixture models. The proof of Theorem 2.6 in Alquier and Ridgway (2017) itself relies mostly on a deviation inequality from Bhattacharya et al. (2016) and on PAC-Bayesian theory (Catoni, 2004; Massart, 2007).

Proof. Fix $0 < \alpha < 1$. Theorem 2.6 from Alquier and Ridgway (2017) gives:

$$\begin{aligned}&\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0}\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \\ &\quad\leq \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int \mathcal{K}\left(\sum_{j=1}^K p_j^0 q_{\theta_j^0}, \sum_{j=1}^K p_j q_{\theta_j}\right)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)}\right\}.\end{aligned}$$

Thanks to Lemma 4.6.1

$$\mathcal{K}\left(\sum_{j=1}^K p_j^0 q_{\theta_j^0}, \sum_{j=1}^K p_j q_{\theta_j}\right) \leq \mathcal{K}(p^0, p) + \sum_{j=1}^K p_j^0 \mathcal{K}(q_{\theta_j^0}, q_{\theta_j}).$$

Then

$$\mathcal{K}(\rho, \pi) = \mathcal{K}\left(\rho_p \bigotimes_{j=1}^K \rho_j, \pi_p \bigotimes_{j=1}^K \pi_j\right) = \mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_j, \pi_j)$$

the last inequality being obtained thanks to Theorem 28 in Van Erven and Harremos (2014). Gathering all the pieces together leads to

$$\begin{aligned}&\mathbb{E}\left[\int D_\alpha\left(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0}\right)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \\ &\quad\leq \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\left[\int \mathcal{K}(p^0, p)\rho_p(dp) + \sum_{j=1}^K \int \mathcal{K}(q_{\theta_j^0}, q_{\theta_j})\rho_j(d\theta_j)\right] + \frac{\mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_j, \pi_j)}{n(1-\alpha)}\right\}\end{aligned}$$

that is the result stated in Theorem 4.3.1. □

4.6.3 Proof of Lemma 4.3.2

Proof. Let us define $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ by the following formula $\rho_{p,n}(dp) \propto \mathbf{1}(p \in \mathcal{B})\pi(dp)$ with

$$\mathcal{B} = \left\{ p \in \mathcal{S}_K / \mathcal{K}(p^0, p) \leq K r'_{n,K} \right\}$$

and

$$r'_{n,K} = \max\left(\frac{1}{K(n-1)}, \frac{\log(n(K-1)\Gamma(A)^{\frac{K}{K-1}}/M_p^0)}{n} \right)$$

where $A = \frac{2}{K}$ and $M_p^0 = \max\{p_j^0 / j = 1, ..., K\}$. We adopt the notation $S = \sum_{j=1}^K \alpha_j$ in the following. We recall that we consider that $K \geq 2$ and then $A = \frac{2}{K} \leq 1$.

First, $\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K r'_{n,K}$.

Then, let us show that $\mathcal{K}(\rho_{p,n}, \pi_p) \leq K n r'_{n,K}$. For that, let us define

$$\mathcal{A} = \left\{ p \in \mathbb{R}^K / p_j^0 e^{-K r'_{n,K}} \leq p_j \leq p_j^0 e^{-K r'_{n,K}} + \frac{p_K^0}{n(K-1)} \text{ for } j = 1, ..., K-1,\ p_K = 1 - \sum_{j=1}^{K-1} p_j \right\}$$

where K is such that $p_K^0 = \max\{p_j^0 / j = 1, ..., K\}$ (this assumption can always be held by relabelling the components of the vector). Then, $p_K^0 \geq \frac{1}{K}$ (otherwise, the sum of the components of p^0 would be strictly lower than 1 and the vector would not be included in $\mathcal{S}_K$). We will show that $\mathcal{A} \subset \mathcal{B}$ and that $\pi_p(\mathcal{A}) \geq e^{-K n r'_{n,K}}$. Then, we will conclude thanks to the following formula: $\mathcal{K}(\rho_{p,n}, \pi_p) = -\log(\pi_p(\mathcal{B}))$.

First, let us show that $\mathcal{A} \subset \mathcal{B}$.

Let $p \in \mathcal{A}$. As $p_K = 1 - \sum_{j=1}^K p_j$, we just need to check that $\mathcal{K}(p^0, p) \leq K r'_{n,K}$ and that $p_j \geq 0$ for each $j = 1, ..., K$.

The first part can be proven using the definition of $\mathcal{A}$. According to the $K-1$ left-hand side inequalities in the definition of $\mathcal{A}$,

$$\begin{aligned}
\mathcal{K}(p^0, p) = \sum_{j=1}^{K-1} p_j^0 \log\left(\frac{p_j^0}{p_j}\right) + p_K^0 \log\left(\frac{p_K^0}{p_K}\right) &\leq \sum_{j=1}^{K-1} p_j^0 \log(e^{K r'_{n,K}}) + p_K^0 \log\left(\frac{p_K^0}{p_K}\right) \\
&= \sum_{j=1}^{K-1} p_j^0 K r'_{n,K} + p_K^0 \log\left(\frac{p_K^0}{p_K}\right) \\
&= (1 - p_K^0) K r'_{n,K} + p_K^0 \log\left(\frac{p_K^0}{p_K}\right).
\end{aligned}$$

All we need to show now is that $\log\left(\frac{p_K^0}{p_K}\right) \leq K r'_{n,K}$. This comes from the following inequalities:

$$\log\left(\frac{p^0_K}{p_K}\right) = \log\left(\frac{p^0_K}{1-\sum_{j=1}^{K-1} p_j}\right) \leq \log\left(\frac{p^0_K}{1-\sum_{j=1}^{K-1} p^0_j e^{-Kr'_{n,K}} - \frac{p^0_K}{n}}\right)$$
$$= \log\left(\frac{p^0_K}{1-(1-p^0_K)e^{-Kr'_{n,K}} - \frac{p^0_K}{n}}\right)$$
$$\leq \frac{p^0_K}{1-(1-p^0_K)e^{-Kr'_{n,K}} - \frac{p^0_K}{n}} - 1$$
$$= \frac{p^0_K - 1 + (1-p^0_K)e^{-Kr'_{n,K}} + \frac{p^0_K}{n}}{1-(1-p^0_K)e^{-Kr'_{n,K}} - \frac{p^0_K}{n}}$$

i.e.

$$\log\left(\frac{p^0_K}{p_K}\right) \leq \frac{p^0_K - 1 + (1-p^0_K)e^{-Kr'_{n,K}} + \frac{p^0_K}{n}}{1-(1-p^0_K)e^{-Kr'_{n,K}} - \frac{p^0_K}{n}}$$
$$= \frac{\frac{p^0_K}{n} - (1-p^0_K)(1-e^{-Kr'_{n,K}})}{p^0_K(1-\frac{1}{n}) + (1-p^0_K)(1-e^{-Kr'_{n,K}})}$$
$$= \frac{\frac{1}{n} - (\frac{1}{p^0_K}-1)(1-e^{-Kr'_{n,K}})}{(1-\frac{1}{n}) + (\frac{1}{p^0_K}-1)(1-e^{-Kr'_{n,K}})}$$
$$\leq \frac{\frac{1}{n}}{1-\frac{1}{n}} = \frac{1}{n-1}$$
$$\leq Kr'_{n,K}.$$

Hence $\mathcal{K}(p^0,p) \leq (1-p^0_K)Kr'_{n,K} + p^0_K \log\left(\frac{p^0_K}{p_K}\right) \leq (1-p^0_K)Kr'_{n,K} + p^0_K Kr'_{n,K} = Kr'_{n,K}$.

On the other hand, for $j = 1, ..., K-1$, $p_j \geq p^0_j e^{-Kr'_{n,K}} \geq 0$ and:

$$p_K = 1 - \sum_{j=1}^{K-1} p_j \geq 1 - \sum_{j=1}^{K-1} \left(p^0_j e^{-Kr'_{n,K}} + \frac{p^0_K}{n(K-1)}\right)$$
$$= 1 - \left((1-p^0_K)e^{-Kr'_{n,K}} + \frac{p^0_K}{n}\right)$$
$$\geq 1 - (1-p^0_K)e^{-Kr'_{n,K}} - p^0_K$$
$$= (1-p^0_K)(1-e^{-Kr'_{n,K}})$$
$$\geq 0.$$

Then, $p \in \mathcal{B}$, and finally $\mathcal{A} \subset \mathcal{B}$.

Now, let us show that $\pi_p(\mathcal{A}) \geq e^{-Knr'_{n,K}}$.

Let us denote f the density of the $\pi_p = \mathcal{D}_K(\alpha_1, ..., \alpha_K)$ Dirichlet distribution:

$$f(p) = \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \mathbf{1}(p \in \mathcal{B}).$$

Thus, we can lower bound $\pi_p(\mathcal{A})$:

$$\pi_p(\mathcal{A}) = \int_{\mathcal{A}} f(p_1, ..., p_K)dp = \int_{\mathcal{A}} \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}\, \mathbf{1}(p \in \mathcal{B})dp$$

$$\geq \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K-1} \int_{p_j^0 e^{-Kr'_{n,K}}}^{p_j^0 e^{-Kr'_{n,K}} + \frac{p_K^0}{n(K-1)}} p_j^{\alpha_j - 1}\, dp_j$$

as for $p \in \mathcal{A}$, $0 \leq p_j^0 e^{-Kr'_{n,K}} \leq p_j \leq p_j^0 e^{-Kr'_{n,K}} + \frac{p_K^0}{n(K-1)} \leq 1$ for each $j = 1, ..., K$ (as $\mathcal{A} \subset \mathcal{B}$), and then $p_j^{\alpha_j - 1} \geq 1$.

Then, by definition of $r'_{n,K}$, $\frac{p_K^0}{n(K-1)} \geq \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}$, and using inequalities $\Gamma(A) \geq \Gamma(\alpha_j)$ as $A \leq \alpha_j \leq 1$ and $\Gamma(S) \geq 1$ as $S \geq 2$,

$$\pi_p(\mathcal{A}) \geq \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K-1} \int_{p_j^0 e^{-Kr'_{n,K}}}^{p_j^0 e^{-Kr'_{n,K}} + \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}} p_j^{\alpha_j - 1}\, dp_j$$

$$\geq \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K-1} \int_{p_j^0 e^{-Kr'_{n,K}}}^{p_j^0 e^{-Kr'_{n,K}} + \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}} dp_j$$

$$= \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \prod_{j=1}^{K-1} \Gamma(A)^{\frac{K}{K-1}} e^{-nr'_{n,K}}$$

$$= \frac{\Gamma\Big(S\Big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \Gamma(A)^K e^{-n(K-1)r'_{n,K}}$$

$$\geq e^{-nKr'_{n,K}}.$$

Hence, as $\mathcal{A} \subset \mathcal{B}$, $\pi_p(\mathcal{B}) \geq \pi_p(\mathcal{A}) \geq e^{-nKr'_{n,K}}$, and finally, $\mathcal{K}(\rho_{p,n}, \pi_p) = -\log(\pi_p(\mathcal{B})) \leq Knr'_{n,K}$.

We just proved the lemma but with the rate $r'_{n,K}$ instead of the value $r_{n,K}$ used in the lemma. We can conlude by noticing that the result is available for every r such that $r'_{n,K} \leq r$, and that in particular $r'_{n,K} \leq r_{n,K}$. This last result comes from the inequality:

$$\Gamma(A) \leq \frac{\Gamma(1 + \frac{A}{2})}{\Big(\frac{A}{2}\Big)^{1 - \frac{A}{2}}}$$

which is a direct application of the left-hand side of inequality (3.2) part 3 in Laforgia and Natalin (2013) with $x = \frac{A}{2} > 0$ and $\lambda = \frac{A}{2} \in (0, 1)$. As, $1 + \frac{A}{2} \in [1, 2]$, then $\Gamma(1 + \frac{A}{2}) \leq 1$, and $\frac{1}{\big(\frac{A}{2}\big)^{1 - \frac{A}{2}}} = K^{1 - \frac{A}{2}} \leq K$. Thus:

$$\Gamma(A) \leq K$$

and as $K \geq 2$ and $p_K^0 \geq \frac{1}{K}$, it follows that

$$\log((K-1)\Gamma(A)^{\frac{K}{K-1}}/p_K^0) \leq \log(K(K)^{\frac{K}{K-1}}K) \leq \log(K(K)^2 K) \leq \log(K^4)$$

i.e. $r'_{n,K} \leq \max(\frac{1}{K(n-1)}, \frac{\log(nK^4)}{n}) \leq \max(\frac{1}{K(n-1)}, \frac{4\log(nK)}{n})$. Besides, $\frac{n}{n-1} = 1 + \frac{1}{n-1} \leq 2$ implies $\frac{1}{K(n-1)} \leq \frac{1}{2(n-1)} \leq \frac{1}{n} \leq \frac{4\log(2)}{n} \leq \frac{4\log(nK)}{n}$, and finally $r'_{n,K} \leq \frac{4\log(nK)}{n} = r_{n,K}$.

□

4.6.4 Proof of Corollary 4.3.3

Proof. According to the Lemma 4.3.2, there exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that

$$\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K\frac{4\log(nK)}{n}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq Kn\frac{4\log(nK)}{n}.$$

Similarly, the same result states that there exists distributions $\rho_{j,n} \in \mathcal{M}_1^+(\mathcal{S}_V)$ for $j = 1, ..., K$ such that

$$\int \mathcal{K}(\theta_j^0, \theta_j)\rho_{j,n}(d\theta_j) \leq \frac{4V\log(nV)}{n}$$

and

$$\mathcal{K}(\rho_{j,n}, \pi_j) \leq n\frac{4V\log(nV)}{n}.$$

We conclude using theorem 4.3.1:

$$\mathbb{E}\left[\int D_\alpha(\sum_{j=1}^K p_j q_{\theta_j}, \sum_{j=1}^K p_j^0 q_{\theta_j^0})\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}\left[\frac{8KV\log(nV)}{n} \bigvee \frac{8K\log(nK)}{n}\right].$$

□

4.6.5 Proof of Corollary 4.3.4

Proof. For $R_{j,n} = \frac{1}{n} \bigvee \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2}\right]$ (for $j = 1, ..., K$), there exists distributions $\rho_{j,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ for $j = 1, ..., K$ such that

$$\int \mathcal{K}(q_{\mu_j^0}, q_{\mu_j})\rho_{j,n}(d\mu_j) \leq R_{j,n}$$

and

$$\mathcal{K}(\rho_{j,n}, \pi_j) \leq nR_{j,n}.$$

Indeed, let us define $\rho_{j,n}$ as a Gaussian distribution of mean μ_j^0 and variance $\frac{2V^2}{n}$. According to Lemma 4.6.2:

$$\mathcal{K}(q_{\mu_j^0}, q_{\mu_j}) = \frac{(\mu_j - \mu_j^0)^2}{2V^2}.$$

Then, $\int \mathcal{K}(q_{\mu_j^0}, q_{\mu_j})\rho_{j,n}(d\mu_j) = \frac{1}{2V^2}\mathbb{E}_{\mu_j\sim\rho_{j,n}}[(\mu_j - \mu_j^0)^2]$

$$\begin{aligned}
&= \frac{1}{2V^2} \times \frac{2V^2}{n} \\
&= \frac{1}{n} \\
&\leq R_{j,n}.
\end{aligned}$$

We can finally conclude using the formula using again Lemma 4.6.2:

$$\begin{aligned}
\mathcal{K}(\rho_{j,n}, \pi_j) &= \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{2V^2}\right) + \frac{V^2}{n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \\
&= \frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} \\
&= n \times \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2}\right] \\
&\leq nR_{j,n}.
\end{aligned}$$

In addition, let us recall that there also exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that

$$\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K\frac{4\log(nK)}{n}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq nK\frac{4\log(nK)}{n}.$$

For $r_{n,K} = \frac{4\log(nK)}{n}\bigvee_{j=1}^K R_{j,n} = \frac{4\log(nK)}{n} \vee \frac{1}{n}\bigvee_{j=1}^K \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2}\right]$ i.e. $r_{n,K} = \frac{4\log(nK)}{n}\bigvee_{j=1}^K \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{V^2}{n\mathcal{V}^2} + \log\left(\frac{\mathcal{V}}{V}\right) + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2}\right]$, we finally obtain the required inequality using theorem 4.3.1.

□

4.6.6 Proof of Corollary 4.3.5

Normal-Inverse-Gamma prior

Proof. First, let us focus on the first result, when the chosen prior is the Normal-Inverse-Gamma $\pi_j = \mathcal{NIG}(0, \mathcal{V}^2, 1, \gamma^2)$ for each $j = 1, ..., K$. In order to obtain the required rate

$$r_{n,K} = \frac{4\log(nK)}{n}\bigvee_{j=1}^K \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - (\sigma_j^0)^2}{(\sigma_j^0)^2}\right],$$

we proceed as previously and find a variational density on both the mean and the variance such that the two different terms $\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j, d\sigma_j^2)$ and $\mathcal{K}(\rho_{j,n}, \pi_j)$ are upper bounded for $j = 1, ..., K$.

Let us define $\rho_{j,n}$ as a Normal-Inverse-Gamma distribution $\mathcal{NIG}(\mu_j^0, \lambda_n, a_n, b_n)$ where λ_n, a_n and b_n are hyperparameters that we will precise later. Using Lemma 4.6.2:

$$\mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)}) = \frac{1}{2}\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2\sigma_j^2} + \frac{(\mu_j-\mu_j^0)^2}{2\sigma_j^2} - \frac{1}{2}.$$

$$\text{Then, } \int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = \frac{1}{2}\mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right)\right] + \mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\frac{(\sigma_j^0)^2}{2\sigma_j^2}\right] + \mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\frac{(\mu_j-\mu_j^0)^2}{2\sigma_j^2}\right] - \frac{1}{2}.$$

As

$$\mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\frac{(\sigma_j^0)^2}{2\sigma_j^2}\right] = \frac{(\sigma_j^0)^2}{2}\mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\frac{1}{\sigma_j^2}\right] = \frac{(\sigma_j^0)^2}{2}\frac{a_n}{b_n},$$

$$\mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right)\right] = \frac{1}{2}\Big(\log(b_n) - \psi(a_n)\Big) - \frac{1}{2}\log((\sigma_j^0)^2)$$

and

$$\mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\frac{(\mu_j-\mu_j^0)^2}{2\sigma_j^2}\right] = \mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\frac{1}{2\sigma_j^2}.\mathbb{E}_{\mu_j\sim\mathcal{N}(\mu_j^0,\frac{\sigma_j^2}{\lambda_n})}[(\mu_j-\mu_j^0)^2]\right]$$

i.e.

$$\mathbb{E}_{(\mu_j,\sigma_j^2)\sim\rho_{j,n}}\left[\frac{(\mu_j-\mu_j^0)^2}{2\sigma_j^2}\right] = \mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\frac{1}{2\sigma_j^2}.\frac{\sigma_j^2}{\lambda_n}\right] = \frac{1}{2\lambda_n},$$

we get:

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = -\frac{1}{2} + \frac{(\sigma_j^0)^2}{2}\frac{a_n}{b_n} + \frac{1}{2\lambda_n} + \frac{1}{2}\Big(\log(b_n) - \psi(a_n)\Big) - \frac{1}{2}\log((\sigma_j^0)^2).$$

Now, we compute the term $\mathcal{K}(\rho_{j,n}, \pi_j)$ using the fomula giving the Kullback-Leibler divergence between two Gaussian-Inverse-Gamma distributions. Using Lemma 4.6.2:

$$\mathcal{K}(\rho_{j,n}, \pi_j) = \frac{1}{2}\log\left(\frac{\lambda_n}{\mathcal{V}^{-2}}\right) + \frac{\mathcal{V}^{-2}}{2\lambda_n} + \frac{\mathcal{V}^{-2}(\mu_j^0)^2}{2}\frac{a_n}{b_n} - \frac{1}{2} + (a_n-1)\psi(a_n) + \log\left(\frac{1}{\Gamma(a_n)}\right) + \log\left(\frac{b_n}{\gamma^2}\right) + a_n\frac{\gamma^2-b_n}{b_n}.$$

Then, for $\lambda_n = n$, $a_n = 1$ and $b_n = (\sigma_j^0)^2$:

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = \frac{1}{2n} \leq R_{j,n}$$

and

$$\begin{aligned}\mathcal{K}(\rho_{j,n}, \pi_j) &= \frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2-(\sigma_j^0)^2}{(\sigma_j^0)^2} \\ &= n \times \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2-(\sigma_j^0)^2}{(\sigma_j^0)^2}\right] \\ &\leq nR_{j,n}.\end{aligned}$$

with $R_{j,n} = \frac{1}{2n} \bigvee \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2-(\sigma_j^0)^2}{(\sigma_j^0)^2}\right]$

We end the proof by remaining that there also exists a distribution $\rho_{p,n} \in \mathcal{M}_1^+(\mathcal{S}_K)$ such that

$$\int \mathcal{K}(p^0, p)\rho_{p,n}(dp) \leq K\frac{4\log(nK)}{n}$$

and

$$\mathcal{K}(\rho_{p,n}, \pi_p) \leq nK\frac{4\log(nK)}{n}.$$

We can finally conclude using again Theorem 4.3.1 with

$$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} R_{j,n}$$
$$= \frac{4\log(nK)}{n} \bigvee \frac{1}{2n} \bigvee_{j=1}^{K} \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2-(\sigma_j^0)^2}{(\sigma_j^0)^2}\right]$$

i.e.

$$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^{K} \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2(\sigma_j^0)^2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2-(\sigma_j^0)^2}{(\sigma_j^0)^2}\right].$$

□

Factorized prior

Proof. Let us focus now on the case of independant priors $\pi_j = \mathcal{N}(0, \mathcal{V}^2) \bigotimes \mathcal{IG}(1, \gamma^2)$ for $j = 1, ..., K$. The proof is almost the same as previously.

We define here $\rho_{j,n}$ as the product measure of Normal distribution $\mathcal{N}(\mu_j^0, \theta_n^2)$ and of an Inverse-Gamma distribution $\mathcal{IG}(a_n, b_n)$ where θ_n^2, a_n and b_n are hyperparameters to be detailed later. Then, we have again:

$$\mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)}) = \frac{1}{2}\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right) + \frac{(\sigma_j^0)^2}{2\sigma_j^2} + \frac{(\mu_j-\mu_j^0)^2}{2\sigma_j^2} - \frac{1}{2}.$$

Hence,

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = \frac{1}{2}\mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\log\left(\frac{\sigma_j^2}{(\sigma_j^0)^2}\right)\right] + \mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\frac{(\sigma_j^0)^2}{2\sigma_j^2}\right]$$
$$+ \mathbb{E}_{\mu_j\sim\mathcal{N}(\mu_j^0,\theta_n^2)}\left[(\mu_j-\mu_j^0)^2\right]\mathbb{E}_{\sigma_j^2\sim\mathcal{IG}(a_n,b_n)}\left[\frac{1}{2\sigma_j^2}\right] - \frac{1}{2}.$$

i.e.

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = -\frac{1}{2} + \frac{a_n}{2b_n}\left((\sigma_j^0)^2 + \theta_n^2\right) + \frac{1}{2}\left(\log(b_n) - \psi(a_n)\right) - \frac{1}{2}\log((\sigma_j^0)^2).$$

Then we compute the term $\mathcal{K}(\rho_{j,n}, \pi)$ as the sum of the Kullback-Leibler divergence between two Gaussian distributions and between two Inverse-Gamma distributions:

$$\mathcal{K}(\rho_{j,n}, \pi_j) = \frac{1}{2}\log\left(\frac{\mathcal{V}^2}{\theta_n^2}\right) + \frac{\theta_n^2}{2\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + (a_n - 1)\psi(a_n) + \log\left(\frac{1}{\Gamma(a_n)}\right) + \log\left(\frac{b_n}{\gamma^2}\right) + a_n\frac{\gamma^2 - b_n}{b_n}.$$

Then, for $\theta_n^2 = \frac{1}{n}$, $a_n = 1$ and $b_n = (\sigma_j^0)^2$:

$$\int \mathcal{K}(q_{(\mu_j^0,(\sigma_j^0)^2)}, q_{(\mu_j,\sigma_j^2)})\rho_{j,n}(d\mu_j) = \frac{1}{2(\sigma_j^0)^2 n} \leq R_{j,n}$$

and

$$\begin{aligned}\mathcal{K}(\rho_{j,n}, \pi_j) &= \frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - (\sigma_j^0)^2}{(\sigma_j^0)^2} \\ &= n \times \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2 - (\sigma_j^0)^2}{(\sigma_j^0)^2}\right] \\ &\leq nR_{j,n}\end{aligned}$$

with $R_{j,n} = \frac{1}{2(\sigma_j^0)^2 n} \vee \frac{1}{n}\left[\frac{1}{2}\log\left(n\mathcal{V}^2\right) + \frac{1}{2n\mathcal{V}^2} + \frac{(\mu_j^0)^2}{2\mathcal{V}^2} - \frac{1}{2} + \log\left(\frac{(\sigma_j^0)^2}{\gamma^2}\right) + \frac{\gamma^2-(\sigma_j^0)^2}{(\sigma_j^0)^2}\right]$.

The end of the proof is the same as the one used in the Normal-Inverse-Gamma case.

□

4.6.7 Proof of Theorem 4.3.6

Proof. We assume that $\Theta_K(r_{n,K})$ is not empty (otherwise, this is obvious). Applying Theorem 2.7 in Alquier and Ridgway (2017) for any $\alpha \in (0,1)$, $\theta^* \in \Theta_K(r_{n,K})$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{\alpha}{1-\alpha}\mathcal{K}(P^0, P_{\theta^*}) + \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int \mathbb{E}\left[\log\frac{P_{\theta^*}(X_i)}{P_\theta(X_i)}\right]\rho(d\theta) + \frac{\mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_j, \pi_j)}{n(1-\alpha)}\right\}.$$

Let us take $\rho_{j,n}$ and $\mathcal{A}_{n,K}$ from the definition of $\Theta_K(r_{n,K})$, and $\rho_{p,n}(dp) \propto \mathbf{1}(p \in \mathcal{A}_{n,K})\pi(dp)$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{\alpha}{1-\alpha}\mathcal{K}(P^0, P_{\theta^*}) + \frac{\alpha}{1-\alpha}\int \mathbb{E}\left[\log\frac{P_{\theta^*}(X_i)}{P_\theta(X_i)}\right]\rho_{p,n}(dp)\prod_{j=1}^K \rho_{j,n}(d\theta_j) + \frac{\mathcal{K}(\rho_{p,n}, \pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_{j,n}, \pi_j)}{n(1-\alpha)}.$$

We have $\mathcal{K}(\rho_{p,n}, \pi_p) = -\log(\pi_p(\mathcal{A}_{n,K})) \leq nKr_{n,K}$ and $\mathcal{K}(\rho_{j,n}, \pi_j) \leq nr_{n,K}$ for each j by definition of $\Theta_K(r_{n,K})$. Moreover, as for Lemma 4.6.1:

$$\begin{aligned}
\log \frac{P_{\theta^*}(X)}{P_\theta(X)} &= \frac{1}{P_{\theta^*}(X)} P_{\theta^*}(X) \log \frac{P_{\theta^*}(X)}{P_\theta(X)} \\
&\leq \frac{1}{P_{\theta^*}(X)} \sum_{j=1}^K p_j^* q_{\theta_j^*}(X) \log \frac{p_j^* q_{\theta_j^*}(X)}{p_j q_{\theta_j}(X)} \\
&= \sum_{j=1}^K \frac{p_j^* q_{\theta_j^*}(X)}{P_{\theta^*}(X)} \log \frac{p_j^*}{p_j} + \sum_{j=1}^K \frac{p_j^* q_{\theta_j^*}(X)}{P_{\theta^*}(X)} \log \frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)} \\
&\leq \sum_{j=1}^K \frac{p_j^* q_{\theta_j^*}(X)}{P_{\theta^*}(X)} \log \frac{p_j^*}{p_j} + \sum_{j=1}^K \log \frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}
\end{aligned}$$

and thus, as the support of $\rho_{p,n}$ is on $\mathcal{A}_{n,K}$ on which $\log \frac{p_j^*}{p_j} \leq Kr_{n,K}$,

$$\begin{aligned}
\int \mathbb{E}\left[\log \frac{P_{\theta^*}(X)}{P_\theta(X)}\right] \rho_{p,n}(dp) \prod_{j=1}^K \rho_{j,n}(d\theta_j) &\leq \int P^0 \sum_{j=1}^K \frac{p_j^* q_{\theta_j^*}}{P_{\theta^*}} \log \frac{p_j^*}{p_j} \, d\mu \, \rho_{p,n}(dp) \\
&\quad + \sum_{j=1}^K \int \mathbb{E}\left[\log \frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right] \rho_{j,n}(d\theta_j) \\
&\leq \int P^0 \sum_{j=1}^K \frac{p_j^* q_{\theta_j^*}}{P_{\theta^*}} Kr_{n,K} \, d\mu \, \rho_{p,n}(dp) + Kr_{n,K} \\
&= 2Kr_{n,K}
\end{aligned}$$

which ends the proof as it holds for any $\theta^* \in \Theta_K(r_{n,K})$.

□

4.6.8 Proof of Corollary 4.3.7

Proof. It is sufficient to show that $\mathcal{S}_K \times [-L, L]^K \subset \Theta_K(r_{n,K})$ for

$$r_{n,K} = \frac{4\log(nK)}{n} \bigvee_{j=1}^K \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log\left(\mathcal{V}\right) + \frac{L^2}{2\mathcal{V}^2} - \frac{1}{2}\right],$$

the oracle inequality being a direct corollary. For that, let us take any $\theta^* \in \mathcal{S}_K \times [-L, L]^K$ and show that it satisfies the conditions in the definition of $\Theta_K(r_{n,K})$.

The existence of a set $\mathcal{A}_{n,K}$ filling the first condition has already been done in the proof of Lemma 4.3.2 as $\frac{4\log(nK)}{n} \leq r_{n,K}$.

We define distributions $\rho_{j,n} \in \mathcal{M}_1^+(\Theta)$ by Gaussians of mean θ_j^* and variance $\frac{2}{n}$ ($j = 1, ..., K$) and we show that for $j = 1, ..., K$:

$$\int \mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right] \rho_{j,n}(d\theta_j) \leq r_{n,K} \quad , \quad \mathcal{K}(\rho_{j,n}, \pi_j) \leq nr_{n,K}.$$

We start from

$$\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right) = \frac{(\theta_j - \theta_j^*)^2}{2} - (X - \theta_j^*)(\theta_j - \theta_j^*)$$

and if we take the mean of this quantity with respect to P^0, we obtain:

$$\mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right] = \frac{(\theta_j - \theta_j^*)^2}{2} - (\mathbb{E}X - \theta_j^*)(\theta_j - \theta_j^*)$$

and as $\theta_j - \theta_j^*$ is a zero-mean random variable, we have:

$$\begin{aligned}\int \mathbb{E}\left[\log\left(\frac{q_{\theta_j^*}(X)}{q_{\theta_j}(X)}\right)\right]\rho_{j,n}(d\theta_j) &= \frac{1}{2}\mathbb{E}_{\theta_j\sim\rho_{j,n}}[(\theta_j - \theta_j^*)^2] - (\mathbb{E}X - \theta_j^*)\mathbb{E}_{\theta_j\sim\rho_{j,n}}[\theta_j - \theta_j^*] \\ &= \frac{1}{2} \times \frac{2}{n} \\ &\leq r_{n,K}.\end{aligned}$$

Then, we conclude according to Lemma 4.6.2:

$$\begin{aligned}\mathcal{K}(\rho_{j,n}, \pi) &= \frac{1}{2}\log\left(\frac{n\mathcal{V}^2}{2}\right) + \frac{1}{n\mathcal{V}^2} + \frac{(\theta_j^*)^2}{2\mathcal{V}^2} - \frac{1}{2} \\ &= \frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log\left(\mathcal{V}\right) + \frac{(\theta_j^*)^2}{2\mathcal{V}^2} - \frac{1}{2} \\ &\leq n \times \frac{1}{n}\left[\frac{1}{2}\log\left(\frac{n}{2}\right) + \frac{1}{n\mathcal{V}^2} + \log\left(\mathcal{V}\right) + \frac{L^2}{2\mathcal{V}^2} - \frac{1}{2}\right] \\ &\leq n r_{n,K}.\end{aligned}$$

□

4.6.9 Proof of Theorem 4.4.1

Here, we cannot directly use a result from Alquier and Ridgway (2017). So we start the proof from scratch, using the main lines of Bhattacharya et al. (2016); Alquier and Ridgway (2017) with adequate adaptation.

Proof. For any $\alpha \in (0,1)$ and $\theta \in \Omega$, by definition of the Renyi divergence and using $D_\alpha(P^{\otimes n}, R^{\otimes n}) = nD_\alpha(P,R)$ as data are i.i.d.:

$$\mathbb{E}\left[\exp\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha)nD_\alpha(P_\theta, P^0)\right)\right] = 1$$

Thus, integrating and using Fubini's theorem,

$$\mathbb{E}\left[\int \exp\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha)nD_\alpha(P_\theta, P^0)\right)\pi(d\theta)\right] = 1$$

Using Lemma 4.2.1,

$$\mathbb{E}\left[\exp\left(\sup_{\rho\in\mathcal{M}_1^+(\Omega)}\left\{\int\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha)nD_\alpha(P_\theta, P^0)\right)\rho(d\theta) - \mathcal{K}(\rho,\pi)\right\}\right)\right] = 1.$$

Note that Bhattacharya et al. (2016); Alquier and Ridgway (2017) also used Lemma 4.2.1 in their proofs, this is inspired by the PAC-Bayesian theory (Catoni, 2004, 2007).

It is interesting to note that Lemma 4.2.1 is at the core of VB: it is used to provide approximation algorithms, and also to prove the consistency of VB. Thanks to Jensen's inequality,

$$\mathbb{E}\Bigg[\sup_{\rho\in\mathcal{M}_1^+(\Omega)}\Bigg\{\int\Bigg(-\alpha r_n(P_\theta,P^0)+(1-\alpha)nD_\alpha(P_\theta,P^0)\Bigg)\rho(d\theta)-\mathcal{K}(\rho,\pi)\Bigg\}\Bigg]\leq 0$$

Therefore, when considering $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n)$ as a distribution on $\mathcal{M}_1^+(\Omega)$ with all its mass on $\Theta_{\widehat{K}}$,

$$\mathbb{E}\Bigg[\int\Bigg(-\alpha r_n(P_\theta,P^0)+(1-\alpha)nD_\alpha(P_\theta,P^0)\Bigg)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)-\mathcal{K}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\pi)\Bigg]\leq 0$$

Using $\mathcal{K}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\pi)=\mathcal{K}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\Pi_{\widehat{K}})+\log(\frac{1}{\pi_{\widehat{K}}})$, we rearrange terms:

$$\mathbb{E}\Bigg[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\Bigg]$$
$$\leq\mathbb{E}\Bigg[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)+\frac{\mathcal{K}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\Pi_{\widehat{K}})}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_{\widehat{K}}})}{n(1-\alpha)}\Bigg]$$

Thus, by definition of $\widehat{K}$,

$$\mathbb{E}\Bigg[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\Bigg]$$
$$\leq\mathbb{E}\Bigg[\inf_{K\geq 1}\Bigg\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n)+\frac{\mathcal{K}(\widetilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Bigg\}\Bigg]$$

which leads to

$$\mathbb{E}\Bigg[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\Bigg]$$
$$\leq\inf_{K\geq 1}\Bigg\{\mathbb{E}\Bigg[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n)+\frac{\mathcal{K}(\widetilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Bigg]\Bigg\}$$

and by definition of $\widetilde{\pi}_{n,\alpha}^{K}(.|X_1^n)$,

$$\mathbb{E}\Bigg[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\Bigg]$$
$$\leq\inf_{K\geq 1}\Bigg\{\mathbb{E}\Bigg[\inf_{\rho\in\mathcal{M}_1^+(\Theta_K)}\Bigg\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\rho(d\theta)+\frac{\mathcal{K}(\rho,\Pi_K)}{n(1-\alpha)}\Bigg\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\Bigg]\Bigg\}.$$

Then,

$$\mathbb{E}\Bigg[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\Bigg]$$

$$\leq \inf_{K\geq 1} \inf_{\rho\in\mathcal{M}_1^+(\Theta_K)} \left\{ \mathbb{E}\left[\frac{\alpha}{1-\alpha} \int \frac{r_n(P_\theta, P^0)}{n} \rho(d\theta) + \frac{\mathcal{K}(\rho, \Pi_K)}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)} \right] \right\}.$$

And finally,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n) \right] \leq \inf_{K\geq 1} \inf_{\rho\in\mathcal{M}_1^+(\Theta_K)} \left\{ \frac{\alpha}{1-\alpha} \int \mathcal{K}(P^0, P_\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho, \Pi_K)}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)} \right\}.$$

To conclude, we just need to upper bound the function inside the infimum over all integers K's by $\frac{\alpha}{1-\alpha} \inf_{\theta^*\in\Theta_K(r_{n,K})} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} 2K r_{n,K} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}$. This is direct: if the set $\Theta_K(r_{n,K})$ is not empty (otherwise the inequality is obvious) we notice that $\mathcal{K}(P^0, P_\theta) = \mathcal{K}(P^0, P_{\theta^*}) + \mathbb{E}\left[\log \frac{P_{\theta^*}(X_i)}{P_\theta(X_i)} \right]$ for any $\theta^* \in \Theta_K(r_{n,K})$ and then we follow the sketch of the proof of Theorem 4.3.6.

□

4.6.10 Algorithms

We now provide the derivations leading to the algorithms described in the paper.

Algorithm 1

We apply a coordinate descent on variables $\omega^1 \in \mathcal{S}_K$,..., $\omega^n \in \mathcal{S}_K$, $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$, $\rho_1 \in \mathcal{M}_1^+(\Theta)$,..., and $\rho_K \in \mathcal{M}_1^+(\Theta)$ in order to solve the optimization program:

$$\min_{\rho\in\mathcal{F},\, w\in\mathcal{S}_K^n} \left\{ -\alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \left(\int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) \right) \right.$$
$$\left. + \alpha \sum_{i=1}^n \sum_{j=1}^K \omega_j^i \log(\omega_j^i) + \mathcal{K}(\rho_p, \pi_p) + \sum_{j=1}^K \mathcal{K}(\rho_j, \pi_j) \right\}.$$

We explain how to obtain Algorithm 1.

Optimization with respect to $\omega^i \in \mathcal{S}_K$:

First, we fix $\omega^\ell \in \mathcal{S}_K$ for $\ell \neq i$, $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\rho_j \in \mathcal{M}_1^+(\Theta)$ for $j = 1, ..., K$, and we solve the program with respect to $\omega^i \in \mathcal{S}_K$, which becomes:

$$\min_{\omega^i\in\mathcal{S}_K} \left\{ \sum_{j=1}^K \omega_j^i \left(\log(\omega_j^i) - \int \log(p_j)\rho_p(dp) - \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) \right) \right\}.$$

Put $\mathbf{E} = \{1, ..., K\}$, $\lambda = \left(\frac{1}{K}, ..., \frac{1}{K}\right)$ and $h(j) = \int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j)$ and use Lemma 4.2.1 to obtain:

$$w_j^i \propto \exp\left(\int \log(p_j)\rho_p(dp) + \int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) \right).$$

Optimization with respect to $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$:

Now, we fix $\omega^i \in \mathcal{S}_K$ for $i = 1, ..., n$, and $\rho_j \in \mathcal{M}_1^+(\Theta)$ for $j = 1, ..., K$, and we solve the program with respect to $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$, which becomes:

$$\min_{\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)} \left\{ -\alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \int \log(p_j) \rho_p(dp) + \mathcal{K}(\rho_p, \pi_p) \right\}.$$

Using Lemma 4.2.1 for $\mathbf{E} = \mathcal{S}_K$, $\lambda = \pi_p$ and $h(p) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \log(p_j)$, we get directly the solution:

$$\rho_p(dp) \propto \exp\left(\alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \log(p_j) \right) \pi_p(dp).$$

Optimization with respect to $\rho_j \in \mathcal{M}_1^+(\Theta)$:

Now, we fix $\omega^i \in \mathcal{S}_K$ for $i = 1, ..., n$, $\rho_p \in \mathcal{M}_1^+(\mathcal{S}_K)$ and $\rho_\ell \in \mathcal{M}_1^+(\Theta)$ for $\ell \neq j$, and we solve the program with respect to $\rho_j \in \mathcal{M}_1^+(\Theta)$, which becomes:

$$\min_{\rho_j \in \mathcal{M}_1^+(\Theta)} \left\{ -\alpha \sum_{i=1}^{n} \omega_j^i \int \log(q_{\theta_j}(X_i)) \rho_j(d\theta_j) + \mathcal{K}(\rho_j, \pi_j) \right\}.$$

Using Lemma 4.2.1 for $\mathbf{E} = \Theta$, $\lambda = \pi_j$ and $h(\theta_j) = \alpha \sum_{i=1}^{n} \omega_j^i \log(q_{\theta_j}(X_i))$, we get directly the solution:

$$\rho_j(d\theta_j) \propto \exp\left(\alpha \sum_{i=1}^{n} \omega_j^i \log(q_{\theta_j}(X_i)) \right) \pi_j(d\theta_j)$$

Application to multinomial mixture models

We simply use

$$\int \log(p_j) \rho_p(dp) = \mathbb{E}_{p \sim \rho_p}[\log(p_j)] = \psi(\phi_j) - \psi(\sum_{\ell=1}^{K} \phi_\ell),$$

$$\int \log(q_{\theta_j}(X_i)) \rho_j(d\theta_j) = \mathbb{E}_{\theta_j \sim \rho_j}[\log(\theta_{X_i,j})] = \psi(\gamma_{X_i,j}) - \psi\Big(\sum_{v=1}^{V} \gamma_{vj} \Big),$$

$$\exp\left(\alpha \sum_{i=1}^{n} \sum_{j=1}^{K} \omega_j^i \log(p_j) \right) \pi_p(p) = \prod_{j=1}^{K} p_j^{\alpha_j + \alpha \sum_{i=1}^{n} \omega_j^i - 1},$$

$$\exp\left(\alpha \sum_{i=1}^{n} \omega_j^i \log(q_{\theta_j}(X_i)) \right) \pi_j(\theta_j) = \prod_{v=1}^{V} \theta_{vj}^{\beta_v + \alpha \sum_{i=1}^{n} \omega_j^i \mathbb{1}(X_i = v) - 1}.$$

We recognize a Dirichlet distribution.

Application to Gaussian mixture models

For Gaussian mixtures, use

$$\int \log(p_j)\rho_p(dp) = \mathbb{E}_{p\sim\rho_p}[\log(p_j)] = \psi(\phi_j) - \psi(\sum_{\ell=1}^{K}\phi_\ell),$$

$$\int \log(q_{\theta_j}(X_i))\rho_j(d\theta_j) = -\frac{1}{2}\mathbb{E}_{\theta_j\sim\rho_j}[(\theta_j - X_i)^2] + \text{cst} = -\frac{1}{2}\Big\{s_j^2 + (n_j - X_i)^2\Big\} + \text{cst},$$

$$\exp\left(\alpha\sum_{i=1}^{n}\sum_{j=1}^{K}\omega_j^i\log(p_j)\right)\pi_p(p) = \prod_{j=1}^{K}p_j^{\alpha_j+\alpha\sum_{i=1}^{n}\omega_j^i-1},$$

$$\exp\left(\alpha\sum_{i=1}^{n}\omega_j^i\log(q_{\theta_j}(X_i))\right)\pi_j(\theta_j) \propto \exp\left(-\frac{\alpha}{2}\sum_{i=1}^{n}\omega_j^i(\theta_j - X_i)^2\right)\exp\left(-\frac{1}{2\mathcal{V}^2}\theta_j^2\right)$$

$$\propto \exp\left(-\frac{1/\mathcal{V}^2 + \alpha\sum_{i=1}^{n}\omega_j^i}{2}\left(\theta_j - \frac{\alpha\sum_{i=1}^{n}\omega_j^i X_i}{1/\mathcal{V}^2 + \alpha\sum_{i=1}^{n}\omega_j^i}\right)^2\right).$$

We recognize a Gaussian distribution.

Supplementary material

We provide in this supplementary material a very short simulation study. Our objective is not to compare extensively EM to CAVI as this was already done in many papers (mentioned in the main body of the paper). We just show on a low-dimensional example that the properties of VB with $\alpha = 1/2$ and $\alpha = 1$ (CAVI) are very close one to each other, and to the ones of EM.

We compare our algorithm for $\alpha = 0.5$ and $\alpha = 1$ (equivalent to CAVI) to EM algorithm for unit-variance Gaussian mixture parameters estimation. We consider 10 different unit-variance Gaussian mixtures which parameters $(p^0, \theta_1^0, \theta_2^0, \theta_3^0)$ are generated independently from a Dirichlet distribution $p^0 \sim \mathcal{D}_K(2/3, 2/3, 2/3)$ and Gaussians $\theta_j^0 \sim \mathcal{N}(0, 10)$ for $j = 1, 2, 3$. From these mixtures, we create 10 different datasets which contain 1000 i.i.d. realizations of the corresponding mixtures. We compare our algorithms using the Mean Average Error (MAE) between the estimates and the true parameters. For each dataset, we run each algorithm 5 times and keep the one with the lowest MAE in order to avoid situations where the initialization leads to a local optimum. Then, we average the resulting MAEs over the different datasets to obtain the final values of the MAE. We also record the standard deviation of the MAE over the different datasets. The following table summarizes the results. Values in brackets represent the standard deviations of the series of MAEs, and the three components are ordered in ascending values. The three estimations are comparable both in terms of estimation precision and computational efficiency :

Algorithm	p	θ_1	θ_2	θ_3
VB ($\alpha = 0.5$)	0.033 (0.020)	0.137 (0.297)	0.383 (1.108)	0.054 (0.047)
VB ($\alpha = 1$)	0.033 (0.020)	0.139 (0.207)	0.364 (0.968)	0.056 (0.039)
EM	0.033 (0.021)	0.141 (0.219)	0.364 (0.968)	0.059 (0.047)

Chapter 5

Convergence Rates of Variational Inference in Sparse Deep Learning

Variational inference is becoming more and more popular for approximating intractable posterior distributions in Bayesian statistics and machine learning. Meanwhile, a few recent works have provided theoretical justification and new insights on deep neural networks for estimating smooth functions in usual settings such as nonparametric regression. In this chapter, we show that variational inference for sparse deep learning retains the same generalization properties than exact Bayesian inference. In particular, we highlight the connection between estimation and approximation theories via the classical bias-variance trade-off and show that it leads to near-minimax rates of convergence for Hölder smooth functions. Additionally, we show that the model selection framework over the neural network architecture via ELBO maximization does not overfit and adaptively achieves the optimal rate of convergence.

5.1 Introduction

Deep learning (DL) is a field of machine learning that aims to model data using complex architectures combining several nonlinear transformations with hundreds of parameters called Deep Neural Networks (DNN) (LeCun et al., 2015; Goodfellow et al., 2016). Although generalization theory that explains why DL generalizes so well is still an open problem, it is widely acknowledged that it mainly takes advantage of large datasets containing millions of samples and a huge computing power coming from clusters of graphics processing units. Very popular architectures for deep neural networks such as the multilayer perceptron, the convolutional neural network (Lecun et al., 1998), the recurrent neural network (Rumelhart et al., 1986) or the generative adversarial network (Goodfellow et al., 2014) have shown impressive results and have enabled to perform better than humans in various important areas in artificial intelligence such as image recognition, game playing, machine translation, computer vision or natural language processing, to name a few prominent examples. An outstanding example is AlphaGo (Silver et al., 2017), an artificial intelligence developed by Google that learned to play the game of Go using deep learning techniques and even defeated the world champion in 2016.

The Bayesian approach, leading to popular methods such as Hidden Markov Models (Baum and Petrie, 1966) and Particle Filtering (Doucet and Johansen, 2009), provides a natural way to model uncertainty. Some prior distribution is put over the space of parameters and represents the prior belief as to which parameters are likely to have generated the data before any datapoint is observed. Then this prior distribution is updated using the Bayes rule when new data arrive in order to capture the more likely parameters given the observations. Unfortunately, exact Bayesian inference is computationally challenging for complex models as the normalizing constant of the posterior distribution is often intractable. In such cases, approximate inference methods such as variational inference (VI) (Jordan et al., 1999) and expectation propagation (Minka, 2001) are popular to overcome intractability in Bayesian modeling. The idea of VI is to minimize the Kullback-Leibler (KL) divergence with respect to the posterior given a set of tractable distributions, which is also equivalent to maximizing a numerical criterion called the Evidence Lower Bound (ELBO). Recent advances of VI have shown great performance in practice and have been applied to many machine learning problems (Hoffman et al., 2013; Kingma and Welling, 2013).

The Bayesian approach to learning in neural networks has a long history. Bayesian Neural Networks (BNN) have been first proposed in the 90s and widely studied since then (MacKay, 1992b; Neal, 1995). They offer a probabilistic interpretation and a measure of uncertainty for DL models. They are more robust to overfitting than classical neural networks and still achieve great performance even on small datasets. A prior distribution is put on the parameters of the network, namely the weight matrices and the bias vectors, for instance a Gaussian or a uniform distribution, and Bayesian inference is done through the likelihood specification. Nevertheless, state-of-the-art neural networks may contain millions of parameters and the form of a neural network is not adapted to exact integration, which makes the posterior distribution be intractable in practice. Modern approximate inference mainly relies on VI, with sometimes a flavor of sampling techniques. A lot of recent papers have investigated variational inference for DNNs (Hinton and van Camp, 1993; Graves, 2011; Blundell et al., 2015) to fit an approximate posterior that maximizes the evidence lower bound. For instance, Blundell et al. (2015) introduced Bayes by Backprop, one of the most famous techniques of VI applied to neural networks, which derives a fully factorized Gaussian approximation to the posterior: using the reparameterization trick (Opper and Archambeau, 2008), the gradients of ELBO towards parameters of the Gaussian approximation can be computed by backpropagation, and then be used for updates. Another point of interest in DNNs is the choice of the prior. Blundell et al. (2015) introduced a mixture of Gaussians prior on the weights, with one mixture tightly concentrated around zero, imitating the sparsity-inducing spike-and-slab prior. This offers a Bayesian alternative to the dropout regularization procedure (Srivastava et al., 2014) which injects sparsity in the network by switching off randomly some of the weights of the network. This idea goes back to David MacKay who discussed in his thesis the possibility of choosing a spike-and-slab prior over the weights of the neural network (MacKay, 1992a). More recently, Rockova and Polson (2018) introduced Spike-and-Slab Deep Learning (SS-DL), a fully Bayesian alternative to dropout for improving generalizability of deep ReLU networks.

5.1.1 Related work

Although deep learning is extremely popular, the study of generalization properties of DNNs is still an open problem. Some works have been conducted in order to investigate the theoretical properties of neural networks from different points of view. The literature developed in the past decades can be shared in three parts. First, the approximation theory wonders how well a function can be approximated by neural networks. The first studies were mostly conducted to obtain approximation guarantees for shallow neural nets with a single hidden layer (Cybenko, 1989; Barron, 1993). Since then, modern research has focused on the expressive power of depth and extended the previous results to deep neural networks with a larger number of layers (Bengio and Delalleau, 2011; Yarotsky, 2016; Petersen and Voigtländer, 2017; Grohs et al., 2019). Indeed, even though the universal approximation theorem (Cybenko, 1989) states that a shallow neural network containing a finite number of neurons can approximate any continuous function on compact sets under mild assumptions on the activation function, recent advances showed that a shallow network requires exponentially many neurons in terms of the dimension to represent a monomial function, whereas linearly many neurons are sufficient for a deep network (Rolnick and Tegmark, 2018). Second, as the objective function in deep learning is known to be nonconvex, the optimization community has discussed the landscape of the objective as well as the dynamics of some learning algorithms such as Stochastic Gradient Descent (SGD) (Baldi and Hornik, 1989; Stanford et al., 2000; Soudry and Carmon, 2016; Kawaguchi, 2016; Kawaguchi et al., 2019; Nguyen et al., 2019; Allen-Zhu et al., 2019; Du et al., 2019). Finally, the statistical learning community has investigated generalization properties of DNNs, see Barron (1994); Zhang et al. (2017); Schmidt-Hieber (2017); Suzuki (2018); Imaizumi and Fukumizu (2019); Suzuki (2019). In particular, Schmidt-Hieber (2017) and Suzuki (2019) showed that estimators in nonparametric regression based on sparsely connected DNNs with ReLU activation function and wisely chosen architecture achieve the minimax estimation rates (up to logarithmic factors) under classical smoothness assumptions on the regression function. In the same time, Bartlett et al. (2017) and Neyshabur et al. (2018) respectively used Rademacher complexity and covering number, and PAC-Bayes theory to get spectrally-normalized margin bounds for deep ReLU networks. More recently, Imaizumi and Fukumizu (2019) and Hayakawa and Suzuki (2019) showed the superiority of DNNs over linear operators in some situations when DNNs achieve the minimax rate of convergence while alternative methods fail. From a Bayesian point of view, Rockova and Polson (2018) and Suzuki (2018) studied the concentration of the posterior distribution while Vladimirova et al. (2019) investigated the regularization effect of prior distributions at the level of the units.

Such as for generalization properties of DNNs, only little attention has been put in the literature towards the theoretical properties of VI until recently. Alquier et al. (2016) studied generalization properties of variational approximations of Gibbs distributions in machine learning for bounded loss functions. Alquier and Ridgway (2017); Zhang and Gao (2017); Sheth and Khardon (2017); Bhattacharya et al. (2018); Chérief-Abdellatif and Alquier (2018); Chérief-Abdellatif (2019a); Jaiswal et al. (2019a) extended the previous guarantees to more general statistical models and studied the concentration of variational approximations of the posterior distribution, while Wang and Blei (2018) pro-

vided Bernstein-von-Mises' theorems for variational approximations in parametric models. Huggins et al. (2018); Campbell and Li (2019); Jaiswal et al. (2019b) discussed theoretical properties of variational inference algorithms based on various divergences (respectively Wasserstein and Hellinger distances, and Rényi divergence). More recently, Chérief-Abdellatif et al. (2019) presented generalization bounds for online variational inference. All these works show that under mild conditions, the variational approximation is consistent and achieves the same rate of convergence than the Bayesian posterior distribution it approximates. Note that Alquier and Ridgway (2017); Bhattacharya et al. (2018); Chérief-Abdellatif and Alquier (2018); Chérief-Abdellatif (2019a) restricted their studies to tempered versions of the posterior distribution where the likelihood is raised to an α-power ($\alpha < 1$) as it is known to require less stringent assumptions to obtain consistency and to be robust to misspecification, see respectively Bhattacharya et al. (2016) and Grünwald et al. (2017). Nevertheless, some questions remain unanswered, as the theoretical study of generalization of variational inference for deep neural networks.

5.1.2 Contributions

this chapter aims at filling the gap between theory and practice when using variational approximations for tempered Bayesian Deep Neural Networks. To the best of our knowledge, this is the first paper to present theoretical generalization error bounds of variational inference for Bayesian deep learning. Inspired by the related literature, our work is motivated by the following questions:

- Do consistency of Bayesian DNNs still hold when an approximation is used instead of the exact posterior distribution, and can we obtain the same rates of convergence than those obtained for the regular posterior distribution and frequentist estimators?
- Is it possible to obtain a nonasymptotic generalization error bound that holds for (almost) any generating distribution function and that gives a general formula?
- What about the consistency of numerical algorithms used to compute these variational approximations?
- Can we obtain new insights on the structure of the networks?

The main contribution of this chapter, a nonasymptotic generalization error bound for variational inference in sparse DL in the nonparametric regression framework, answers the first two questions. This generalization result is similar to theoretical inequalities in the seminal works of Suzuki (2018); Imaizumi and Fukumizu (2019); Rockova and Polson (2018) on generalization properties of deep neural networks, and is inspired by the general literature on the consistency of variational approximations (Alquier and Ridgway, 2017; Bhattacharya et al., 2018). In particular, it states that under the same conditions, sparse variational approximations of posterior distributions of deep neural networks are consistent at the same rate of convergence than the exact posterior.

It also raises the question of finding a relevant general definition of consistency that can be used to provide theoretical properties for the exact Bayesian DNNs distribution and

their variational approximations. Indeed, a classical criterion used to assess frequentist guarantees for Bayesian estimators is the concentration of the posterior (to the true distribution) which is defined as the asymptotic concentration of the Bayesian estimator to the true distribution (Ghosal et al., 2000). Nevertheless, posterior concentration to the true distribution only applies when the model is well specified, or at least when the model contains distributions in the neighborhood of the true distribution, which is problematic for misspecified models e.g. when the neural network does not sufficiently approximate the generating distribution. And although the posterior distribution may concentrate to the best approximation of the true distribution in KL divergence in such misspecified models, there exists pathological cases where the regular Bayesian posterior is not consistent at all, see Grünwald et al. (2017). This is the reason why we focus here on tempered posteriors which are robust to such misspecification. Therefore, we introduce in Section 5.2 a notion of consistency of a Bayesian estimator which is closely related to the notion of concentration - even stronger - and which enables a more robust formulation of generalization error bounds for variational approximations. See Appendix 5.6.1 for more details on the connection between the notions of consistency and concentration.

Then we focus on optimization aspects. We no longer assume an ideal optimization, as done for instance in Schmidt-Hieber (2017); Imaizumi and Fukumizu (2019). We address in this chapter the question of the consistency of numerical algorithms used to compute our ideal approximations. We consider an optimization error given by any algorithm and independent to the statistical error, and we show how it affects our generalization result. Our upper bound highlights the connection between the consistency of the variational approximation and the convergence of the ELBO.

We also provide insights on the structure of the network which leads to optimal rates of convergence, i.e. its depth, its width and its sparsity. Indeed, in our first generalization error bound, the structure of the network is ideally tuned for some choice of the generating function, and we show how to choose such a structure. Nevertheless, the characteristics of the regression function may be unknown, e.g. we may know that the regression function is Hölder continuous but we ignore its level of smoothness. We propose here an automated method for choosing the architecture of the network. We introduce a classical model selection framework based on the ELBO criterion (Chérief-Abdellatif, 2019a), and we show that the variational approximation associated with the selected structure does not overfit and adaptively achieves the optimal rate of convergence even without any oracle information.

The rest of this chapter is organized as follows. Section 5.2 introduces the notations and the framework that will be considered in the paper, and presents sparse spike-and-slab variational inference for deep neural networks. Section 5.3 provides theoretical generalization error bounds for variational approximations of DNNs and shows the optimality of the method for estimating Hölder smooth functions. Finally, insights on the choice of the architecture of the network are given in Section 5.4 via the ELBO maximization framework. All the proofs are deferred to the appendix.

5.2 Sparse deep variational inference

Let us introduce the notations and the statistical framework we adopt in this chapter. For any vector $x = (x_1, ..., x_d) \in [-1, 1]^d$ and any real-valued function f defined on $[-1, 1]^d$, $d > 0$, we denote:

$$\|x\|_\infty = \max_{1 \leq i \leq d} |x_i| \quad , \quad \|f\|_2 = \left(\int f^2 \right)^{1/2} \quad \text{and} \quad \|f\|_\infty = \sup_{y \in [-1,1]^d} |f(y)|.$$

For any $\mathbf{k} \in \{0, 1, 2, ...\}^d$, we define $|\mathbf{k}| = \sum_{i=1}^d k_i$ and the mixed partial derivatives when all partial derivatives up to order $|\mathbf{k}|$ exist:

$$D^{\mathbf{k}} f(x) = \frac{\partial^{|\mathbf{k}|} f}{\partial^{k_1} x_1 ... \partial^{k_d} x_d}(x).$$

We also introduce the notion of β-Hölder continuity for $\beta > 0$. We denote $\lfloor \beta \rfloor$ the largest integer strictly smaller than β. Then f is said to be β-Hölder continuous (Tsybakov, 2008) if all partial derivatives up to order $\lfloor \beta \rfloor$ exist and are bounded, and if:

$$\|f\|_{\mathcal{C}_\beta} := \max_{|\mathbf{k}| \leq \lfloor \beta \rfloor} \|D^{\mathbf{k}} f\|_\infty + \max_{|\mathbf{k}| = \lfloor \beta \rfloor} \sup_{x, y \in [-1,1]^d, x \neq y} \frac{|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(y)|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} < +\infty.$$

$\|f\|_{\mathcal{C}_\beta}$ is the norm of the Hölder space $\mathcal{C}_\beta = \{f / \|f\|_{\mathcal{C}_\beta} < +\infty\}$.

5.2.1 Nonparametric regression

We consider the nonparametric regression framework. We have a collection of random variables $(X_i, Y_i) \in [-1, 1]^d \times \mathbb{R}$ for $i = 1, ..., n$ which are independent and identically distributed (i.i.d.) with the generating process:

$$\begin{cases} X_i \sim \mathcal{U}([-1, 1]^d), \\ Y_i = f_0(X_i) + \zeta_i \end{cases}$$

where $\mathcal{U}([-1, 1]^d)$ is the uniform distribution on the interval $[-1, 1]^d$, $\zeta_1, ..., \zeta_n$ are i.i.d. Gaussian random variables with mean 0 and known variance σ^2, and $f_0 : [-1, 1]^d \to \mathbb{R}$ is the true unknown function. For instance, the true regression function f_0 may belong to the set $\mathcal{C}_\beta$ of Hölder functions with level of smoothness β.

5.2.2 Deep neural networks

We call deep neural network any map $f_\theta : \mathbb{R}^d \to \mathbb{R}$ defined recursively as follows:

$$\begin{cases} x^{(0)} := x, \\ x^{(\ell)} := \rho(A_\ell x^{(\ell-1)} + b_\ell) \quad \text{for } \ell = 1, ..., L - 1, \\ f_\theta(x) := A_L x^{(L-1)} + b_L \end{cases}$$

where $L \geq 3$. ρ is an activation function acting componentwise. For instance, we can choose the ReLU activation function $\rho(u) = \max(u, 0)$. Each $A_\ell \in \mathbb{R}^{D_\ell \times D_{\ell-1}}$ is a weight matrix such that its (i, j) coefficient, called edge weight, connects the j-th neuron of the $(\ell - 1)$-th layer to the i-th neuron of the ℓ-th layer, and each $b_\ell \in \mathbb{R}^{D_\ell}$ is a shift vector such that its i-th coefficient, called node vector, represents the weight associated with the i-th node of layer ℓ. We set $D_0 = d$ the number of units in the input layer, $D_L = 1$ the number of units in the output layer and $D_\ell = D$ the number of units in the hidden layers. The architecture of the network is characterized by its number of edges S, i.e. the total number of nonzero entries in matrices A_ℓ and vectors b_ℓ, its number of layers $L \geq 3$ (excluding the input layer), and its width $D \geq 1$. We have $S \leq T$ where $T = \sum_{\ell=1}^{L} D_\ell(D_{\ell-1} + 1)$ is the total number of coefficients in a fully connected network. By now, we consider that S, L and D are fixed, and $d = \mathcal{O}(1)$ as $n \to +\infty$. In particular, we assume that $d \leq D$, which implies that $T \leq LD(D + 1)$. We also suppose that the absolute values of all coefficients are upper bounded by some positive constant $B \geq 2$. This boundedness assumption will be relaxed in the appendix, see Appendix 5.6.7. Then, the parameter of a DNN is $\theta = \{(A_1, b_1), ..., (A_L, b_L)\}$, and we denote $\Theta_{S,L,D}$ the set of all possible parameters. We will also alternatively consider the stacked coefficients parameter $\theta = (\theta_1, ..., \theta_T)$.

5.2.3 Bayesian modeling

We adopt a Bayesian approach, and we place a spike-and-slab prior π (Castillo et al., 2015) over the parameter space $\Theta_{S,L,D}$ (equipped with some suited sigma-algebra) that is defined hierarchically. The spike-and-slab prior is known to be a relevant alternative to dropout for Bayesian deep learning, see Rockova and Polson (2018). First, we sample a vector of binary indicators $\gamma = (\gamma_1, ..., \gamma_T) \in \{0, 1\}^T$ uniformly among the set $\mathcal{S}_T^S$ of T-dimensional binary vectors with exactly S nonzero entries, and then given γ_t for each $t = 1, ..., T$, we put a spike-and-slab prior on θ_t that returns 0 if $\gamma_t = 0$ and a random sample from a uniform distribution on $[-B, B]$ otherwise:

$$\begin{cases} \gamma \sim \mathcal{U}(\mathcal{S}_T^S), \\ \theta_t | \gamma_t \sim \gamma_t \, \mathcal{U}([-B, B]) + (1 - \gamma_t)\delta_{\{0\}}, \quad t = 1, ..., T \end{cases}$$

where $\delta_{\{0\}}$ is a point mass at 0 and $\mathcal{U}([-B, B])$ is a uniform distribution on $[-B, B]$. We recall that the sparsity level S is fixed here and that this assumption will be relaxed in Section 5.4.

Remark 5.2.1. *We consider uniform distributions for simplicity as in similar works (Rockova and Polson, 2018; Suzuki, 2018), but Gaussian distributions can be used as well when working on an unbounded parameter set $\Theta_{S,L,D}$, see Theorem 5.6.2 in Appendix 5.6.7.*

Then we define the tempered posterior distribution $\pi_{n,\alpha}$ on parameter $\theta \in \Theta_{S,L,D}$ using prior π for any $\alpha \in (0, 1)$:

$$\pi_{n,\alpha}(d\theta) \propto \exp\left(-\frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2 \right) \pi(d\theta),$$

which is a slight variant of the definition of the regular Bayesian posterior (for which $\alpha = 1$). This distribution is known to be easier to sample from, to require less stringent assumptions to obtain concentration, and to be robust to misspecification, see respectively Behrens et al. (2012), Bhattacharya et al. (2016) and Grünwald et al. (2017).

5.2.4 Sparse variational inference

The variational Bayes approximation $\widetilde{\pi}_{n,\alpha}$ of the tempered posterior is defined as the projection (with respect to the Kullback-Leibler divergence) of the tempered posterior onto some set $\mathcal{F}_{S,L,D}$:

$$\widetilde{\pi}_{n,\alpha} = \arg \min_{q \in \mathcal{F}_{S,L,D}} \mathrm{KL}(q \| \pi_{n,\alpha}).$$

which is equivalent to:

$$\widetilde{\pi}_{n,\alpha} = \arg \min_{q \in \mathcal{F}_{S,L,D}} \left\{ \frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} \int (Y_i - f_\theta(X_i))^2 q(d\theta) + \mathrm{KL}(q \| \pi) \right\} \tag{5.1}$$

where the function inside the argmin operator in (5.1) is the opposite of the evidence lower bound $\mathcal{L}_n(q)$.

We choose a sparse spike-and-slab variational set $\mathcal{F}_{S,L,D}$ - see for instance Tonolini et al. (2019) - which can be seen as an extension of the popular mean-field variational set with a dependence assumption specifying the number of active neurons. The mean-field approximation is based on a decomposition of the space of parameters $\Theta_{S,L,D}$ as a product $\theta = (\theta_1, ..., \theta_T)$ and consists in compatible product distributions on each parameter θ_t, $t = 1, ..., T$. Here, we fit a distribution in the family that matches the prior: we first choose a distribution π_γ on the set $\mathcal{S}_T^S$ that selects a T-dimensional binary vector γ with S nonzero entries, and then we place a spike-and-slab variational approximation on each θ_t given γ_t:

$$\begin{cases} \gamma \sim \pi_\gamma, \\ \theta_t | \gamma_t \sim \gamma_t \, \mathcal{U}([l_t, u_t]) + (1 - \gamma_t) \delta_{\{0\}} \quad \text{for each } \; t = 1, ..., T \end{cases}$$

where $-1 \leq l_t \leq u_t \leq 1$, with the distribution π_γ and the intervals $[l_t, u_t]$, $t = 1, ..., T$ as the hyperparameters of the variational set $\mathcal{F}_{S,L,D}$. In particular, if we choose a deterministic $\pi_\gamma = \delta_{\{\gamma'\}}$ with $\gamma' \in \mathcal{S}_T^S$, then we will obtain a parametric mean-field approximation. See Section 6.6 of the PhD thesis of Gal (2016) for a more detailed discussion on the connection between Gaussian mean-field and sparse spike-and-slab posterior approximations.

The generalization error of the tempered posterior $\pi_{n,\alpha}$ and of its variational approximation $\widetilde{\pi}_{n,\alpha}$ is the expected average of the squared L_2-distance to the true generating function over the Bayesian estimator:

$$\mathbb{E}\left[\int \| f_\theta - f_0 \|_2^2 \pi_{n,\alpha}(d\theta) \right] \quad \text{and} \quad \mathbb{E}\left[\int \| f_\theta - f_0 \|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta) \right].$$

We say that a Bayesian estimator is consistent at rate $r_n \to 0$ if its generalization error is upper bounded by r_n. Notice that consistency of the Bayesian estimator implies concentration to f_0. Again, see Appendix 5.6.1 for the connection between these two notions.

5.3 Generalization of variational inference for neural networks

The first result of this section is an extension of the result of Rockova and Polson (2018) on the Bayesian distribution for Hölder regression functions. Indeed, we provide a concentration result on the posterior distribution for the expected L_2-distance instead of the empirical L_2-distance, which enables generalization instead of reconstruction on the training datapoints. This result is then extended again to the variational approximation for our definition of consistency: we show that we can still achieve near-optimality using an approximation of the posterior without any additional assumption. Finally, we explain how we can incorporate optimization error in our generalization results.

5.3.1 Concentration of the posterior

Rockova and Polson (2018) gives the first posterior concentration result for deep ReLU networks when estimating Hölder smooth functions in nonparametric regression with empirical L_2-distance. The authors highlight the flexibility of DNNs over other methods for estimating β-Hölder smooth functions as there is a large range of values of the level of smoothness β for which one can obtain concentration, e.g. $0 < \beta < d$ for a DNN against $0 < \beta < 1$ for a Bayesian tree.

The following theorem provides the concentration of the tempered posterior distribution $\pi_{n,\alpha}$ for deep ReLU neural networks when using the expected L_2-distance for some suitable architecture of the network:

Theorem 5.3.1. *Let us assume that $\alpha \in (0,1)$, that f_0 is β-Hölder smooth with $0 < \beta < d$ and that the activation function is ReLU. We consider the architecture of Rockova and Polson (2018) for some positive constant C_D independent of n:*

$$L = 8 + (\lfloor \log_2 n \rfloor + 5)(1 + \lceil \log_2 d \rceil),$$

$$D = C_D \lfloor n^{\frac{d}{2\beta+d}} / \log n \rfloor,$$

$$S \leq 94 d^2 (\beta+1)^{2d} D (L + \lceil \log_2 d \rceil).$$

Then the tempered posterior distribution $\pi_{n,\alpha}$ concentrates at the minimax rate $r_n = n^{\frac{-2\beta}{2\beta+d}}$ up to a (squared) logarithmic factor for the expected L_2-distance in the sense that:

$$\pi_{n,\alpha}\left(\theta \in \Theta_{S,L,D} \,\middle/\, \|f_\theta - f_0\|_2^2 > M_n \cdot n^{\frac{-2\beta}{2\beta+d}} \cdot \log^2 n \right) \xrightarrow[n \to +\infty]{} 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

In order to prove Theorem 5.3.1, we actually have to check that the so-called *prior mass* condition is satisfied:

$$\pi\left(\theta \in \Theta_{S,L,D} \,\middle/\, \|f_\theta - f_0\|_2^2 \leq r_n \right) \geq e^{-n r_n}. \tag{5.2}$$

This assumption, introduced in Ghosal et al. (2000) in order to obtain the concentration of the regular posterior distribution states that the prior must give enough mass to some neighborhood of the true parameter. As shown in Bhattacharya et al. (2016), this condition is even sufficient for tempered posteriors. Actually, this inequality was first stated using the KL divergence instead of the expected L_2-distance (see Condition 2.4 in Theorem 2.1 in Ghosal et al. (2000)), but the KL metric is equivalent to the squared L_2-metric in regression problems with Gaussian noise. This prior mass condition gives us the rate of convergence of the tempered posterior $r_n = n^{\frac{-2\beta}{2\beta+d}}$ (up to a squared logarithmic factor) which is known to be optimal when estimating β-Hölder smooth functions (Tsybakov, 2008). Note that the $\log^2 n$ term is common in the theoretical deep learning literature (Imaizumi and Fukumizu, 2019; Suzuki, 2019; Schmidt-Hieber, 2017).

Remark 5.3.1. *The number of parameters of order* $n^{\frac{2d}{2\beta+d}}/\log n \in [n^{2/3}/\log(n), n^2/\log(n)]$ *is high compared to standard machine learning methods, which may lead to overfitting and hence prevent the procedure from achieving the minimax rate of convergence. The sparsity parameter* S *which gives a network with a small number of nonzero parameters along with the spike-and-slab prior help us tackle this issue and obtain optimal rates of convergence (up to logarithmic factors).*

5.3.2 A generalization error bound

The result we state in this subsection applies to a wide range of activation functions, including the popular ReLU activation and the identity map:

Assumption 5.3.1. *In the following, we assume that the activation function* ρ *is* 1*-Lispchitz continuous (with respect to the aboluste value) and is such that for any* $x \in \mathbb{R}$, $|\rho(x)| \leq |x|$.

We do not assume any longer that the regression function is β-Hölder and we consider any structure (S, L, D). The following theorem gives a generalization error bound when using variational approximations instead of exact tempered posteriors for DNNs. The proof is given in Appendix 5.6.2 and is based on PAC-Bayes theory (Massart, 2007; Guedj, 2019):

Theorem 5.3.2. *For any* $\alpha \in (0, 1)$,

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha} \inf_{\theta^* \in \Theta_{S,L,D}} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D}, \quad (5.3)$$

with

$$r_n^{S,L,D} = \frac{LS}{n}\log(BD) + \frac{2S}{n}\log(BLD) + \frac{S}{n}\log\left(7dL\max\left(\frac{n}{S}, 1\right)\right).$$

The oracle inequality (5.3) ensures consistency of variational Bayes for estimating neural networks and provides the associated rate of convergence given the structure (S, L, D).

Indeed, if f_0 is a neural network with structure (S, L, D), then the infimum term on the right hand side of the inequality vanishes and we obtain a rate of convergence of order

$$r_n^{S,L,D} \sim \max\left(\frac{S\log(nL/S)}{n}, \frac{LS\log D}{n}\right),$$

which underlines a linear dependence on the number of layers and the sparsity. In fact, this rate of convergence is determined by the *extended prior mass condition* (Alquier and Ridgway, 2017; Chérief-Abdellatif and Alquier, 2018; Chérief-Abdellatif, 2019a), which requires that in addition to the previous prior mass condition of Ghosal et al. (2000) and Bhattacharya et al. (2016), the variational set $\mathcal{F}_{S,L,D}$ must contain probability distributions q that are concentrated enough around the true generating function f_0. One of the main findings of Theorem 5.3.2 is that our choice of the sparse spike-and-slab variational set $\mathcal{F}_{S,L,D}$ is rich enough and that both conditions are actually similar and lead to the same rate of convergence. Hence, the rate of convergence is the one that satisfies the prior mass condition (5.2). In particular, as the prior distribution is uniform over the parameter space, the negative logarithm of the prior mass of the neighborhood of the true regression function in Equation (5.2) is a local covering entropy, that is the logarithm of the number of $r_n^{S,L,D}$-balls needed to cover a neighborhood of the true regression function. Especially, it has been shown in previous studies that this local covering entropy fully characterizes the rate of convergence of the empirical risk minimizer for DNNs (Schmidt-Hieber, 2017; Suzuki, 2019). The rate $r_n^{S,L,D}$ we obtain in this work is exactly of the same order than the upper bound on the covering entropy number given in Lemma 5 in Schmidt-Hieber (2017) and in Lemma 3 in Suzuki (2019) which derive rates of convergence for the empirical risk minimizer using different proof techniques. Note that replacing a uniform by a Gaussian in the prior and variational distributions leads to the same rate of convergence, see Appendix 5.6.7.

Nevertheless, deep neural networks are mainly used for their computational efficiency and their ability to approach complex functions, which makes the task of estimating a neural network not so popular in machine learning. As said earlier, Imaizumi and Fukumizu (2019) used neural networks for estimating non-smooth functions. In such a context where the neural network model is misspecified, our generalization error bound is robust and still holds, and satisfies the best possible balance between bias and variance.

Indeed, the upper bound on the generalization error on the right-hand-side of (5.3) is mainly divided in two parts: the approximation error of f_0 by a DNN f_{θ^*} in $\Theta_{S,L,D}$ (i.e. the bias) and the estimation error $r_n^{S,L,D}$ of a neural network f_{θ^*} in $\Theta_{S,L,D}$ (i.e. the variance). For instance, even if the generalization power is decreasing linearly with respect to the number of layers compared to the logarithmic dependence on the width due to the variance term, this effect is compensated by the benefits of depth in the approximation theory of deep learning. Then, as there exists relationships between the bias/the variance and the architecture of a neural network (respectively due to the approximation theory/the form of $r_n^{S,L,D}$), Theorem 5.3.2 gives both a general formula for deriving rates of convergence for variational approximations and insight on the way to choose the architecture. We choose the architecture that minimizes the right-hand-side of (5.3), which can lead to minimax estimators for smooth functions. It also connects the approximation and estimation theories following previous studies. This was done for

instance by Schmidt-Hieber (2017); Suzuki (2019); Imaizumi and Fukumizu (2019) who exploited the effectiveness of ReLU activation function in terms of approximation ability (Yarotsky, 2016; Petersen and Voigtländer, 2017) for Hölder/Besov smooth and piecewise smooth generating functions.

Now we illustrate Theorem 5.3.2 on Hölder smooth functions. The following result shows that the variational approximation achieves the same rate of convergence than the posterior distribution it approximates, and even the minimax rate of convergence if the architecture is well chosen. We present both consistency and concentration results.

Corollary 5.3.3. *Let us fix $\alpha \in (0,1)$. We consider the ReLU activation function. Assume that f_0 is β-Hölder smooth with $0 < \beta < d$. Then with L, D and S defined as in Theorem 5.3.1, the variational approximation of the tempered posterior distribution $\widetilde{\pi}_{n,\alpha}$ is consistent and hence concentrates at the minimax rate $r_n = n^{\frac{-2\beta}{2\beta+d}}$ (up to a squared logarithmic factor):*

$$\widetilde{\pi}_{n,\alpha}\left(\theta \in \Theta_{S,L,D} \,\middle/\, \|f_\theta - f_0\|_2^2 > M_n \cdot n^{\frac{-2\beta}{2\beta+d}} \cdot \log^2 n\right) \xrightarrow[n\to+\infty]{} 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

5.3.3 Optimization error

In this subsection, we discuss the effect of an optimization error that is independent on the previous statistical error. Indeed, in the variational Bayes community, people use approximate algorithms in practice to solve the optimization problem (5.1) when the model is non-conjugate, i.e. the VB solution is not available in closed-form. This is the case here when considering a sparse spike-and-slab variational approximation in $\mathcal{F}_{S,L,D}$ for DNNs with hyperparameters $\phi = (\pi_\gamma, (\phi_t)_{1\le t\le T})$ and an algorithm that gives a sequence of hyperparameters $(\phi^k)_{k\ge1}$ and associated variational approximations $(\widetilde{\pi}^k_{n,\alpha})_{k\ge1}$. The following theorem gives a statistical guarantee for any approximation $\widetilde{\pi}^k_{n,\alpha}$, $k \ge 1$:

Theorem 5.3.4. *For any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}^k_{n,\alpha}(d\theta)\right] \le \frac{2}{1-\alpha}\inf_{\theta^*}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D} + \frac{2\sigma^2}{\alpha(1-\alpha)} \cdot \frac{\mathbb{E}[\mathcal{L}_n^* - \mathcal{L}_n^k]}{n},$$

where $\mathcal{L}_n^*$ is the maximum of the evidence lower bound i.e. the ELBO evaluated at $\widetilde{\pi}_{n,\alpha}$, while $\mathcal{L}_n^k$ is the ELBO evaluated at $\widetilde{\pi}^k_{n,\alpha}$.

We establish a clear connection between the convergence (in mean) of the ELBO $\mathcal{L}_n^k$ to $\mathcal{L}_n^*$ and the consistency of our algorithm $\widetilde{\pi}^k_{n,\alpha}$. Indeed, as soon as the ELBO $\mathcal{L}_n^k$ converges at rate $c_{k,n}$, then our variational approximation $\widetilde{\pi}^k_{n,\alpha}$ is consistent at rate:

$$\max\left(\frac{c_{k,n}}{n}, \frac{S\log(nL/S)}{n}, \frac{SL\log D}{n}\right).$$

In particular, as soon as k is such that $c_{k,n} \leq \max(S \log n, S \log D)$, then we obtain consistency of $\widetilde{\pi}^k_{n,\alpha}$ at rate $r_n^{S,L,D}$, i.e. $\widetilde{\pi}^k_{n,\alpha}$ and $\widetilde{\pi}_{n,\alpha}$ have the same rate of convergence.

However, deriving the convergence of the ELBO is a hard task. For instance, when considering a simple Gaussian mean-field approximation without sparsity, the variational objective $\mathcal{L}_n$ can be maximized using either stochastic (Graves, 2011; Blundell et al., 2015) or natural gradient methods (Khan et al., 2018) on the parameters of the Gaussian approximation. The convergence of the ELBO is often met in practice (Buchholz et al., 2018; Mishkin et al., 2018) and the recent work of Osawa et al. (2019) even showed that Bayesian deep learning enables practical deep learning and matches the performance of standard methods while preserving benefits of Bayesian principles. Nevertheless, the objective is nonconvex and hence it is difficult to prove the convergence to a global maximum in theory. Some recent papers studied global convergence properties of gradient descent algorithms for frequentist classification and regression losses (Du et al., 2019; Allen-Zhu et al., 2019) that we may extend to gradient descent algorithms for the ELBO objective such as Variational Online Gauss Newton or Vadam (Khan et al., 2018; Osawa et al., 2019).

Another point is to develop and study more complex algorithms than simple gradient descent that deal with spike-and-slab sparsity-inducing variational inference, as for instance Titsias and Lázaro-Gredilla (2011) did for multi-task and multiple kernel learning. Also, Louizos et al. (2018) connected sparse spike-and-slab variational inference with L_0-norm regularization for neural networks and proposed a solution to the intractability of the L_0-penalty term through the use of non-negative stochastic gates, while Bellec et al. (2018) proposed an algorithm preserving sparsity during training. Nevertheless, these optimization concerns fall beyond the scope of this chapter and are left for further research.

5.4 Architecture design via ELBO maximization

We saw in Section 5.3 that the choice of the architecture of the neural network is crucial and can lead to faster convergence and better approximation. In this section, we formulate the architecture design of DNNs as a model selection problem and we investigate the ELBO maximization strategy which is very popular in the variational Bayes community. This approach is different from Rockova and Polson (2018) which is fully Bayesian and treats the parameters of the network architecture, namely the depth, the width and the sparsity, as random variables. We show that the ELBO criterion does not overfit and is adaptive: it provides a variational approximation with the optimal rate of convergence, and it does not require the knowledge of the unknown aspects of the regression function f_0 (e.g. the level of smoothness for smooth functions) to select the optimal variational approximation.

We denote $\mathcal{M}_{S,L,D}$ the statistical model associated with the parameter set $\Theta_{S,L,D}$. We consider a countable number of models, and we introduce prior beliefs $\pi_{S,L,D}$ over the sparsity, the depth and the width of the network, that can be defined hierarchically and that are known beforehand. For instance, the prior beliefs can be chosen such that $\pi_L = 2^{-L}$, $\pi_{D|L}$ follows a uniform distribution over $\{d, ..., \max(e^L, d)\}$ given L, and $\pi_{S|L,D}$

a uniform distribution over $\{1, ..., T\}$ given L and D (we recall that T is the number of coefficients in a fully connected network). This particular choice is sensible as it allows to consider any number of hidden layers and (at most) an exponentially large width with respect to the depth of the network. We still consider spike-and-slab priors on $\theta_{S,L,D} \in \Theta_{S,L,D}$ given model $\mathcal{M}_{S,L,D}$.

Each tempered posterior associated with model $\mathcal{M}_{S,L,D}$ is denoted $\pi_{n,\alpha}^{S,L,D}$. We recall that the variational approximation $\widetilde{\pi}_{n,\alpha}^{S,L,D}$ associated with model $\mathcal{M}_{S,L,D}$ is defined as the distribution into the variational set $\mathcal{F}_{S,L,D}$ that maximizes the Evidence Lower Bound:

$$\widetilde{\pi}_{n,\alpha}^{S,L,D} = \arg \max_{q^{S,L,D} \in \mathcal{F}_{S,L,D}} \mathcal{L}_n(q^{S,L,D}).$$

We will simply denote in the following $\mathcal{L}_n^*(S, L, D)$ the closest approximation to the log-evidence i.e., the value of the ELBO evaluated at its maximum:

$$\mathcal{L}_n^*(S, L, D) = \mathcal{L}_n(\widetilde{\pi}_{n,\alpha}^{S,L,D}).$$

The model selection criterion we use here to select the architecture of the network is a slight penalized variant of the classical ELBO criterion (Blei et al., 2017) with strong theoretical guarantees (Chérief-Abdellatif, 2019a) :

$$(\widehat{S}, \widehat{L}, \widehat{D}) = \arg \max_{S,L,D} \left\{ \mathcal{L}_n^*(S, L, D) - \log\left(\frac{1}{\pi_{S,L,D}}\right) \right\}.$$

For any choice of the prior beliefs $\pi_{S,L,D}$, compute the ELBO for each model $\mathcal{M}_{S,L,D}$ using an algorithm that will converge to $\mathcal{L}_n^*(S, L, D)$ and choose the architecture that maximizes the penalized ELBO criterion. It is possible to restrict to a finite number of layers in practice (for instance, a factor of n or $\log n$).

The following theorem shows that this ELBO criterion leads to a variational approximation with the optimal rate of convergence:

Theorem 5.4.1. *For any $\alpha \in (0, 1)$,*

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}^{\widehat{S},\widehat{L},\widehat{D}}(d\theta)\right] \leq \inf_{S,L,D} \left\{ \frac{2}{1-\alpha} \inf_{\theta^* \in \Theta_{S,L,D}} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D} + \frac{2\sigma^2}{\alpha(1-\alpha)} \frac{\log(\frac{1}{\pi_{S,L,D}})}{n} \right\}.$$

This inequality shows that as soon as the complexity term $\log(1/\pi_{S,L,D})/n$ that reflects the prior beliefs is lower than the effective rate of convergence that balances the accuracy and the estimation error $r_n^{S,L,D}$, the selected variational approximation adaptively achieves the best possible rate. For instance, it leads to (near-)minimax rates for Hölder smooth functions and selects the optimal architecture even without the knowledge of β, which was required in the previous section. Note that for the previous choice of prior beliefs $\pi_L = 2^{-L}$, $\pi_{D|L} = 1/(\max(e^L, d) - d + 1)$, $\pi_{S|L,D} = 1/T$, we get:

$$\frac{\log(\frac{1}{\pi_{S,L,D}})}{n} \leq \frac{2\log(D+1) + \log L + \max(L, \log d) + L\log 2}{n}$$

that is lower than $r_n^{S,L,D}$ (up to a factor) and hence the ELBO criterion does not overfit.

5.5 Discussion

In this chapter, we provided theoretical justifications for neural networks from a Bayesian point of view using sparse variational inference. We derived new generalization error bounds and we showed that sparse variational approximations of DNNs achieve (near-)minimax optimality when the regression function is Hölder smooth. All our results directly imply concentration of the approximation of the posterior distribution. We also proposed an automated method for selecting an architecture of the network with optimal consistency guarantees via the ELBO maximization framework.

We think that one of the main challenges here is the design of new computational algorithms for spike-and-slab deep learning in the wake of the work of Titsias and Lázaro-Gredilla (2011) for multi-task and multiple kernel learning, or those of Louizos et al. (2018) and Bellec et al. (2018). In the latter paper, the authors designed an algorithm for training deep networks while simultaneously learning their sparse connectivity allowing for fast and computationally efficient learning, whereas most approaches have focused on compressing already trained neural networks.

In the same time, a future point of interest is the study of the global convergence of these approximate algorithms in nonconvex settings i.e. study of the theoretical convergence of the ELBO. This work was conducted for frequentist gradient descent algorithms (Allen-Zhu et al., 2019; Du et al., 2019). Such studies should be investigated for Bayesian gradient descents, as well as for algorithms that preserve the sparsity of the network during training.

5.6 Proofs and additional results

5.6.1 Connection between concentration and consistency

In this appendix, we show the connection between the notions of *consistency* and *concentration*.

The Bayesian estimator ρ (e.g. the tempered posterior $\pi_{n,\alpha}$ or its variational approximation $\widetilde{\pi}_{n,\alpha}$) is said to be consistent if its generalization error goes to zero as $n \to +\infty$:

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \rho(d\theta)\right] \xrightarrow[n\to+\infty]{} 0.$$

We say that the Bayesian estimator ρ concentrates at rate r_n (Ghosal et al., 2000) if in probability (with respect to the random variables distributed according to the generating process), the estimator concentrates asymptotically around the true distribution as $n \to +\infty$, i.e.:

$$\rho\left(\theta \in \Theta_{S,L,D} \,\middle/\, \|f_\theta - f_0\|_2^2 > M_n r_n\right) \xrightarrow[n\to+\infty]{} 0.$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

The consistency of the Bayesian distribution ρ at rate r_n implies its concentration at rate r_n. Indeed, if we we assume that ρ is consistent at rate r_n, i.e.:

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \rho(d\theta)\right] \leq r_n,$$

then, using Markov's inequality for any $M_n \to +\infty$ as $n \to +\infty$:

$$\mathbb{E}\left[\rho\left(\theta \in \Theta_{S,L,D} \,\middle/\, \|f_\theta - f_0\|_2^2 > M_n r_n\right)\right] \leq \frac{\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \rho(d\theta)\right]}{M_n r_n} \leq \frac{r_n}{M_n r_n} = \frac{1}{M_n} \to 0.$$

Hence, we have the convergence in mean of $\rho\Big(\theta \in \Theta_{S,L,D} \,\Big/\, \|f_\theta - f_0\|_2^2 > M_n r_n\Big)$ to 0, and then the convergence in probability of $\rho\Big(\theta \in \Theta_{S,L,D} \,\Big/\, \|f_\theta - f_0\|_2^2 > M_n r_n\Big)$ to 0, i.e. the concentration of ρ to f_0 at rate r_n.

5.6.2 Proof of Theorem 5.3.2

The structure of the proof of Theorem 5.3.2 is composed of three main steps. The first one consists in obtaining the general shape of the inequality using PAC-Bayes inequalities, and the two others in finding a rate that satisfies the extended prior mass condition.

First step: *we obtain the general inequality*

We start from inequality 2.6 in Alquier and Ridgway (2017) that provides an upper bound on the generalization error but in α-Rényi divergence. We denote P^0 the generating distribution of any (X_i, Y_i) and P_θ the distribution characterizing the model. Then, for any $\alpha \in (0, 1)$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}(d\theta)\right] \leq \inf_{q \in \mathcal{F}_{S,L,D}} \left\{\frac{\alpha}{1-\alpha}\int \mathrm{KL}(P^0, P_\theta) q(\mathrm{d}\theta) + \frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\}.$$

Moreover, the α-Rényi divergence is equal to $D_\alpha(P_\theta, P^0) = \frac{\alpha}{2\sigma^2}\|f_\theta - f_0\|_2^2$ and the KL divergence is $\mathrm{KL}(P^0\|P_\theta) = \frac{1}{2\sigma^2}\|f_\theta - f_0\|_2^2$, and for any θ^*, $\|f_\theta - f_0\|_2^2 \leq 2\|f_\theta - f_{\theta^*}\|_2^2 + 2\|f_{\theta^*} - f_0\|_2^2$. Hence, for any $\theta^* \in \Theta_{S,L,D}$:

$$\begin{aligned}
&\mathbb{E}\left[\int \frac{\alpha}{2\sigma^2}\|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] \\
&\quad \leq \frac{\alpha}{1-\alpha}\frac{2}{2\sigma^2}\|f_{\theta^*} - f_0\|_2^2 + \inf_{q \in \mathcal{F}_{S,L,D}} \left\{\frac{\alpha}{1-\alpha}\int \frac{2}{2\sigma^2}\|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) + \frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\},
\end{aligned}$$

i.e. for any $\theta^* \in \Theta_{S,L,D}$,

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right]$$

$$\leq \frac{2}{1-\alpha}\|f_{\theta^*} - f_0\|_2^2 + \inf_{q\in\mathcal{F}_{S,L,D}}\left\{\frac{2}{1-\alpha}\int \|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) + \frac{2\sigma^2}{\alpha}\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\}.$$

From now on, the rest of the proof consists in finding a distribution $q_n^* \in \mathcal{F}_{S,L,D}$ that satisfies for $\theta^* = \arg\min_{\theta\in\Theta_{S,L,D}} \|f_\theta - f_0\|_2$ the extended prior mass condition, i.e. that satisfies both:

$$\int \|f_\theta - f_{\theta^*}\|_2^2 q_n^*(\mathrm{d}\theta) \leq r_n \tag{5.4}$$

and

$$\mathrm{KL}(q_n^*\|\pi) \leq n r_n \tag{5.5}$$

with $r_n = \frac{SL}{n}\log(BD) + \frac{S}{n}\log(BL(D+1)^2) + \frac{S}{2n}\log\left(\frac{4n}{S}\left\{3+(d+2)^2L^2\right\}\right)$ that is smaller than $r_n^{S,L,D}$ as $3+(x+2)^2L^2 \leq 10x^2L^2$ for $x \geq 1$ and $L \geq 3$. This will lead to:

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha}\inf_{\theta^*\in\Theta_{S,L,D}}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right)r_n^{S,L,D}.$$

Second step: *we prove Inequality (5.4)*

To begin with, we define the loss of the ℓ^{th} layer of the neural network f_θ:

$$r_\ell(\theta) = \sup_{x\in[-1,1]^d}\sup_{1\leq i\leq D}|f_\theta^\ell(x)_i - f_{\theta^*}^\ell(x)_i|$$

where f_θ^ℓs are defined as the partial networks:

$$\begin{cases} f_\theta^0(x) := x, \\ f_\theta^\ell(x) := \rho(A_\ell f_\theta^{\ell-1}(x) + b_\ell) \quad \text{for } \ell = 1, ..., L. \end{cases}$$

We also define the loss of the output layer:

$$r_\ell(\theta) = \sup_{x\in[-1,1]^d}|f_\theta^L(x) - f_{\theta^*}^L(x)| = \sup_{x\in[-1,1]^d}|f_\theta(x) - f_{\theta^*}(x)|.$$

We will prove by induction that for any $\ell = 1, ..., L$:

$$r_\ell(\theta) \leq (BD)^\ell\left(d+1+\frac{1}{BD-1}\right)\sum_{u=1}^{\ell}\widetilde{A}_u + \sum_{u=1}^{\ell}(BD)^{\ell-u}\widetilde{b}_u$$

where $\widetilde{A}_u = \sup_{i,j}|A_{u,i,j} - A_{u,i,j}^*|$ and $\widetilde{b}_u = \sup_j|b_{u,j} - b_{u,j}^*|$. To do so, we will also prove by induction that:

$$c_\ell \leq B^\ell D^{\ell-1}\left(d+1+\frac{1}{BD-1}\right)$$

where

$$\begin{cases} c_\ell = \sup_{x\in[-1,1]^d}\sup_{1\leq i\leq D}|f_{\theta^*}^\ell(x)_i| \quad \text{for } \ell = 1, ..., L, \\ c_L = \sup_{x\in[-1,1]^d}|f_{\theta^*}(x)|, \end{cases}$$

using the formula:

$$x_n \leq u_n x_{n-1} + v_n \implies x_n \leq \sum_{i=2}^{n} \left(\prod_{j=i+1}^{n} u_j \right) v_i + \left(\prod_{j=2}^{n} u_j \right) x_1 \tag{5.6}$$

for any $n \geq 2$ with the convention $\prod_{j=n+1}^{n} u_j = 1$.

Indeed, we have according to Assumption 5.3.1:

- Initialization:

$$\begin{aligned} c_1 &= \sup_{x \in [-1,1]^d} \sup_{1 \leq i \leq D} |f^1_{\theta^*}(x)_i| \\ &\leq \sup_{x \in [-1,1]^d} \sup_{1 \leq i \leq D} \left| \sum_{j=1}^{d} A^*_{1ij} x_j + b^*_{1i} \right| \\ &\leq \sup_{x \in [-1,1]^d} \sup_{1 \leq i \leq D} \left\{ \sum_{j=1}^{d} |A^*_{1ij}| \cdot |x_j| + |b^*_{1i}| \right\} \\ &\leq d \cdot B \cdot 1 + B \\ &= (d+1)B. \end{aligned}$$

- For any layer ℓ:

$$\begin{aligned} c_\ell &\leq \sup_{x \in [-1,1]^d} \sup_{1 \leq i \leq D} \left| \sum_{j=1}^{D} A^*_{\ell ij} f^{\ell-1}_{\theta^*}(x)_j + b^*_{\ell i} \right| \\ &\leq \sup_{x \in [-1,1]^d} \sup_{1 \leq i \leq D} \left\{ \sum_{j=1}^{D} |A^*_{\ell ij}| \cdot |f^{\ell-1}_{\theta^*}(x)_j| + |b^*_{\ell i}| \right\} \\ &\leq D \cdot B \cdot c_{\ell-1} + B. \end{aligned}$$

- Hence, using Formula (5.6), we get:

$$\begin{aligned} c_\ell &\leq \sum_{u=2}^{\ell} \left(\prod_{v=u+1}^{\ell} DB \right) B + \left(\prod_{v=2}^{\ell} BD \right) c_1 \\ &\leq B \sum_{u=2}^{\ell} (DB)^{\ell-u} + (BD)^{\ell-1}(d+1)B \\ &= B \sum_{u=0}^{\ell-2} (DB)^u + (d+1)D^{\ell-1}B^\ell \\ &= B \frac{(BD)^{\ell-1} - 1}{BD - 1} + (d+1)D^{\ell-1}B^\ell \\ &\leq B^\ell D^{\ell-1} \left(d + 1 + \frac{1}{BD-1} \right). \end{aligned}$$

Let us now come back to finding an upper bound on losses of the partial networks f^ℓ_θs. As previously, we have:

- Initialization:

$$\begin{aligned} r_1(\theta) &= \sup_{x\in[-1,1]^d} \sup_{1\le i\le D} |f^1_{\theta^*}(x)_i - f^1_\theta(x)_i| \\ &\le \sup_{x\in[-1,1]^d} \sup_{1\le i\le D} \left\{ \sum_{j=1}^{d} |A_{1ij} - A^*_{1ij}| \cdot |x_j| + |b_{1i} - b^*_{1i}| \right\} \\ &\le d \cdot \widetilde{A}_1 + \widetilde{b}_1. \end{aligned}$$

- For any layer ℓ:

$$\begin{aligned} r_\ell(\theta) &\le \sup_{x\in[-1,1]^d} \sup_{1\le i\le D} \left\{ \sum_{j=1}^{D} |A_{\ell ij} f^{\ell-1}_\theta(x)_j - A^*_{\ell ij} f^{\ell-1}_{\theta^*}(x)_j| + |b_{\ell i} - b^*_{\ell i}| \right\} \\ &\le \sup_{x\in[-1,1]^d} \sup_{1\le i\le D} \left\{ \sum_{j=1}^{D} \left[|A_{\ell ij} - A^*_{\ell ij}| \cdot |f^{\ell-1}_{\theta^*}(x)_j| + |A_{\ell ij}| \cdot |f^{\ell-1}_{\theta^*}(x)_j - f^{\ell-1}_\theta(x)_j| \right] \right. \\ &\qquad \left. + |b_{\ell i} - b^*_{\ell i}| \right\} \\ &\le Dc_{\ell-1}\widetilde{A}_\ell + BDr_{\ell-1}(\theta) + \widetilde{b}_\ell \\ &\le BDr_{\ell-1}(\theta) + \widetilde{A}_\ell B^{\ell-1} D^{\ell-1} \left(d + 1 + \frac{1}{BD-1} \right) + \widetilde{b}_\ell. \end{aligned}$$

- Finally, using Formula (5.6):

$$\begin{aligned} r_\ell(\theta) &\le \sum_{u=2}^{\ell} \left(\prod_{v=u+1}^{\ell} BD \right) \left(\widetilde{A}_u (BD)^{u-1} \left\{ d + 1 + \frac{1}{BD-1} \right\} + \widetilde{b}_u \right) + \left(\prod_{v=2}^{\ell} BD \right) r_1(\theta) \\ &= \sum_{u=2}^{\ell} (BD)^{\ell-u} \widetilde{A}_u (BD)^{u-1} \left(d + 1 + \frac{1}{BD-1} \right) + \sum_{u=2}^{\ell} (BD)^{\ell-u} \widetilde{b}_u + (BD)^{\ell-1} r_1(\theta) \\ &\le \left(d + 1 + \frac{1}{BD-1} \right) \sum_{u=2}^{\ell} (BD)^{\ell-1} \widetilde{A}_u + \sum_{u=2}^{\ell} (BD)^{\ell-u} \widetilde{b}_u + (BD)^{\ell-1} d\widetilde{A}_1 \\ &\qquad + (BD)^{\ell-1} \widetilde{b}_1 \\ &\le (BD)^{\ell-1} \left(d + 1 + \frac{1}{BD-1} \right) \sum_{u=1}^{\ell} \widetilde{A}_u + \sum_{u=1}^{\ell} (BD)^{\ell-u} \widetilde{b}_u. \end{aligned}$$

Then, for any distribution q:

$$\begin{aligned} \int \|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) &\le \int \|f_\theta - f_{\theta^*}\|_\infty^2 q(\mathrm{d}\theta) = \int r_L(\theta)^2 q(\mathrm{d}\theta) \\ &\le \int 2(BD)^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \left(\sum_{\ell=1}^{L} \widetilde{A}_\ell \right)^2 q(\mathrm{d}\theta) + \int 2 \left(\sum_{\ell=1}^{L} (BD)^{L-\ell} \widetilde{b}_u \right)^2 q(\mathrm{d}\theta) \\ &= 2(BD)^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \left(\int \sum_{\ell=1}^{L} \widetilde{A}_\ell^2 q(\mathrm{d}\theta) + 2 \int \sum_{\ell=1}^{L} \sum_{k=1}^{\ell-1} \widetilde{A}_\ell \widetilde{A}_k q(\mathrm{d}\theta) \right) \\ &\quad + 2 \left(\int \sum_{\ell=1}^{L} (BD)^{2(L-\ell)} \widetilde{b}_l^2 q(\mathrm{d}\theta) + 2 \int \sum_{\ell=1}^{L} \sum_{k=1}^{\ell-1} (BD)^{L-\ell} (BD)^{L-k} \widetilde{b}_\ell \widetilde{b}_k q(\mathrm{d}\theta) \right). \end{aligned}$$

Here, we define $q_n^*(\theta)$ as follows:

$$\begin{cases}\gamma_t^* = \mathbb{I}(\theta_t^* \neq 0), \\ \theta_t \sim \gamma_t^* \, \mathcal{U}([\theta_t^* - s_n, \theta_t^* + s_n]) + (1-\gamma_t^*)\delta_{\{0\}} \quad \text{for each } \ t = 1, ..., T.\end{cases}$$

with $s_n^2 = \frac{S}{4n}(BD)^{-2L}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 \frac{L^2}{(BD)^2} + \frac{1}{(BD)^2-1} + \frac{2}{(BD-1)^2}\right\}^{-1}$. Hence:

$$\int \widetilde{A}_\ell^2 q_n^*(\mathrm{d}\theta) = \int \sup_{i,j}(A_{\ell,i,j} - A_{\ell,i,j}^*)^2 q_n^*(\mathrm{d}A_{\ell,i,j}) \leq s_n^2,$$

and

$$\begin{aligned}\int \widetilde{A}_\ell \widetilde{A}_k q_n^*(\mathrm{d}\theta) &= \left(\int \sup_{i,j} |A_{\ell,i,j} - A_{\ell,i,j}^*| q_n^*(\mathrm{d}\theta)\right)\left(\int \sup_{i,j} |A_{k,i,j} - A_{k,i,j}^*| q_n^*(\mathrm{d}\theta)\right) \\ &\leq |s_n| \cdot |s_n| = s_n^2,\end{aligned}$$

and similarly, $\int \widetilde{b}_\ell^2 q_n^*(\mathrm{d}\theta) \leq s_n^2$ and $\int \widetilde{b}_\ell \widetilde{b}_k q_n^*(\mathrm{d}\theta) \leq s_n^2$.

Then

$$\begin{aligned}&\int \|f_\theta - f_{\theta^*}\|_2^2 q_n^*(\mathrm{d}\theta) \\ &\leq 2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2\left(\int \sum_{\ell=1}^{L} \widetilde{A}_\ell^2 q(\mathrm{d}\theta) + 2\int \sum_{\ell=1}^{L}\sum_{k=1}^{\ell-1} \widetilde{A}_\ell \widetilde{A}_k q(\mathrm{d}\theta)\right) \\ &\quad + 2\left(\int \sum_{\ell=1}^{L} (BD)^{2(L-\ell)} \widetilde{b}_l^2 q(\mathrm{d}\theta) + 2\int \sum_{\ell=1}^{L}\sum_{k=1}^{\ell-1} (BD)^{L-\ell}(BD)^{L-k} \widetilde{b}_\ell \widetilde{b}_k q(\mathrm{d}\theta)\right) \\ &\leq 2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 s_n^2\left(L + 2\sum_{\ell=0}^{L-1}\ell\right) \\ &\quad + 2s_n^2 \sum_{\ell=0}^{L-1}(BD)^{2\ell} + 4s_n^2 \sum_{\ell=1}^{L}\sum_{k=L-\ell+1}^{L-1}(BD)^{L-\ell}(BD)^k \\ &= 2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 s_n^2 L^2 \\ &\quad + 2s_n^2 \frac{(BD)^{2L}-1}{(BD)^2-1} + 4s_n^2 \sum_{\ell=1}^{L}\sum_{k=0}^{\ell-2}(BD)^{L-\ell}(BD)^k (BD)^{L-\ell+1} \\ &= 2s_n^2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 L^2 \\ &\quad + 2s_n^2 \frac{(BD)^{2L}-1}{(BD)^2-1} + 4s_n^2 \sum_{\ell=1}^{L}(BD)^{L-\ell}\frac{(BD)^{\ell-1}-1}{BD-1}(BD)^{L-\ell+1} \\ &\leq 2s_n^2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 L^2 \\ &\quad + 2s_n^2 \frac{(BD)^{2L}-1}{(BD)^2-1} + 4s_n^2 \frac{1}{BD-1}\sum_{\ell=1}^{L}(BD)^{2L-\ell}\end{aligned}$$

$$
\begin{aligned}
&= 2s_n^2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 L^2 \\
&\quad + 2s_n^2\frac{(BD)^{2L}-1}{(BD)^2-1} + 4s_n^2\frac{1}{BD-1}(BD)^L\frac{(BD)^L-1}{BD-1} \\
&\leq 2s_n^2(BD)^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 L^2 + 2s_n^2\frac{(BD)^{2L}-1}{(BD)^2-1} + 4s_n^2\frac{1}{(BD-1)^2}(BD)^{2L} \\
&\leq 2s_n^2(BD)^{2L}\left\{\left(d+1+\frac{1}{BD-1}\right)^2\frac{L^2}{(BD)^2} + \frac{1}{(BD)^2-1} + \frac{2}{(BD-1)^2}\right\} \\
&= \frac{S}{2n} \\
&\leq r_n
\end{aligned}
$$

which proves Equation (5.4).

Third step: *we prove Inequality* (5.5)

We will use the fact that for any K, any $p, p^0 \in [0,1]^K$ such that $\sum_{k=1}^K p_k = \sum_{k=1}^K p_k^0 = 1$ and any distributions Q_k, Q_k^0 for $k = 1, ..., K$, we have:

$$
\mathcal{K}\left(\sum_{k=1}^K p_k^0 Q_k^0 \middle\| \sum_{k=1}^K p_k Q_k\right) \leq \mathcal{K}(p^0\|p) + \sum_{k=1}^K p_k^0 \mathcal{K}(Q_k^0\|Q_k). \tag{5.7}
$$

Please refer to Lemma 6.1 in Chérief-Abdellatif and Alquier (2018) for a proof. Then we write q_n^* and π as mixtures of independent products of mixtures of two components:

$$
q_n^* = \sum_{\gamma\in\mathcal{S}_T^S} \mathbb{I}(\gamma=\gamma^*) \bigotimes_{t=1}^T \Big\{\gamma_t\,\mathcal{U}([l_t,u_t]) + (1-\gamma_t)\delta_{\{0\}}\Big\}
$$

and

$$
\pi = \sum_{\gamma\in\mathcal{S}_T^S} \binom{T}{S^*}^{-1} \bigotimes_{t=1}^T \Big\{\gamma_t\,\mathcal{U}([-B,B]) + (1-\gamma_t)\delta_{\{0\}}\Big\}
$$

Hence, using Inequality 5.7 twice and the additivity of KL for independent distributions:

$$
\begin{aligned}
\mathrm{KL}(q_n^*\|\pi) &\leq \mathrm{KL}\left(\{\mathbb{I}(\gamma=\gamma^*)\}_{\gamma\in\mathcal{S}_T^S} \middle\| \left\{\binom{T^*}{S^*}^{-1}\right\}_{\gamma\in\mathcal{S}_T^S}\right) + \sum_{\gamma\in\mathcal{S}_T^S}\mathbb{I}(\gamma=\gamma^*) \\
&\qquad \mathrm{KL}\left(\bigotimes_{t=1}^T \Big\{\gamma_t\,\mathcal{U}([l_t,u_t]) + (1-\gamma_t)\delta_{\{0\}}\Big\} \middle\| \bigotimes_{t=1}^T \Big\{\gamma_t\,\mathcal{U}([-B,B]) + (1-\gamma_t)\delta_{\{0\}}\Big\}\right) \\
&= \log\binom{T}{S} + \sum_{t=1}^T \mathrm{KL}\Big(\gamma_t^*\,\mathcal{U}([l_t,u_t]) + (1-\gamma_t^*)\delta_{\{0\}} \Big\| \gamma_t^*\,\mathcal{U}([-B,B]) + (1-\gamma_t^*)\delta_{\{0\}}\Big) \\
&\leq \log\binom{T}{S} + \sum_{t=1}^T \gamma_t^*\mathrm{KL}\Big(\mathcal{U}([l_t,u_t]) \Big\| \mathcal{U}([-B,B])\Big) + \sum_{t=1}^T (1-\gamma_t^*)\mathrm{KL}\Big(\delta_{\{0\}}\|\delta_{\{0\}}\Big)
\end{aligned}
$$

$$
\begin{aligned}
&\leq S\log(T) + \sum_{t=1}^{T} \gamma_t^* \log\left(\frac{2B}{u_t - l_t}\right) \\
&= S\log(T) + \sum_{t=1}^{T} \gamma_t^* \log\left(\frac{2B}{2s_n}\right) \\
&= S\log(T) + S\log(B) + \frac{S}{2}\log\left(\frac{1}{s_n^2}\right) \\
&= S\log(T) + S\log(B) \\
&\quad + \frac{S}{2}\log\left(\frac{4n}{S}(BD)^{2L}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 L^2 + \frac{1}{(BD)^2-1} + \frac{2}{(BD-1)^2}\right\}\right),
\end{aligned}
$$

and hence,

$$
\begin{aligned}
\mathrm{KL}(q_n^* \| \pi) &\leq S\log(T) + S\log(B) \\
&\quad + \frac{S}{2}\log\left(\frac{4n}{S}(BD)^{2L}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 L^2 + \frac{1}{(BD)^2-1} + \frac{2}{(BD-1)^2}\right\}\right) \\
&\leq S\log(L(D+1)^2) + S\log(B) + LS\log(BD) \\
&\quad + \frac{S}{2}\log\left(\frac{4n}{S}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 L^2 + \frac{1}{(BD)^2-1} + \frac{2}{(BD-1)^2}\right\}\right) \\
&\leq n r_n,
\end{aligned}
$$

which ends the proof.

5.6.3 Proof of Corollary 5.3.3

Corollary 5.3.3 is a direct consequence of Theorem 5.3.2, and we just need to find an upper bound on $\inf_{\theta^* \in \Theta_{S,L,D}} \|f_{\theta^*} - f_0\|_\infty^2$ and $r_n^{S,L,D}$. Indeed, according to Theorem 5.3.2:

$$
\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha} \inf_{\theta^* \in \Theta_{S,L,D}} \|f_{\theta^*} - f_0\|_\infty^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n. \tag{5.8}
$$

We directly use the rate r_n in the proof of Theorem 5.3.2 rather than $r_n^{S,L,D}$.

Let us assume that f_0 is β-Hölder smooth with $0 < \beta < d$. Then according to Lemma 5.1 in Rockova and Polson (2018), we have for some positive constant C_D independent of n (see Theorem 6.1 in Rockova and Polson (2018)) a neural network with architecture :

$$
\begin{aligned}
L &= 8 + (\lfloor \log_2 n \rfloor + 5)(1 + \lceil \log_2 d \rceil), \\
D &= C_D \lfloor n^{\frac{d}{2\beta+d}} / \log n \rfloor, \\
S &\leq 94 d^2 (\beta+1)^{2d} D (L + \lceil \log_2 d \rceil),
\end{aligned}
$$

with an error $\|f - f_0\|_\infty$ that is at most a constant multiple of $\frac{D}{n} + D^{-\beta/d} \leq C_D n^{\frac{-2\beta}{2\beta+d}} / \log n + C_D^{-\beta/d} n^{\frac{-\beta}{2\beta+d}} \log^{\beta/d} n \leq (C_D / \log n + C_D^{-\beta/d} \log n) n^{\frac{-\beta}{2\beta+d}}$, which gives an upper bound on the first term of the right-hand-side of Inequality 5.8 of order $n^{\frac{-2\beta}{2\beta+d}} \log^2 n$.

In the same time, we have for some constants C, C' that do not depend on n:

$$\begin{aligned} r_n &\leq \frac{SL}{n}\log(BD) + \frac{S}{n}\log(2BL(D+1)^2) + \frac{S}{2n}\log\left(\frac{4n}{S}\left\{3+(d+2)^2L^2\right\}\right) \\ &\leq C\left(\frac{DL^2}{n}\log D + \frac{DL}{n}\log(LD) + \frac{DL}{n}\log n\right) \\ &\leq C'\frac{n^{\frac{d}{2\beta+d}}}{n}\log^2 n = C' n^{\frac{-2\beta}{2\beta+d}}\log^2 n. \end{aligned}$$

Then the tempered posterior distribution $\pi_{n,\alpha}$ concentrates at the minimax rate $r_n = n^{\frac{-2\beta}{2\beta+d}}$ up to a (squared) logarithmic factor for the expected L_2-distance in the sense that:

$$\pi_{n,\alpha}\left(\theta \in \Theta_{S,L,D} \,\middle/\, \|f_\theta - f_0\|_2^2 > M_n n^{\frac{-2\beta}{2\beta+d}}\log^2 n\right) \xrightarrow[n\to+\infty]{} 0.$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

5.6.4 Proof of Theorem 5.3.1

We could prove Theorem 5.3.1 using the prior mass condition (5.2) but we will use instead the same proof than for Theorem 5.3.2. Indeed, we can easily show that for any $\theta^* \in \Theta_{S,L,D}$,

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \pi_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha}\|f_{\theta^*} - f_0\|_2^2 + \inf_q\left\{\frac{2}{1-\alpha}\int \|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) + \frac{2\sigma^2}{\alpha}\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\}$$

where the infimum is taken over all the probability distributions on $\Theta_{S,L,D}$. We have:

$$\begin{aligned} &\inf_q\left\{\frac{2}{1-\alpha}\int \|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) + \frac{2\sigma^2}{\alpha}\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\} \\ &\qquad \leq \inf_{q\in\mathcal{F}_{S,L,D}}\left\{\frac{2}{1-\alpha}\int \|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) + \frac{2\sigma^2}{\alpha}\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\} \\ &\qquad \leq \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right) r_n^{S,L,D}, \end{aligned}$$

which implies

$$\begin{aligned} \mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] &\leq \frac{2}{1-\alpha}\inf_{\theta^*\in\Theta_{S,L,D}}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right) r_n^{S,L,D} \\ &\leq \frac{2}{1-\alpha}\inf_{\theta^*\in\Theta_{S,L,D}}\|f_{\theta^*} - f_0\|_\infty^2 + \frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right) r_n^{S,L,D}. \end{aligned}$$

The rest of the proof follows the same lines than the one of Corollary 5.3.3.

5.6.5 Proof of Theorem 5.3.4

First, we need Donsker and Varadhan's variational formula. Refer to Lemma 1.1.3. in Massart (2007) for a proof.

Theorem 5.6.1. *For any probability λ on some measurable space $(\boldsymbol{E}, \mathcal{E})$ and any measurable function $h : \boldsymbol{E} \to \mathbb{R}$ such that $\int \mathrm{e}^h \mathrm{d}\lambda < \infty$,*

$$\log \int \mathrm{e}^h \mathrm{d}\lambda = \sup_q \left\{ \int h \mathrm{d}q - KL(q, \lambda) \right\},$$

where the supremum is taken over all probability distributions over $\boldsymbol{E}$ and with the convention $\infty - \infty = -\infty$. Moreover, if h is upper-bounded on the support of λ, then the supremum is reached by the distribution of the form:

$$\lambda_h(d\beta) = \frac{\mathrm{e}^{h(\beta)}}{\int \mathrm{e}^h \mathrm{d}\lambda} \lambda(\mathrm{d}\beta).$$

Let us come back to the proof of Theorem 5.3.4. Here, we can not directly use Theorem 2.6 in Alquier and Ridgway (2017). Thus we begin from scratch. For any $\alpha \in (0, 1)$ and $\theta \in \Theta_{S,L,D}$, using the definition of Rényi divergence and $D_\alpha(P^{\otimes n}, R^{\otimes n}) = nD_\alpha(P, R)$ as data are i.i.d.

$$\mathbb{E}\left[\exp\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \right] = 1$$

where $r_n(P_\theta, P^0) = \frac{1}{2\sigma^2} \sum_{i=1}^n \{(Y_i - f_\theta(X_i))^2 - (Y_i - f_0(X_i))^2\}$ is the negative log-likelihood ratio. Then we integrate and use Fubini's theorem,

$$\mathbb{E}\left[\int \exp\left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \pi(d\theta) \right] = 1.$$

According to Theorem 5.6.1,

$$\mathbb{E}\left[\exp\left(\sup_q \left\{ \int \left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) q(d\theta) - \mathrm{KL}(q||\pi) \right\} \right) \right] = 1$$

where the supremum is taken over all probability distributions over $\Theta_{S,L,D}$. Then, using Jensen's inequality,

$$\mathbb{E}\left[\sup_q \left\{ \int \left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) q(d\theta) - \mathrm{KL}(q||\pi) \right\} \right] \leq 0,$$

and then,

$$\mathbb{E}\left[\int \left(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0) \right) \widetilde{\pi}^k_{n,\alpha}(d\theta) - \mathrm{KL}(\widetilde{\pi}^k_{n,\alpha}||\pi) \right] \leq 0.$$

We rearrange terms:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \widetilde{\pi}^k_{n,\alpha}(d\theta) \right] \leq \mathbb{E}\left[\frac{\alpha}{1-\alpha} \int \frac{r_n(P_\theta, P^0)}{n} \widetilde{\pi}^k_{n,\alpha}(d\theta) + \frac{\mathrm{KL}(\widetilde{\pi}^k_{n,\alpha}||\pi)}{n(1-\alpha)} \right],$$

that we can write:

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^k_{n,\alpha}(d\theta)\right] \leq \mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}(d\theta)+\frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}||\pi)}{n(1-\alpha)}\right] + \mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}^k_{n,\alpha}(d\theta)+\frac{\mathrm{KL}(\widetilde{\pi}^k_{n,\alpha}||\pi)}{n(1-\alpha)}\right] - \mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}(d\theta)+\frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}||\pi)}{n(1-\alpha)}\right].$$

Let us precise that $\mathbb{E}\left[\frac{r_n(P_\theta,P^0)}{n}\right] = \mathrm{KL}(P^0||P_\theta) = \frac{\|f_0-f_\theta\|_2^2}{2\sigma^2}$, and:

$$\mathcal{L}_n(q) = -\frac{\alpha}{2\sigma^2}\sum_{i=1}^n\int(Y_i-f_\theta(X_i))^2q(d\theta)-\mathrm{KL}(q\|\pi) \quad \text{up to a constant.}$$

Then:

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^k_{n,\alpha}(d\theta)\right] \leq \mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}(d\theta)+\frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}||\pi)}{n(1-\alpha)}\right]+\frac{\mathbb{E}[\mathcal{L}^*_n-\mathcal{L}^k_n]}{n(1-\alpha)}.$$

We conclude by interverting the infimum and the expectation and the same inequalities than in Theorem 5.3.2:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}(d\theta)+\frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}\|\pi)}{n(1-\alpha)}\right]\\
&\qquad=\mathbb{E}\left[\inf_{q\in\mathcal{F}_{S,L,D}}\left\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}q(\mathrm{d}\theta)+\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\}\right]\\
&\qquad\leq\inf_{q\in\mathcal{F}_{S,L,D}}\left\{\mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}q(\mathrm{d}\theta)+\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right]\right\}\\
&\qquad\leq\frac{\alpha}{1-\alpha}\frac{2}{2\sigma^2}\inf_{\theta^*\in\Theta_{S,L,D}}\|f_{\theta^*}-f_0\|_2^2+\frac{\alpha}{2\sigma^2}\frac{2}{1-\alpha}\left(1+\frac{\sigma^2}{\alpha}\right)r_n^{S,L,D}.
\end{aligned}$$

5.6.6 Proof of Theorem 5.4.1

We start from the last inequality obtained in the proof of Theorem 3 in Chérief-Abdellatif (2019a) that provides an upper bound in α-Rényi divergence for the ELBO model selection framework. We still denote P^0 the generating distribution and P_θ the distribution characterizing the model. Then, for any $\alpha\in(0,1)$:

$$\begin{aligned}
&\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^{\widehat{S},\widehat{L},\widehat{D}}_{n,\alpha}(d\theta)\right]\\
&\leq\inf_{S,L,D}\left\{\inf_{q\in\mathcal{F}_{S,L,D}}\left\{\frac{\alpha}{1-\alpha}\int\mathrm{KL}(P^0,P_{\theta_{S,L,D}})q(d\theta_{S,L,D})+\frac{\mathrm{KL}(q,\Pi^{S,L,D})}{n(1-\alpha)}\right\}+\frac{\log(\frac{1}{\pi_{S,L,D}})}{n(1-\alpha)}\right\}
\end{aligned}$$

where $\Pi^{S,L,D}$ denotes the prior over the parameter set $\Theta_{S,L,D}$ and $\pi_{S,L,D}$ the prior belief over model (S,L,D).

As for the proof of Theorem 5.3.2, for any S, L, D and any $\theta^* \in \Theta_{S,L,D}$:

$$\mathbb{E}\left[\int \frac{\alpha}{2\sigma^2}\|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}^{\widehat{S},\widehat{L},\widehat{D}}(d\theta)\right]$$
$$\leq \frac{\alpha}{1-\alpha}\frac{2}{2\sigma^2}\|f_{\theta^*} - f_0\|_2^2 + \inf_{q \in \mathcal{F}_{S,L,D}}\left\{\frac{\alpha}{1-\alpha}\int \frac{2}{2\sigma^2}\|f_\theta - f_{\theta^*}\|_2^2 q(d\theta) \quad + \frac{\mathrm{KL}(q, \Pi^{S,L,D})}{n(1-\alpha)}\right\}$$
$$+ \frac{\log(\frac{1}{\pi_{S,L,D}})}{n(1-\alpha)},$$

and then for any S, L, D and any $\theta^* \in \Theta_{S,L,D}$,

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}^{\widehat{S},\widehat{L},\widehat{D}}(d\theta)\right] \leq \frac{2}{1-\alpha}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D}$$
$$+ \frac{2\sigma^2}{\alpha(1-\alpha)}\frac{\log(\frac{1}{\pi_{S,L,D}})}{n},$$

which finally leads to Theorem 5.4.1.

5.6.7 Result for sparse Gaussian approximations

In this appendix, we consider non-bounded parameter sets $\Theta_{S,L,D}$ and Gaussians instead of uniform distributions in spike-and-slab priors on $\theta \in \Theta_{S,L,D}$:

$$\begin{cases} \gamma \sim \mathcal{U}(\mathcal{S}_T^S), \\ \theta_t|\gamma_t \sim \gamma_t\, \mathcal{N}(0,1) + (1-\gamma_t)\delta_{\{0\}}, \quad t = 1, ..., T \end{cases}$$

and Gaussian-based sparse spike-and-slab approximations:

$$\begin{cases} \gamma \sim \pi_\gamma, \\ \theta_t|\gamma_t \sim \gamma_t\, \mathcal{N}(m_t, s_n^2) + (1-\gamma_t)\delta_{\{0\}} \quad \text{for each } t = 1, ..., T. \end{cases}$$

The following theorem states that using Gaussians instead of uniform distributions still leads to consistency with the same rate of convergence. Note that the infimum in the RHS of the inequality is taken over a bounded neural network model.

Theorem 5.6.2. *Let us introduce the sets $\Theta_{S,L,D}^B$ that contain the neural network parameters upper bounded by B (in L_∞-norm). Then for any $\alpha \in (0,1)$, for any $B \geq 2$,*

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha}\inf_{\theta^* \in \Theta_{S,L,D}^B}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D}$$

with

$$r_n^{S,L,D} = \frac{SL}{n}\log(2BD) + \frac{S}{4n}\left(12\log(LD) + B^2\right) + \frac{S}{n}\log\left(11d\max(\frac{n}{S}, 1)\right).$$

Proof. The proof follows the same structure than for Theorem 5.3.2. We fix $B \geq 2$.

First step: *we obtain the general inequality*

We can directly write for any $\theta^* \in \Theta_{S,L,D}$,

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \widetilde{\pi}_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha}\|f_{\theta^*} - f_0\|_2^2 + \inf_{q \in \mathcal{F}_{S,L,D}} \left\{\frac{2}{1-\alpha}\int \|f_\theta - f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) + \frac{2\sigma^2}{\alpha}\frac{\mathrm{KL}(q\|\pi)}{n(1-\alpha)}\right\}.$$

We define $\theta^* = \arg\min_{\theta \in \Theta^B_{S,L,D}} \|f_\theta - f_0\|_2$. Again, the rest of the proof consists in finding a distribution $q_n^* \in \mathcal{F}_{S,L,D}$ that satisfies the extended prior mass condition:

$$\int \|f_\theta - f_{\theta^*}\|_2^2 q_n^*(\mathrm{d}\theta) \leq r_n \tag{5.9}$$

and

$$\mathrm{KL}(q_n^*\|\pi) \leq n r_n \tag{5.10}$$

with $r_n = \frac{SL}{n}\log(2BD) + \frac{S}{n}\log(L(D+1)^2) + \frac{S\log\log(3D)}{n} + \frac{SB^2}{4n} + \frac{S}{2n}\log\left(\frac{16n}{S}\left\{3+(d+2)^2\right\}\right) \leq r_n^{S,L,D}$ as $3 + (x+2)^2 \leq 7x^2$ for $x \geq 1$.

Second step: *we prove Inequality (5.9)*

All coefficients of parameter θ^* are upper bounded by B. Hence, we still have:

$$c_\ell \leq B^\ell D^{\ell-1}\left(d + 1 + \frac{1}{BD-1}\right).$$

However, the upper bound on $r_\ell(\theta)$ is not the same, as $|A_{\ell,i,j}|$ can not be upper bounded by B directly and must be upper bounded by $|A^*_{\ell,i,j}| + \widetilde{A}_\ell \leq B + \widetilde{A}_\ell$:

$$\begin{aligned}
r_\ell(\theta) &\leq \sup_{x\in[-1,1]^d} \sup_{1\leq i\leq D} \left\{\sum_{j=1}^{D}\left[|A_{\ell ij} - A^*_{\ell ij}| \cdot |f^{\ell-1}_{\theta^*}(x)_j| + |A_{\ell ij}| \cdot |f^{\ell-1}_{\theta^*}(x)_j - f^{\ell-1}_{\theta}(x)_j|\right] \right. \\
&\qquad \left. + |b_{\ell i} - b^*_{\ell i}|\right\} \\
&\leq \sup_{x\in[-1,1]^d} \sup_{1\leq i\leq D} \left\{\sum_{j=1}^{D}\left[|A_{\ell ij} - A^*_{\ell ij}| \cdot |f^{\ell-1}_{\theta^*}(x)_j| + (B + \widetilde{A}_\ell) \cdot |f^{\ell-1}_{\theta^*}(x)_j - f^{\ell-1}_{\theta}(x)_j|\right] \right. \\
&\qquad \left. + |b_{\ell i} - b^*_{\ell i}|\right\} \\
&\leq D c_{\ell-1}\widetilde{A}_\ell + (B + \widetilde{A}_\ell) D r_{\ell-1}(\theta) + \widetilde{b}_\ell \\
&\leq (B + \widetilde{A}_\ell) D r_{\ell-1}(\theta) + \widetilde{A}_\ell B^{\ell-1} D^{\ell-1}\left(d + 1 + \frac{1}{BD-1}\right) + \widetilde{b}_\ell.
\end{aligned}$$

Then, using Formula 5.6:

$$r_\ell(\theta) \le \sum_{u=2}^{\ell}\left(\prod_{v=u+1}^{\ell}(B+\widetilde{A_v})D\right)\left(\widetilde{A}_u(BD)^{u-1}\left\{d+1+\frac{1}{BD-1}\right\}+\widetilde{b}_u\right) + \left(\prod_{v=2}^{\ell}(B+\widetilde{A_v})D\right)r_1(\theta)$$
$$\le \sum_{u=2}^{\ell} D^{\ell-u}\prod_{v=u+1}^{\ell}(B+\widetilde{A_v})\widetilde{A}_u(BD)^{u-1}\left(d+1+\frac{1}{BD-1}\right) + \sum_{u=2}^{\ell} D^{\ell-u}\prod_{v=u+1}^{\ell}(B+\widetilde{A_v})\widetilde{b}_u + D^{\ell-1}\prod_{v=2}^{\ell}(B+\widetilde{A_v})r_1(\theta),$$

and using inequality $r_1(\theta) \le d\cdot\widetilde{A}_1+\widetilde{b}_1$:

$$r_\ell(\theta) \le D^{\ell-1}\left(d+1+\frac{1}{BD-1}\right)\sum_{u=2}^{\ell}B^{u-1}\prod_{v=u+1}^{\ell}(B+\widetilde{A}_v)\widetilde{A}_u + \sum_{u=2}^{\ell}D^{\ell-u}\prod_{v=u+1}^{\ell}(B+\widetilde{A}_v)\widetilde{b}_u + dD^{\ell-1}\prod_{v=2}^{\ell}(B+\widetilde{A}_v)\widetilde{A}_1 + D^{\ell-1}\prod_{v=2}^{\ell}(B+\widetilde{A}_v)\widetilde{b}_1$$
$$\le D^{\ell-1}\left(d+1+\frac{1}{BD-1}\right)\sum_{u=1}^{\ell}B^{u-1}\prod_{v=u+1}^{\ell}(B+\widetilde{A}_v)\widetilde{A}_u + \sum_{u=1}^{\ell}D^{\ell-u}\prod_{v=u+1}^{\ell}(B+\widetilde{A}_v)\widetilde{b}_u.$$

Then we have for any distribution $q(\theta)=q_1(\theta_1)\times\ldots\times q_T(\theta_T)$:

$$\int\|f_\theta-f_{\theta^*}\|_2^2 q(\mathrm{d}\theta) \le \int\|f_\theta-f_{\theta^*}\|_\infty^2 q(\mathrm{d}\theta) = \int r_L(\theta)^2 q(\mathrm{d}\theta)$$
$$\le \int 2D^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2\left(\sum_{\ell=1}^{L}B^{\ell-1}\prod_{v=\ell+1}^{L}(B+\widetilde{A}_v)\widetilde{A}_\ell\right)^2 q(\mathrm{d}\theta) + \int 2\left(\sum_{\ell=1}^{L}D^{L-\ell}\prod_{v=\ell+1}^{L}(B+\widetilde{A}_v)\widetilde{b}_\ell\right)^2 q(\mathrm{d}\theta)$$
$$= 2D^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2\left(\int\sum_{\ell=1}^{L}B^{2\ell-2}\prod_{v=\ell+1}^{L}(B+\widetilde{A}_v)^2\widetilde{A}_\ell^2 q(\mathrm{d}\theta) + 2\int\sum_{\ell=1}^{L}\sum_{k=1}^{\ell-1}B^{\ell-1}B^{k-1}\prod_{v=\ell+1}^{L}(B+\widetilde{A}_v)\widetilde{A}_\ell\prod_{v=k+1}^{L}(B+\widetilde{A}_v)\widetilde{A}_k q(\mathrm{d}\theta)\right)$$
$$+ 2\left(\int\sum_{\ell=1}^{L}D^{2(L-\ell)}\prod_{v=\ell+1}^{L}(B+\widetilde{A}_v)^2\widetilde{b}_\ell^2 q(\mathrm{d}\theta) + 2\int\sum_{\ell=1}^{L}\sum_{k=1}^{\ell-1}D^{L-\ell}D^{L-k}\prod_{v=\ell+1}^{L}(B+\widetilde{A}_v)\widetilde{b}_\ell\prod_{v=k+1}^{L}(B+\widetilde{A}_v)\widetilde{b}_k q(\mathrm{d}\theta)\right)$$
$$= 2D^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2\left(\sum_{\ell=1}^{L}B^{2\ell-2}\prod_{v=\ell+1}^{L}\int(B+\widetilde{A}_v)^2 q(\mathrm{d}\theta)\int\widetilde{A}_\ell^2 q_\ell(\mathrm{d}\theta_\ell)\right.$$
$$\left.+ 2\sum_{\ell=1}^{L}\sum_{k=1}^{\ell-1}B^{\ell-1}B^{k-1}\prod_{v=\ell+1}^{L}\int(B+\widetilde{A}_v)^2 q(\mathrm{d}\theta)\int\widetilde{A}_\ell q_\ell(\mathrm{d}\theta_\ell)\prod_{v=k+1}^{\ell}\int(B+\widetilde{A}_v)q(\mathrm{d}\theta)\int\widetilde{A}_k q(\mathrm{d}\theta)\right)$$

$$+2\Bigg(\sum_{\ell=1}^{L} D^{2(L-\ell)} \prod_{v=\ell+1}^{L} \int (B+\widetilde{A}_v)^2 q(\mathrm{d}\theta) \int \widetilde{b}_\ell^2 q(\mathrm{d}\theta)$$
$$+2\sum_{\ell=1}^{L}\sum_{k=1}^{\ell-1} D^{L-\ell} D^{L-k} \prod_{v=\ell+1}^{L} \int (B+\widetilde{A}_v)^2 q(\mathrm{d}\theta) \int \widetilde{b}_\ell q(\mathrm{d}\theta) \prod_{v=k+1}^{\ell} \int (B+\widetilde{A}_v) q(\mathrm{d}\theta) \int \widetilde{b}_k q(\mathrm{d}\theta)\Bigg).$$

Here, we define $q_n^*(\theta)$ as follows:

$$\begin{cases} \gamma_t^* = \mathbb{I}(\theta_t^* \neq 0), \\ \theta_t \sim \gamma_t^* \, \mathcal{N}(\theta_t^*, s_n^2) + (1-\gamma_t^*)\delta_{\{0\}} \quad \text{for each } \ t = 1, ..., T. \end{cases}$$

with $s_n^2 = \frac{S}{16n} \log(3D)^{-1} (2BD)^{-2L} \left\{ \left(d + 1 + \frac{1}{BD-1} \right)^2 + \frac{1}{(2BD)^2-1} + \frac{2}{(2BD-1)^2} \right\}^{-1}$.

We upper bound the expectation of the supremum of absolute values of Gaussian variables:

$$\int \widetilde{A}_\ell q_n^*(\mathrm{d}\theta) = \int \sup_{i,j} |A_{\ell,i,j} - A_{\ell,i,j}^*| q_n^*(\mathrm{d}\theta) \leq \sqrt{2 s_n^2 \log(2D^2)} = \sqrt{4 s_n^2 \log(3D)},$$

and use Example 2.7 in Boucheron et al. (2003):

$$\int \widetilde{A}_\ell^2 q_n^*(\mathrm{d}\theta) = \int \sup_{i,j} (A_{\ell,i,j} - A_{\ell,i,j}^*)^2 q_n^*(\mathrm{d}\theta) \leq s_n^2 (1 + 2\sqrt{\log(D^2)} + \log(D^2)) = 4 s_n^2 \log(3D),$$

which also give:

$$\int (B + \widetilde{A}_\ell) q_n^*(\mathrm{d}\theta) = B + \int \widetilde{A}_\ell q_n^*(\mathrm{d}\theta) \leq B + \sqrt{4 s_n^2 \log(3D)} \leq 2B,$$

and

$$\begin{aligned} \int (B + \widetilde{A}_\ell)^2 q_n^*(\mathrm{d}\theta) &= B^2 + 2B \int \widetilde{A}_\ell q_n^*(\mathrm{d}\theta) + \int \widetilde{A}_\ell^2 q_n^*(\mathrm{d}\theta) \\ &\leq B^2 + 2B\sqrt{4 s_n^2 \log(3D)} + 4 s_n^2 \log(3D) \\ &\leq 4B^2 \end{aligned}$$

as $\sqrt{4 s_n^2 \log(3D)} \leq B$ ($s_n^2 \leq \frac{LD(D+1)}{16n} (2BD)^{-2L} \leq \frac{2LD^2}{16n} 4^{-2L} D^{-2L} \leq 1$).

Similarly,

$$\int \widetilde{b}_\ell q_n^*(\mathrm{d}\theta) \leq \sqrt{4 s_n^2 \log(3D)}$$

and

$$\int \widetilde{b}_\ell^2 q_n^*(\mathrm{d}\theta) \leq 4 s_n^2 \log(3D).$$

Then

$$\begin{aligned} \int \|f_\theta - f_{\theta^*}\|_2^2 q_n^*(\mathrm{d}\theta) \leq 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \Bigg(& \sum_{\ell=1}^{L} B^{2\ell-2} (4B^2)^{L-\ell} 4 s_n^2 \log(3D) \\ & + 2 \sum_{\ell=1}^{L} \sum_{k=1}^{\ell-1} B^{\ell-1} B^{k-1} (4B^2)^{L-\ell} \sqrt{4 s_n^2 \log(3D)} (2B)^{\ell-k} \sqrt{4 s_n^2 \log(3D)} \Bigg) \end{aligned}$$

$$
\begin{aligned}
&+ 2\Bigg(\sum_{\ell=1}^{L} D^{2(L-\ell)} (4B^2)^{L-\ell} 4 s_n^2 \log(3D) \\
&+ 2 \sum_{\ell=1}^{L} \sum_{k=1}^{\ell-1} D^{L-\ell} D^{L-k} (4B^2)^{L-\ell} \sqrt{4 s_n^2 \log(3D)} (2B)^{\ell-k} \sqrt{4 s_n^2 \log(3D)} \Bigg),
\end{aligned}
$$

i.e.

$$
\begin{aligned}
&\int \|f_\theta - f_{\theta^*}\|_2^2 q_n^*(\mathrm{d}\theta) \\
&\quad \le 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \Bigg(B^{2L-2} 4 s_n^2 \log(3D) \sum_{\ell=0}^{L-1} 4^\ell \\
&\qquad\qquad + 2B^{2L-2} 4 s_n^2 \log(3D) \sum_{\ell=1}^{L} \sum_{k=1}^{\ell-1} 2^{L-\ell} 2^{L-k} \Bigg) \\
&\qquad + 2 \Bigg(4 s_n^2 \log(3D) \sum_{\ell=1}^{L} (2BD)^{2L-2\ell} + 8 s_n^2 \log(3D) \sum_{\ell=1}^{L} \sum_{k=1}^{\ell-1} (2BD)^{L-\ell} (2BD)^{L-k} \Bigg) \\
&\quad \le 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \Bigg(B^{2L-2} 4 s_n^2 \log(3D) \frac{4^L - 1}{4-1} \\
&\qquad\qquad + 2B^{2L-2} 4 s_n^2 \log(3D) \sum_{\ell=1}^{L} 2^{L-\ell} 2^{L-\ell+1} \sum_{k=0}^{\ell-2} 2^k \Bigg) \\
&\qquad + 2 \Bigg(4 s_n^2 \log(3D) \sum_{\ell=0}^{L-1} (2BD)^{2\ell} + 8 s_n^2 \log(3D) \sum_{\ell=1}^{L} (2BD)^{L-\ell} (2BD)^{L-\ell+1} \sum_{k=0}^{\ell-2} (2BD)^k \Bigg) \\
&\quad \le 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \Bigg(B^{2L-2} 4 s_n^2 \log(3D) \frac{4^L}{3} \\
&\qquad\qquad + 2B^{2L-2} 4 s_n^2 \log(3D) \sum_{\ell=1}^{L} 2^{L-\ell} 2^{L-\ell+1} 2^{\ell-1} \Bigg) \\
&\qquad + 2 \Bigg(4 s_n^2 \log(3D) \frac{(2BD)^{2L}}{(2BD)^2 - 1} + 8 s_n^2 \log(3D) \sum_{\ell=1}^{L} (2BD)^{L-\ell} (2BD)^{L-\ell+1} \frac{(2BD)^{\ell-1}}{2BD-1} \Bigg) \\
&\quad \le 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \Bigg(B^{2L-2} 4 s_n^2 \log(3D) \frac{4^L}{3} + 2B^{2L-2} 4 s_n^2 \log(3D) 2^L \sum_{\ell=0}^{L-1} 2^\ell \Bigg) \\
&\qquad + 2 \Bigg(4 s_n^2 \log(3D) \frac{(2BD)^{2L}}{(2BD)^2 - 1} + 8 s_n^2 \log(3D) \sum_{\ell=0}^{L-1} (2BD)^\ell \frac{(2BD)^L}{2BD-1} \Bigg) \\
&\quad \le 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 \Bigg(B^{2L-2} 4 s_n^2 \log(3D) \frac{4^L}{3} + 2B^{2L-2} 4 s_n^2 \log(3D) 2^{2L} \Bigg) \\
&\qquad + 2 \Bigg(4 s_n^2 \log(3D) \frac{(2BD)^{2L}}{(2BD)^2 - 1} + 8 s_n^2 \log(3D) \frac{(2BD)^{2L}}{(2BD-1)^2} \Bigg) \\
&\quad = 2D^{2L-2} \left(d + 1 + \frac{1}{BD-1} \right)^2 4 s_n^2 \log(3D) \left(B^{2L-2} \frac{4^L}{3} + 2B^{2L-2} 2^{2L} \right) \\
&\qquad + 2 \left(\frac{(2BD)^{2L}}{(2BD)^2 - 1} + 2 \frac{(2BD)^{2L}}{(2BD-1)^2} \right) 4 s_n^2 \log(3D),
\end{aligned}
$$

and consequently, as $BD \geq 2$,

$$
\begin{aligned}
&\int \|f_\theta - f_{\theta^*}\|_2^2 q_n^*(\mathrm{d}\theta) \\
&\quad \leq 8s_n^2 \log(3D)\Bigg\{ D^{2L-2}\left(d+1+\frac{1}{BD-1}\right)^2 \frac{7}{3} B^{2L-2} 2^{2L} \\
&\qquad\qquad + (2BD)^{2L}\left(\frac{1}{(2BD)^2-1} + \frac{2}{(2BD-1)^2}\right)\Bigg\} \\
&\quad = 8s_n^2 \log(3D)\Bigg\{ (2BD)^{2L}\frac{1}{(BD)^2}\left(d+1+\frac{1}{BD-1}\right)^2 \frac{7}{3} \\
&\qquad\qquad + (2BD)^{2L}\left(\frac{1}{(2BD)^2-1} + \frac{2}{(2BD-1)^2}\right)\Bigg\} \\
&\quad \leq 8s_n^2 \log(3D)(2BD)^{2L}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 + \frac{1}{(2BD)^2-1} + \frac{2}{(2BD-1)^2}\right\} \\
&\quad = \frac{S}{2n} \\
&\quad \leq r_n.
\end{aligned}
$$

which ends Step 2.

Third step: *we prove Inequality* (5.10)

We end the proof:

$$
\begin{aligned}
\mathrm{KL}(q_n^* \| \pi) &\leq \log\binom{T}{S} + \sum_{t=1}^T \gamma_t^* \mathrm{KL}\left(\mathcal{N}(\theta_t^*, s_n^2) \middle\| \mathcal{N}(0,1)\right) \\
&\leq S\log(T) + \sum_{t=1}^T \gamma_t^* \left\{\frac{1}{2}\log\left(\frac{1}{s_n^2}\right) + \frac{s_n^2+\theta_t^{*2}}{2} - \frac{1}{2}\right\} \\
&\leq S\log(T) + \sum_{t=1}^T \gamma_t^* \left\{\frac{1}{2}\log\left(\frac{1}{s_n^2}\right) + \frac{s_n^2+B^2}{2} - \frac{1}{2}\right\} \\
&= S\log(T) + \frac{S}{2}s_n^2 + \frac{S}{2}\frac{B^2-1}{2} + \frac{S}{2}\log\left(\frac{1}{s_n^2}\right) \\
&\leq S\log(T) + \frac{S}{2} + \frac{S}{2}\frac{B^2-1}{2} \\
&\quad + \frac{S}{2}\log\left(\frac{16n}{S}\log(3D)(2BD)^{2L}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 + \frac{1}{(2BD)^2-1} + \frac{2}{(2BD-1)^2}\right\}\right) \\
&\leq S\log(L(D+1)^2) + \frac{B^2 S}{4} + LS\log(2BD) + \frac{S}{2}\log\log(3D) \\
&\quad + \frac{S}{2}\log\left(\frac{16n}{S}\left\{\left(d+1+\frac{1}{BD-1}\right)^2 + \frac{1}{(BD)^2-1} + \frac{2}{(BD-1)^2}\right\}\right) \\
&\leq nr_n.
\end{aligned}
$$

□

Chapter 6

Consistency of ELBO maximization for model selection

The Evidence Lower Bound (ELBO) is a quantity that plays a key role in variational inference. It can also be used as a criterion in model selection. However, though extremely popular in practice in the variational Bayes community, there has never been a general theoretic justification for selecting based on the ELBO. In this chapter, we show that the ELBO maximization strategy has strong theoretical guarantees, and is robust to model misspecification while most works rely on the assumption that one model is correctly specified. We illustrate our theoretical results by an application to the selection of the number of principal components in probabilistic PCA.

6.1 Introduction

Approximate Bayesian inference is at the core of modern Bayesian statistics and machine learning. While exact Bayesian inference is often intractable, variational inference has proved to provide an efficient solution when dealing with large datasets and complex probabilistic models. Variational Bayes (VB) aims at maximizing a numerical quantity referred to as Evidence Lower Bound on the marginal likelihood (ELBO), and thus makes use of optimization techniques to converge faster than Monte Carlo sampling approach. Blei et al. (2017) provides a comprehensive survey on variational inference. Although VB is mainly used for its practical efficiency, little attention has been put towards its theoretical properties during the last years. While Alquier et al. (2016) studied the properties of variational approximations of Gibbs distributions used in machine learning for bounded loss functions, Alquier and Ridgway (2017); Zhang and Gao (2017); Wang and Blei (2018); Bhattacharya et al. (2018); Chérief-Abdellatif and Alquier (2018) extended the results to more general statistical models.

At the same time, model selection remains a major problem of interest in statistics that naturally arises in the course of scientific inquiry. The statistician aims at selecting a model among several candidates given an observed dataset. To do so, one can perform cross validation as in Vehtari et al. (2014) or maximize a numerical criterion to make the

final choice, see the review of Rao and Wu (2001). In the literature, penalized criteria such as AIC and BIC respectively introduced by Akaike (1974) and Schwarz (1978) are popular. While AIC aims at optimizing the prediction performance, BIC is more suitable for recovering with high probability the true model (when such a model exists), see Yang (2005). Thus, it is necessary to define a criterion suited to a given objective. Meanwhile, a non-asymptotic theory of penalization using oracle inequalities has been developed during the last two decades, and offers a simple way to assess the quality of a given model selection criterion. We refer the interested reader to Catoni (2007) for more details.

In this chapter, we are interested in finding an estimate of the distribution of the data, and we need to choose from among competing models. Blei et al. (2017) states that "the [evidence lower] bound is a good approximation of the marginal likelihood, which provides a basis for selecting a model. Though this sometimes works in practice, selecting based on a bound is not justified in theory". Since then, authors of Chérief-Abdellatif and Alquier (2018) have provided an analysis of model selection based on the ELBO in the case of mixture models. We extend their result to the general case of independent and identically distributed (i.i.d.) data, and we provide an oracle inequality on the ELBO criterion that justifies the consistency of ELBO maximization when the objective is the estimation of the distribution of the data. In particular, as soon as there exists a true model, we show that the ELBO criterion is adaptive and that the selected estimator achieves the convergence rate of the variational approximation associated with the true model.

The rest of this chapter is organized as follows. Section 6.2 introduces the setting and the key concepts needed to understand our results. In Section 6.3, we prove that the ELBO criterion provides a variational approximation that is consistent with the sample size as soon as there exists a true model. We also extend the result to misspecified models. We finally illustrate the main theorem of this chapter by an application to the selection of the number of principal components in probabilistic Principal Component Analysis (PCA) in Section 6.4. All the proofs are deferred to the appendix.

6.2 Framework

Let us introduce the notations and the framework we adopt in this chapter. We consider a collection of i.i.d. random variables X_1,...,X_n distributed according to some probability distribution P^0 in a measurable space $\left(\mathbb{X}, \mathcal{X}\right)$. We denote $X_1^n = (X_1, ..., X_n)$. We consider a countable collection $\{\mathcal{M}_K / K \geq 1\}$ of statistical mixture models $\mathcal{M}_K = \{P_{\theta_K} \,/\, \theta_K \in \Theta_K\}$ where Θ_K is the parameter set associated with index K. We make no assumptions on Θ_K's nor on P_{θ_K}. Parameter spaces may overlap or have inclusion relationships. Let $\mathcal{M}_1^+(\Theta_K)$ be the set of all probability distributions over Θ_K.

We use a Bayesian approach, and we define a prior π over the full parameter space $\cup_{K\geq 1}\Theta_K$ (equipped with some suited sigma-algebra). First, we specify a prior weight π_K assigned to model $\mathcal{M}_K$, and then a conditional prior $\Pi_K(.)$ on $\theta_K \in \Theta_K$ given model

$\mathcal{M}_K$:

$$\pi = \sum_{K \geq 1} \pi_K \Pi_K.$$

The Kullback-Leibler divergence between two probability distributions P and R is

$$\mathrm{KL}(P, R) = \begin{cases} \int \log\left(\frac{dP}{dR}\right) dP \text{ if } R \text{ dominates } P, \\ +\infty \text{ otherwise.} \end{cases}$$

For any $\alpha \neq 1$, authors of Van Erven and Harremos (2014) detail the properties of the α-Rényi divergence between two probability distributions P and R which is equal to:

$$D_\alpha(P, R) = \begin{cases} \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dR}\right)^{\alpha - 1} dP \text{ if } R \text{ dominates } P, \\ +\infty \text{ otherwise.} \end{cases}$$

We define the tempered posterior distribution $\pi^K_{n,\alpha}(.|X_1^n)$ on parameter $\theta_K \in \Theta_K$ given model $\mathcal{M}_K$ using prior Π_K and likelihood L_n for any $\alpha \in (0, 1)$:

$$\pi^K_{n,\alpha}(d\theta_K | X_1^n) \propto L_n(\theta_K)^\alpha \Pi_K(d\theta_K).$$

This definition is a slight variant of the regular Bayesian posterior (for which $\alpha = 1$), and is also referred to as Bayesian fractional posterior in Bhattacharya et al. (2016). This posterior is easier to sample from, more robust to model misspecification and requires less stringent conditions to obtain consistency, see respectively Behrens et al. (2012), Grünwald et al. (2017) and Bhattacharya et al. (2016).

The Variational Bayes approximation $\widetilde{\pi}^K_{n,\alpha}(.|X_1^n)$ of the tempered posterior associated with model $\mathcal{M}_K$ is then defined as the projection, with respect to the Kullback-Leibler divergence, of the tempered posterior onto some set $\mathcal{F}_K$:

$$\widetilde{\pi}^K_{n,\alpha}(.|X_1^n) = \arg \min_{\rho_K \in \mathcal{F}_K} \mathrm{KL}(\rho_K, \pi^K_{n,\alpha}(.|X_1^n)).$$

The choice of the variational set $\mathcal{F}_K$ is crucial: the variational approximation must be close enough to the target distribution (as an approximation of the tempered posterior) but not too close (in order to be tractable). A classical variational set $\mathcal{F}_K$ is the parametric family which leads to a tractable parametric approximation, e.g. a Gaussian distribution. Another popular set $\mathcal{F}_K$ in the VB community is the mean-field approximation that is based on a partition of the space of parameters, and which consists in a factorization of the variational approximation over the partition.

Alternatively, the variational approximation is often defined as the distribution into $\mathcal{F}_K$ that maximizes the Evidence Lower Bound:

$$\widetilde{\pi}^K_{n,\alpha}(.|X_1^n) = \arg \max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K) \rho_K(d\theta_K) - \mathrm{KL}\Big(\rho_K, \Pi_K\Big) \right\}$$

where the function inside the argmax operator is the ELBO (as a function of K and ρ_K) and ℓ_n is the log-likelihood. In the following, we will just call $\mathrm{ELBO}(K)$ the closest approximation to the log-evidence, i.e. the value of the ELBO evaluated at its maximum:

$$\mathrm{ELBO}(K) = \alpha \int \ell_n(\theta_K) \widetilde{\pi}^K_{n,\alpha}(d\theta_K | X_1^n) - \mathrm{KL}(\widetilde{\pi}^K_{n,\alpha}(.|X_1^n), \Pi_K).$$

In the variational Bayes community, researchers and practitioners use the ELBO in order to select the model from which they will consider the final variational approximation $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n)$, as stated in Blei et al. (2017). We propose to consider a penalized version of the ELBO criterion

$$\widehat{K} = \arg\max_{K \geq 1} \left\{ \text{ELBO}(K) - \log\left(\frac{1}{\pi_K}\right) \right\}$$

which is a slight variant of the classical definition, although choosing a uniform prior over a finite number of models leads to maximizing the ELBO. Note that the penalty term is not just an artefact in order to ease the theoretical proof, but it is a complexity term that reflects our prior beliefs over the different models.

We will provide in the next section a theoretical justification to such a selection criterion and show that the selected variational estimator $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n)$ is consistent under mild conditions as soon as there exists a true model. We will adopt the definition of *consistency* used in Alquier and Ridgway (2017) and Chérief-Abdellatif and Alquier (2018) that is, the Bayesian estimator is said to be consistent if, in expectation (with respect to the random variables distributed according to P^0), the average Rényi loss between a distribution in the selected model and the true distribution (over the Bayesian estimator) goes to zero as $n \to +\infty$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n) \right] \xrightarrow[n\to+\infty]{} 0.$$

This definition is closely related to the notion of *concentration* which is defined in Ghosal et al. (2000) as the asymptotic concentration of the Bayesian estimator around the true distribution, and which is usually used to assess frequentist guarantees for Bayesian estimators. It is sometimes also referred to as *contraction* (or even *consistency*). See Appendix 6.5.1 for more details on the connection between the notions of *consistency* and *concentration*.

6.3 Consistency of the ELBO criterion

In this section, unless explicitly stated otherwise, we assume that there exists a true model $\mathcal{M}_{K_0}$ that contains the true distribution P^0, i.e. that there exists K_0 and $\theta^0 \in \Theta_{K_0}$ such that $P^0 = P_{\theta^0}$.

A key assumption introduced in Ghosal et al. (2000) in order to obtain the concentration of the regular posterior distribution $\pi_{n,1}^{K_0}(.|X_1^n)$ associated with the true model $\mathcal{M}_{K_0}$ is a *prior mass condition* which states that the prior Π_{K_0} must give enough mass to some neighborhood (in the Kullback-Leibler sense) of the true parameter. Bhattacharya et al. (2016) showed that this condition was sufficient when considering tempered posteriors $\pi_{n,\alpha}^{K_0}(.|X_1^n)$. Alquier and Ridgway (2017) extended this assumption in order to obtain the concentration and the consistency of variational approximations of the tempered posteriors $\widetilde{\pi}_{n,\alpha}^{K_0}(.|X_1^n)$. In addition to the previous prior mass condition, this extension requires the variational set $\mathcal{F}_{K_0}$ to contain probability distributions concentrated around the true parameter. Note that when $\mathcal{F}_{K_0} = \mathcal{M}_1^+(\Theta_{K_0})$, this goes back to the standard prior mass

condition. This extended prior mass condition is standard in the variational Bayes community, see Alquier and Ridgway (2017); Chérief-Abdellatif and Alquier (2018), and can be formulated as follows:

Assumption : *We assume that there exists r_n for which there is a distribution $\rho_{K_0,n} \in \mathcal{F}_{K_0}$ such that*:

$$\int \mathrm{KL}(P^0, P_{\theta_{K_0}})\rho_{K_0,n}(d\theta_{K_0}) \leq r_n \text{ and } \mathrm{KL}(\rho_{K_0,n}, \Pi_{K_0}) \leq nr_n. \tag{6.1}$$

Remark 6.3.1. *Define the KL-ball $\mathcal{B}$ centered at θ_0 of radius r_n:*

$$\mathcal{B} = \{\theta \in \Theta_{K_0} / \mathrm{KL}(P_{\theta_0}, P_\theta) \leq r_n\},$$

and consider the restriction $\rho_{K_0,n}$ of Π_{K_0} to $\mathcal{B}$. Then it is clear that when $\rho_{K_0,n} \in \mathcal{F}_{K_0}$, Assumption 6.1 becomes equivalent to the former prior mass condition of Ghosal et al. (2000), i.e. $\Pi_{K_0}(\mathcal{B}) \geq e^{-nr_n}$. The computation of the prior mass $\Pi_{K_0}(\mathcal{B})$ is a major difficulty. It has been raised as a question of interest in Ghosal et al. (2000), and is addressed for categorical distributions and Dirichlet priors in Ghosal et al. (2000) (but for an L_1-ball) and in Chérief-Abdellatif and Alquier (2018) (for a KL-ball). Unfortunately, $\rho_{K_0,n}$ does not belong to $\mathcal{F}_{K_0}$ in general and the computation of the prior mass is no longer sufficient. Nevertheless, the strategy of computing the prior mass of KL-balls remains of interest when dealing with mixture models and mean-field approximation sets, see Chérief-Abdellatif and Alquier (2018) where the authors showed that studying the prior mass condition of Ghosal et al. (2000) independently on the weights and on each component becomes sufficient.

Remark 6.3.2. *When $\mathcal{F}_{K_0}$ is parametric, it is often possible to overcome the difficulty presented above in order to find a rate r_n as in Assumption 6.1. Indeed, the point is to express the distribution $\rho_{K_0,n}$ using the general parametric form of the variational family, and to find relevant values of the parameters that will lead to fast rates of convergence r_n. This is the strategy we follow in Section 6.4 for probabilistic PCA. See Alquier and Ridgway (2017); Chérief-Abdellatif and Alquier (2018) for other examples of such computations.*

Alquier and Ridgway (2017) showed that the variational approximation $\widetilde{\pi}^{K_0}_{n,\alpha}(.|X_1^n)$ associated with a true model is consistent under Assumption 6.1 and that the convergence rate is equal to r_n. Nevertheless, in model selection, we do not necessarily know which model is true and the challenge is to be able to find one such that the corresponding approximation is consistent at a comparable convergence rate. We show that the variational approximation $\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(.|X_1^n)$ associated with the selected model is also consistent at rate r_n as soon as Assumption 6.1 is satisfied:

Theorem 6.3.1. *Assume that Assumption 6.1 is satisfied. Then for any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}r_n + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

The inequality in Theorem 6.3.1 shows the adaptivity of our procedure. Indeed, whatever the value of $\widehat{K}$ (which can be different from K_0), we obtain the consistency of the selected variational approximation at the same rate of convergence than the estimator associated with the true model (as soon as the additional term in the upper bound is lower than r_n, which is the case for prior weights used in practice). We recall that we look for a good estimation of the true distribution P^0 and not for an estimation of the true model index K_0 which is a different task that would require identifiability assumptions that are stronger than those in our theorem. The overall rate is composed of the convergence rate associated with the true model $\mathcal{M}_{K_0}$, and of a complexity term that reflects the prior belief over the (unknown) true model. For example, if we range a countable number of models according to our prior belief, and we take $\pi_K = 2^{-K}$, then the corresponding term will be of order K_0/n. More generally, when $\frac{1}{n} \lesssim r_n$, we obtain the consistency at the rate associated with the true model.

As a short example, Chérief-Abdellatif and Alquier (2018) investigated the case of mixture models. For instance, authors obtained a convergence rate equal to $K_0 \log(nK_0)/n$ for Gaussian mixtures when there exists a true K_0-components mixture model. We study another example in Section 6.4.

We can also extend this result to misspecified models. In the model selection literature, only little attention has been put to misspecification when the true distribution does not belong to any of the models, see Lv and Liu (2013). Now, we do not assume any longer that there exists a true model, and we show that our ELBO criterion is robust to model misspecification:

Theorem 6.3.2. *For each index K, let us define the set $\Theta_K(r_{K,n})$ of parameters $\theta_K^* \in \Theta_K$, for which there is a distribution $\rho_{K,n} \in \mathcal{F}_K$ such that:*

$$\int \mathbb{E}\left[\log \frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right] \rho_{K,n}(d\theta_K) \leq r_{K,n} \text{ and } \mathrm{KL}(\rho_{K,n}, \Pi_K) \leq n r_{K,n}. \tag{6.2}$$

Then for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta | X_1^n)\right] \leq \inf_{K \geq 1} \left\{ \frac{\alpha}{1-\alpha} \inf_{\theta_K^* \in \Theta_K(r_{K,n})} \mathrm{KL}(P^0, P_{\theta_K^*}) + \frac{1+\alpha}{1-\alpha} r_{K,n} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)} \right\}.$$

Note that when there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\theta^0}$ with $\theta^0 \in \Theta_{K_0}$, then under Assumption 6.1, we get $\theta^0 \in \Theta_{K_0}(r_{K_0,n})$, and we recover Theorem 6.3.1. Furthermore, the oracle inequality in Theorem 6.3.2 shows that the selected variational approximation adaptively achieves the best upper bound among the different models $\mathcal{M}_K$, where each upper bound is a trade-off between two terms: a bias due to the error of approximating the true distribution by a distribution in model $\mathcal{M}_K$, and a variance term $r_{K,n}$ (as soon as the penalty term is lower than $r_{K,n}$) that is defined in Condition 6.2.

6.4 Application to probabilistic PCA

We consider here the probabilistic Principal Component Analysis (PCA) problem as an application of our work. From now on, matrices will be denoted in bold capital letters. We assume the model

$$X_i = \boldsymbol{W} Z_i + \sigma^2 \boldsymbol{I}_d$$

with i.i.d. Gaussian random variables $Z_i \sim \mathcal{N}(0, \boldsymbol{I}_K)$, where $\boldsymbol{I}_d$ and $\boldsymbol{I}_K$ are respectively the d- and K-dimensional identity matrices ($K < d$), $\boldsymbol{W} \in \mathbb{R}^{d \times K}$ is the K-rank matrix that contains the principal axes and σ^2 is a noisy term that is known. We suppose here that data are centred. Hence, the distribution of each X_i is

$$P_{\boldsymbol{W}} := \mathcal{N}(0, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d).$$

We are not interested here in estimating the principal axes $\boldsymbol{W}$ and selecting the number of components K, but in estimating the true distribution of the X_i's.

Each model corresponds to a rank K. We place an equal prior weight over each integer $K = 1, ..., d$. Hence the optimization problem is equivalent to maximizing the ELBO as in Blei et al. (2017). Given rank K, we place a prior over the K-rank matrix $\boldsymbol{W}$ to infer a distribution over principal axes. We choose independent Gaussian priors $\mathcal{N}(0, s^2 \boldsymbol{I}_d)$ on the columns $W_1, ..., W_K$ of $\boldsymbol{W}$. We also consider Gaussian independent variational approximations $\mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j)$ for the columns of $\boldsymbol{W}$. Then, as soon as there exists a true model, i.e. there exists K_0 and $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K_0}$ such that the true distribution of each X_i is $P_{\boldsymbol{W}_0} = \mathcal{N}(0, \boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)$, under the assumption that the coefficients of $\boldsymbol{W}_0$ are bounded, then Theorem 6.4.1 provides an explicit rate of convergence of our variational estimator even when K_0 is unknown:

Theorem 6.4.1. *For any $\alpha \in (0, 1)$, as soon as there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K_0}$ and such that the coefficients of $\boldsymbol{W}_0$ are bounded, then:*

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{W}}, P_{\boldsymbol{W}_0}) \widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\,\boldsymbol{W} | X_1^n) \right] = \mathcal{O}\left(\frac{dK_0 \log(dn)}{n} \right).$$

The proof as well as the computation of the ELBO are detailed in the appendix. Note that this corollary can directly lead to a result in Frobenius distance between covariance matrices $\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d$ and $\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d$ instead of the Rényi divergence between the corresponding distributions even when $\boldsymbol{W}$ and $\boldsymbol{W}_0$ are not equal-sized matrices. We denote $\|.\|_F$ the Frobenius norm and $\|.\|_2$ the spectral norm of a matrix, which are respectively defined as the square root of the sum of the absolute squares of the elements of a matrix and as its largest singular value.

The following corollary assesses the consistency of the selected variational approximation to the true covariance matrix in Frobenius norm. The idea, borrowed from Alquier and Ridgway (2017), is to project matrices onto some set of bounded matrices under the assumption that the spectral norm of the true matrix $\boldsymbol{W}_0$ is also bounded:

Corollary 6.4.2. *For any $\alpha \in (0, 1)$, as soon as there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K_0}$ and such that the spectral norm of $\boldsymbol{W}_0$ is upper bounded by*

a positive constant $B > 0$, then:

$$\mathbb{E}\left[\int \left\| clip_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0\boldsymbol{W}_0^T \right\|_F^2 \widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\boldsymbol{W}|X_1^n)\right] = \mathcal{O}\left(\frac{dK_0\log(dn)}{n}\right)$$

where $clip_B(\boldsymbol{A})$ *is the matrix which* (i,j)*-entry is equal to* $\begin{cases} \boldsymbol{A}_{i,j} & \text{if } |\boldsymbol{A}_{i,j}| \leq B^2 \\ B^2 & \text{if } \boldsymbol{A}_{i,j} \geq B^2 \\ -B^2 & \text{otherwise.} \end{cases}$

The requirement in our corollary is that the spectral norm of the true matrix $\boldsymbol{W}_0$ is bounded by some positive constant B, which implies the boundedness of the coefficients of the matrix as required in Theorem 6.4.1. In particular, the coefficients of the matrix $\boldsymbol{W}_0\boldsymbol{W}_0^T$ are bounded by B^2:

$$\begin{aligned} |(\boldsymbol{W}_0\boldsymbol{W}_0^T)_{i,j}| = \left|\sum_{k=1}^{K_0}(\boldsymbol{W}_0)_{i,k}(\boldsymbol{W}_0)_{j,k}\right| &\leq \left(\sum_{k=1}^{K_0}(\boldsymbol{W}_0)_{i,k}^2\right)^{1/2}\left(\sum_{k=1}^{K_0}(\boldsymbol{W}_0)_{j,k}^2\right)^{1/2} \\ &= \frac{\|\boldsymbol{W}_0 e_i\|_2}{\|e_i\|_2}\frac{\|\boldsymbol{W}_0 e_j\|_2}{\|e_j\|_2} \leq \|\boldsymbol{W}_0\|_2^2 \leq B^2 \end{aligned}$$

using Cauchy-Schwarz inequality and the property $\|\boldsymbol{W}_0\|_2 = \max_{x\neq 0}\frac{\|\boldsymbol{W}_0 x\|_2}{\|x\|_2}$ where e_ℓ is the vector of $\mathbb{R}^d$ which components are all equal to 0 except for the ℓ-th one that is set to 1. Hence it seems sensible to project (with respect to the Frobenius distance) any estimator $\boldsymbol{W}\boldsymbol{W}^T$ onto the set of all matrices whose entries lie in the interval $[-B^2, B^2]$, which is exactly what the $clip_B$ application does. Note that the spectral norm of Matrix $\boldsymbol{W}_0$ is equal to the largest eigenvalue of $\boldsymbol{W}_0\boldsymbol{W}_0^T$, so our assumption comes back to upper bounding the eigenvalues of the covariance matrix $\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d$, which is a classical assumption when estimating covariance matrices, see for instance Cai et al. (2015).

It is also possible to obtain a consistent pointwise covariance matrix estimator with the same convergence rate:

Corollary 6.4.3. *For any $\alpha \in (0,1)$, as soon as there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K_0}$ and such that the spectral norm of $\boldsymbol{W}_0$ is bounded by B. Let us define a pointwise estimator of the covariance matrix:*

$$\widehat{\boldsymbol{\Sigma}} = \int clip_B(\boldsymbol{W}\boldsymbol{W}^T)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\boldsymbol{W}|X_1^n) + \sigma^2\boldsymbol{I}_d.$$

Then,

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\Sigma}} - (\boldsymbol{W}_0\boldsymbol{W}_0^T + \sigma^2\boldsymbol{I}_d)\right\|_F^2\right] = \mathcal{O}\left(\frac{dK_0\log(dn)}{n}\right).$$

Discussion

In this chapter we proved the consistency of ELBO maximization in model selection. By penalizing the variational lower bound using our prior beliefs over the different models, we showed that under mild conditions, the variational approximation associated with

the selected model is consistent at the same convergence rate than the approximation associated with the true model. Moreover, the oracle inequality in Theorem 6.3.2 proved that the selected approximation is robust to misspecification. An application to the selection of the number of principal components in probabilistic PCA was provided as a short example.

We discuss in Appendix 6.5.1 the connection between the notions of *consistency* and *concentration*. This justifies the use of the α parameter in the definition of the evidence lower bound, as the regular posterior distribution is not robust to model misspecification. Indeed, authors of Grünwald et al. (2017) explain that there are pathologic cases where the regular posterior does not concentrate to the true distribution.

A point of interest when dealing with model selection is the question of recovering the true model (when it exists). This issue falls beyond the scope of this chapter which treats the question of estimating the true distribution, and can be the object of future works. The true model recovery would require stronger assumptions, but the implementation in Section 5 in Bishop (1999) suggests that those may hold for probabilistic PCA.

Also, it would be interesting to study cross-validation instead of ELBO maximization. However, the tools used in this work such as the theory of penalized criteria and oracle inequalities were particularly suited to the ELBO, and thus a different theory should be used in order to obtain the consistency of validation log-likelihood in the VB framework. This question is left for future research.

6.5 Proofs and additional results

6.5.1 Connection between consistency and concentration.

In this appendix, we highlight the connection between the notions of *consistency* used in Alquier and Ridgway (2017) and Chérief-Abdellatif and Alquier (2018) and *concentration*. We consider a true model $\mathcal{M}_{K_0}$ to which the true distribution $P^0 = P_{\theta^0}$ belongs, $\theta^0 \in \Theta_{K_0}$. We recall that the Bayesian estimator $\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(.|X_1^n)$ is said to be consistent if, in expectation (with respect to the random variables distributed according to P^0), the average Rényi loss between a distribution in the selected model and the true distribution (over the Bayesian estimator) goes to zero as $n \to +\infty$:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right] \xrightarrow[n\to+\infty]{} 0.$$

Similarly, we give the definition of *concentration* at rate s_n of the selected variational approximation to P^0 as stated in Ghosal et al. (2000), that is, in probability (with respect to the random variables distributed according to P^0), the approximation concentrates asymptotically around the true distribution as $n \to +\infty$, i.e. in probability:

$$\widetilde{\pi}^{\widehat{K}}_{n,\alpha}\left(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\right) \xrightarrow[n\to+\infty]{} 0$$

for any constant $M > 0$. The reference metric here is the α-Rényi divergence.

We show in this appendix that the consistency of the selected variational approximation to P^0 at rate r_n implies the concentration of the selected variational approximation to P^0 at any rate s_n such that $r_n = o(s_n)$ and $s_n \to 0$ as $n \to +\infty$, as for instance $s_n = r_n \log(\log(n))$ when the consistency rate r_n is slower than a log-logarithmic one.

To do so, we assume that the selected variational approximation is consistent to P^0 at rate r_n, i.e.:

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\right] \leq r_n.$$

Then, using Markov's inequality for any s_n such that $r_n = o(s_n)$ and $s_n \to 0$ and any constant $M > 0$:

$$\mathbb{E}\left[\widetilde{\pi}_{n,\alpha}^{\widehat{K}}\left(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\right)\right] \leq \frac{\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\right]}{Ms_n} \leq \frac{r_n}{Ms_n} \xrightarrow[n\to+\infty]{} 0.$$

Hence, we obtain the convergence in mean of $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}\left(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\right)$ to 0, which implies the convergence in probability of $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}\left(D_\alpha(P_\theta, P^0) > Ms_n|X_1^n\right)$ to 0, i.e. the concentration of $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n)$ to P^0 at rate s_n.

6.5.2 Proof of Theorem 6.3.1.

First, we need Donsker and Varadhan's famous variational formula. Refer for example to Catoni (2007) for a proof (Lemma 1.1.3).

Lemma 6.5.1. *For any probability λ on some measurable space $(\boldsymbol{E}, \mathcal{E})$ and any measurable function $h : \boldsymbol{E} \to \mathbb{R}$ such that $\int \mathrm{e}^h \mathrm{d}\lambda < \infty$,*

$$\log \int \mathrm{e}^h \mathrm{d}\lambda = \sup_{\rho \in \mathcal{M}_1^+(\boldsymbol{E})} \left\{ \int h \mathrm{d}\rho - KL(\rho, \lambda) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, if h is upper-bounded on the support of λ, then the supremum on the right-hand side is reached by the distribution of the form:

$$\lambda_h(d\beta) = \frac{\mathrm{e}^{h(\beta)}}{\int \mathrm{e}^h \mathrm{d}\lambda} \lambda(\mathrm{d}\beta).$$

We come back to the proof of Theorem 6.3.1. We adapt the proof of Theorem 4.1 in Chérief-Abdellatif and Alquier (2018).

Proof. For any $\alpha \in (0, 1)$ and $\theta \in \Omega := \cup_{K \geq 1} \Theta_K$, using the definition of Rényi divergence and $D_\alpha(P^{\otimes n}, R^{\otimes n}) = nD_\alpha(P, R)$ as data are i.i.d.:

$$\mathbb{E}\left[\exp\left(- \alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0)\right)\right] = 1$$

where $r_n(P_\theta, P^0) = \sum_{i=1}^n \log(P^0(X_i)/P_\theta(X_i))$ is the negative log-likelihood ratio. Then we integrate and use Fubini's theorem,

$$\mathbb{E}\bigg[\int \exp\bigg(-\alpha r_n(P_\theta, P^0) + (1-\alpha) n D_\alpha(P_\theta, P^0)\bigg)\pi(d\theta)\bigg] = 1.$$

Using Lemma 6.5.1,

$$\mathbb{E}\bigg[\exp\bigg(\sup_{\rho\in\mathcal{M}_1^+(\Omega)}\bigg\{\int\bigg(-\alpha r_n(P_\theta, P^0)+(1-\alpha)nD_\alpha(P_\theta, P^0)\bigg)\rho(d\theta)-\mathrm{KL}(\rho,\pi)\bigg\}\bigg)\bigg] = 1.$$

Then, using Jensen's inequality,

$$\mathbb{E}\bigg[\sup_{\rho\in\mathcal{M}_1^+(\Omega)}\bigg\{\int\bigg(-\alpha r_n(P_\theta, P^0) + (1-\alpha)nD_\alpha(P_\theta, P^0)\bigg)\rho(d\theta) - \mathrm{KL}(\rho,\pi)\bigg\}\bigg] \le 0.$$

Now, we consider $\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n)$ as a distribution on $\mathcal{M}_1^+(\Omega)$ with all its mass on $\Theta_{\widehat{K}}$,

$$\mathbb{E}\bigg[\int\bigg(-\alpha r_n(P_\theta, P^0) + (1-\alpha)nD_\alpha(P_\theta, P^0)\bigg)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n) - \mathrm{KL}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\pi)\bigg] \le 0.$$

We use $\mathrm{KL}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\pi) = \mathrm{KL}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\Pi_{\widehat{K}}) + \log(\frac{1}{\pi_{\widehat{K}}})$, and we rearrange terms:

$$\mathbb{E}\bigg[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\bigg]$$
$$\le \mathbb{E}\bigg[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n) + \frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(.|X_1^n),\Pi_{\widehat{K}})}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_{\widehat{K}}})}{n(1-\alpha)}\bigg].$$

By definition of $\widehat{K}$,

$$\mathbb{E}\bigg[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\bigg]$$
$$\le \mathbb{E}\bigg[\inf_{K\ge1}\bigg\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n) + \frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\bigg\}\bigg]$$

which gives

$$\mathbb{E}\bigg[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\bigg]$$
$$\le \inf_{K\ge1}\bigg\{\mathbb{E}\bigg[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\widetilde{\pi}_{n,\alpha}^{K}(d\theta|X_1^n) + \frac{\mathrm{KL}(\widetilde{\pi}_{n,\alpha}^{K}(.|X_1^n),\Pi_K)}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\bigg]\bigg\}$$

and hence, by definition of $\widetilde{\pi}_{n,\alpha}^{K}(.|X_1^n)$,

$$\mathbb{E}\bigg[\int D_\alpha(P_\theta, P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\bigg]$$

$$\leq \inf_{K\geq 1}\left\{\mathbb{E}\left[\inf_{\rho\in\mathcal{F}_K}\left\{\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\rho(d\theta)+\frac{\mathrm{KL}(\rho,\Pi_K)}{n(1-\alpha)}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right]\right\}.$$

which leads to,

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right]$$
$$\leq \inf_{K\geq 1}\inf_{\rho\in\mathcal{F}_K}\left\{\mathbb{E}\left[\frac{\alpha}{1-\alpha}\int\frac{r_n(P_\theta,P^0)}{n}\rho(d\theta)+\frac{\mathrm{KL}(\rho,\Pi_K)}{n(1-\alpha)}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right]\right\}.$$

Finally,

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right]$$
$$\leq \inf_{K\geq 1}\left\{\inf_{\rho_K\in\mathcal{F}_K}\left\{\frac{\alpha}{1-\alpha}\int\mathrm{KL}(P^0,P_{\theta_K})\rho_K(d\theta_K)+\frac{\mathrm{KL}(\rho_K,\Pi_K)}{n(1-\alpha)}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

The theorem is a direct corollary of this inequality as soon as Assumption 6.1 is satisfied.

□

6.5.3 Proof of Theorem 6.3.2.

Proof. Fix $\alpha\in(0,1)$ and let us prove Theorem 6.3.2. Let us recall that $\Theta_K(r_{K,n})$ is defined as the set of parameters $\theta_K^*\in\Theta_K$, for which there is a distribution $\rho_{K,n}\in\mathcal{F}_K$ such that:

$$\int\mathbb{E}\left[\log\frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right]\rho_{K,n}(d\theta_K)\leq r_{K,n}\ \text{ and }\ \mathrm{KL}(\rho_{K,n},\Pi_K)\leq nr_{K,n}.$$

We begin from:

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right]$$
$$\leq \inf_{K\geq 1}\left\{\inf_{\rho_K\in\mathcal{F}_K}\left\{\frac{\alpha}{1-\alpha}\int\mathrm{KL}(P^0,P_{\theta_K})\rho_K(d\theta_K)+\frac{\mathrm{KL}(\rho_K,\Pi_K)}{n(1-\alpha)}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

Then, we write for any K, any $\theta_K\in\Theta_K$, $\theta_K^*\in\Theta_K$:

$$\mathrm{KL}(P^0,P_{\theta_K})=\mathrm{KL}(P^0,P_{\theta_K^*})+\mathbb{E}\left[\log\frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right]$$

which gives:

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(d\theta|X_1^n)\right]$$

$$\leq \inf_{K\geq 1}\left\{\inf_{\theta_K^*\in\Theta_K}\left\{\frac{\alpha}{1-\alpha}\mathrm{KL}(P^0,P_{\theta_K^*})+\inf_{\rho_K\in\mathcal{F}_K}\left\{\frac{\alpha}{1-\alpha}\int\mathbb{E}\left[\log\frac{P_{\theta_K^*}(X_i)}{P_{\theta_K}(X_i)}\right]\rho_K(d\theta_K)\right.\right.\right.$$
$$\left.\left.\left.+\frac{\mathrm{KL}(\rho_K,\Pi_K)}{n(1-\alpha)}\right\}\right\}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

Hence, using the definition of $\Theta_K(r_{K,n})$ and upper bounding the right-hand-side of the previous inequality by an inf over $\Theta_K(r_{K,n})$, we conclure:

$$\mathbb{E}\left[\int D_\alpha(P_\theta,P^0)\widetilde{\pi}_{n,\alpha}^{\widehat{K}}(d\theta|X_1^n)\right]\leq\inf_{K\geq 1}\left\{\frac{\alpha}{1-\alpha}\inf_{\theta^*\in\Theta_K(r_{K,n})}\mathrm{KL}(P^0,P_{\theta_K^*})+\frac{1+\alpha}{1-\alpha}r_{K,n}+\frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}\right\}.$$

□

6.5.4 Proof of Theorem 6.4.1.

Proof. We still consider the framework of probabilistic PCA in Section 6.4. We assume that there exists a true rank K_0 and a matrix $\boldsymbol{W}_0\in\mathbb{R}^{d\times K_0}$ with bounded coefficients such that the true distribution of each X_i is $\mathcal{N}(0,\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)$, and we place a prior $\Pi_{K_0}=\mathcal{N}(0,s^2\boldsymbol{I}_d)^{\otimes K_0}$ and a variational approximation $\rho_{K_0}=\rho^{\otimes K_0}$ on W given $K=K_0$ where we denote $\rho=\mathcal{N}(0,\frac{1}{dn^2}\boldsymbol{I}_d)$. We recall that $\pi_K=\frac{1}{d}$ for any $K=1,...,d$.

To obtain the rate of convergence $r_n=dK_0\log(nd)/n$ for probabilistic PCA, we just need to show that the quantities in Assumption 6.1 are upper bounded by r_n (up to a constant) as we have $\log(1/\pi_{K_0})/n$ much smaller than r_n:

$$\int\mathrm{KL}\Big(\mathcal{N}(0,\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d),\mathcal{N}(0,\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)\Big)\rho_{K_0}(d\theta_K)\quad,\quad\frac{\mathrm{KL}(\rho_{K_0},\Pi_{K_0})}{n}.$$

We have two terms. The first one, i.e. the Kullback-Leibler term, provides a rate of convergence of $dK_0\log(dn)/n$ as:

$$\begin{aligned}\mathrm{KL}(\rho_{K_0},\Pi_{K_0})&=\sum_{j=1}^{K_0}\mathrm{KL}\Big(\mathcal{N}(0,\frac{1}{dn^2}\boldsymbol{I}_d),\mathcal{N}(0,s^2\boldsymbol{I}_d)\Big)\\&=\frac{K_0}{2}\left(\frac{1}{n^2s^2}-d+d\log(s^2)+d\log(dn^2)\right)\\&\leq\frac{K_0}{2n^2s^2}-\frac{dK_0}{2}+\frac{dK_0\log(s^2)}{2}+dK_0\log(dn).\end{aligned}$$

The integral is much more complicated to deal with. We will show that it leads to a rate faster than $dK_0\log(dn)/n$. If we denote $\mathbb{E}$ the expectation with respect to ρ_{K_0}, then the integral will be equal to:

$$\frac{1}{2}\mathbb{E}\left[\mathrm{Tr}\Big((\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)^{-1}(\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)\Big)\right]-\frac{d}{2}+\frac{1}{2}\mathbb{E}\left[\log\left(\frac{\det(\boldsymbol{W}\boldsymbol{W}^T+\sigma^2\boldsymbol{I}_d)}{\det(\boldsymbol{W}_0\boldsymbol{W}_0^T+\sigma^2\boldsymbol{I}_d)}\right)\right].$$

The expectation of the log-ratio is easy to upper bound. We denote $\lambda_1, ..., \lambda_d$ the positive eigenvalues of the positive definite matrix $\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d$. Then for each $j = 1, ..., d$, $\lambda_j \geq \sigma^2$ and using Jensen's inequality and the log-concavity of the determinant:

$$\begin{aligned}
\mathbb{E}\left[\log\left(\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)\right)\right] &\leq \log\left(\det\left(\mathbb{E}[\boldsymbol{W}\boldsymbol{W}^T] + \sigma^2 \boldsymbol{I}_d\right)\right) \\
&= \log\left(\det\left(\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d + \frac{1}{dn^2}\boldsymbol{I}_d\right)\right) \\
&= \sum_{j=1}^{d} \log\left(\lambda_j + \frac{1}{dn^2}\right) \\
&= \sum_{j=1}^{d} \log(\lambda_j) + \sum_{j=1}^{d} \log\left(1 + \frac{1}{\lambda_j dn^2}\right) \\
&= \mathbb{E}\left[\log\left(\det(\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)\right)\right] + \sum_{j=1}^{d} \log\left(1 + \frac{1}{\lambda_j dn^2}\right) \\
&\leq \mathbb{E}\left[\log\left(\det(\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)\right)\right] + \sum_{j=1}^{d} \frac{1}{\lambda_j dn^2} \\
&\leq \mathbb{E}\left[\log\left(\det(\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)\right)\right] + \frac{1}{n^2\sigma^2}
\end{aligned}$$

and then the expectation of the log-ratio provides a rate of convergence of $1/n^2$:

$$\mathbb{E}\left[\log\left(\frac{\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)}{\det(\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)}\right)\right] \leq \frac{1}{n^2\sigma^2}.$$

The remainder can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Tr}\left((\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)^{-1} (\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)\right)\right] - d \\
&\qquad = \mathbb{E}\left[\mathrm{Tr}\left((\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)^{-1} (\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T)\right)\right] \\
&\qquad \leq \mathbb{E}\left[\|(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)^{-1}\|_F \times \|\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F \right] \\
&\qquad \leq \sqrt{d}\mathbb{E}\left[\|(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)^{-1}\|_2 \times \|\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F \right] \\
&\qquad = \sqrt{d}\mathbb{E}\left[\sigma_{\max}\left((\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)^{-1}\right) \times \|\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F \right] \\
&\qquad = \sqrt{d}\mathbb{E}\left[\sigma_{\min}(\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)^{-1} \times \|\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F \right]
\end{aligned}$$

i.e.

$$\mathbb{E}\left[\mathrm{Tr}\left((\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)^{-1} (\boldsymbol{W}_0 \boldsymbol{W}_0^T + \sigma^2 \boldsymbol{I}_d)\right)\right] - d \leq \sqrt{d}\mathbb{E}\left[(\sigma^2)^{-1} \times \|\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F \right]$$

$$= \frac{\sqrt{d}}{\sigma^2}\mathbb{E}\Big[\| \boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F\Big]$$

where $\|.\|_F$ is the Frobenius norm on matrices, $\|.\|_2$ the spectral norm, and $\sigma_{\min}(\boldsymbol{A})$, $\sigma_{\max}(\boldsymbol{A})$ the lowest and largest singular values of a matrix $\boldsymbol{A}$. We use the fact that for a symmetric semi-definite positive matrix: $\sigma_{\max}(\boldsymbol{A}^{-1}) = \Big(\sigma_{\min}(\boldsymbol{A})\Big)^{-1}$ and $\sigma_{\min}(\boldsymbol{A}+\sigma^2\boldsymbol{I}_d) \geq \sigma^2$, as well as the inequality $\|\boldsymbol{A}\|_F \leq \sqrt{d}\|\boldsymbol{A}\|_2$ for any $d \times d$ matrix $\boldsymbol{A}$.

The only thing left to do is to upper bound the expectation of the Frobenius norm of $\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T$ by a multiple of $\frac{\sqrt{d}K_0 \log(dn)}{n}$. We use the triangle and Cauchy-Schwarz's inequalities:

$$\begin{aligned}
\mathbb{E}\Big[\| \boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F\Big] &\leq \mathbb{E}\Big[\| \boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}\boldsymbol{W}_0^T\|_F\Big] + \mathbb{E}\Big[\| \boldsymbol{W}\boldsymbol{W}_0^T - \boldsymbol{W}_0 \boldsymbol{W}_0^T\|_F\Big] \\
&\leq \mathbb{E}\Big[\| \boldsymbol{W}(\boldsymbol{W} - \boldsymbol{W}_0)^T\|_F\Big] + \mathbb{E}\Big[\|(\boldsymbol{W} - \boldsymbol{W}_0)\boldsymbol{W}_0^T\|_F\Big] \\
&\leq \mathbb{E}\Big[\| \boldsymbol{W}\|_F\| \boldsymbol{W} - \boldsymbol{W}_0\|_F\Big] + \mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F\| \boldsymbol{W}_0\|_F\Big] \\
&\leq \sqrt{\mathbb{E}\Big[\| \boldsymbol{W}\|_F^2\Big]\mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2\Big]} + \sqrt{\mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2\Big]\mathbb{E}\Big[\| \boldsymbol{W}_0\|_F^2\Big]} \\
&\leq \sqrt{\mathbb{E}\Big[\| \boldsymbol{W}\|_F^2\Big]\mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2\Big]} + \| \boldsymbol{W}_0\|_F\sqrt{\mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2\Big]}.
\end{aligned}$$

We can upper bound $\| \boldsymbol{W}_0\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^{K_0} (\boldsymbol{W}_0)_{i,j}^2}$ by $\sqrt{dK_0}C$ where C is an upper bound on each of the coefficients of matrix $\boldsymbol{W}_0$.

Also, we can notice that $dn^2\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2 = \sum_{i=1}^d \sum_{j=1}^{K_0} \Big(\sqrt{d}n(\boldsymbol{W}_{i,j} - (\boldsymbol{W}_0)_{i,j})\Big)^2$ is a sum of squares of independent standard normal random variables. Thus $dn^2\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2$ follows a chi-squared distribution with dK_0 degrees of freedom and its expectation is equal to dK_0. Hence:

$$\mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2\Big] = \frac{K_0}{n^2}.$$

Similarly, as $\boldsymbol{W}_{i,j} - (\boldsymbol{W}_0)_{i,j}$ is centered, we get:

$$\begin{aligned}
\mathbb{E}\Big[\| \boldsymbol{W}\|_F^2\Big] &= \mathbb{E}\Big[\sum_{i=1}^d \sum_{j=1}^{K_0} \boldsymbol{W}_{i,j}^2\Big] \\
&= \sum_{i=1}^d \sum_{j=1}^{K_0} \mathbb{E}\Big[\Big(\boldsymbol{W}_{i,j} - (\boldsymbol{W}_0)_{i,j}\Big)^2 + (\boldsymbol{W}_0)_{i,j}^2 - 2(\boldsymbol{W}_0)_{i,j}\Big(\boldsymbol{W}_{i,j} - (\boldsymbol{W}_0)_{i,j}\Big)\Big] \\
&= \mathbb{E}\Big[\| \boldsymbol{W} - \boldsymbol{W}_0\|_F^2\Big] + \| \boldsymbol{W}_0\|_F^2 \\
&\leq \frac{K_0}{n^2} + dK_0C^2 \\
&= \left(dC^2 + \frac{1}{n^2}\right)K_0.
\end{aligned}$$

Thus, we obtain:

$$\begin{aligned}\mathbb{E}\Big[\| \boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T\|_F\Big] &\leq \frac{\sqrt{K_0}}{n}\sqrt{K_0}\sqrt{dC^2+\frac{1}{n^2}} + \sqrt{dK_0}C\frac{\sqrt{K_0}}{n} \\ &= \frac{K_0}{n}\sqrt{dC^2+\frac{1}{n^2}} + \frac{\sqrt{d}K_0 C}{n} \\ &\leq \frac{K_0}{n}\Big(\sqrt{d}C+\frac{1}{n}\Big) + \frac{\sqrt{d}K_0 C}{n} \\ &= \frac{K_0}{n}\Big(2\sqrt{d}C+\frac{1}{n}\Big).\end{aligned}$$

Hence, the order of the upper bound of the expectation of the Fobrenius norm of matrix $\boldsymbol{W}_0 \boldsymbol{W}_0^T - \boldsymbol{W}\boldsymbol{W}^T$ is $\frac{\sqrt{d}K_0}{n} < \frac{\sqrt{d}K_0 \log(dn)}{n}$.

Finally, the consistency rate associated with the integral term is $\frac{dK_0}{n}$, and the overall rate of convergence is $\frac{dK_0 \log(dn)}{n}$. □

6.5.5 Computation of the ELBO for probabilistic PCA.

We consider the framework of probabilistic PCA detailed in Section 6.4. Given rank K, we place independent Gaussian priors on the columns $W_1, ..., W_K$ of $\boldsymbol{W}$ such that $\Pi_K = \mathcal{N}(0, s^2\boldsymbol{I}_d)^{\otimes K}$, and Gaussian independent variational approximations $\mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j)$ for the columns of $\boldsymbol{W}$. The ELBO associated with rank K and variational approximation $\rho_K = \otimes_{j=1}^K \mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j)$ is given by:

$$\mathrm{ELBO}_K(\rho_K) = \alpha \int \ell_n(\boldsymbol{W})\rho_K(d\boldsymbol{W}) - \mathrm{KL}\Big(\rho_K, \Pi_K\Big).$$

The Kullback-Leibler term $\mathrm{KL}\Big(\rho_K, \Pi_K\Big)$ is equal to:

$$\frac{1}{2}\sum_{j=1}^K \left\{ \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_j)}{s^2} + \frac{\mu_j^T\mu_j}{s^2} - \log\Big(\det(\boldsymbol{\Sigma}_j)\Big)\right\} - \frac{dK}{2} + \frac{dK\log(s^2)}{2}$$

while the average log-likelihood $\int \ell_n(\boldsymbol{W})\rho_K(d\boldsymbol{W})$ is:

$$-\frac{dn}{2}\log(2\pi) - \frac{n}{2}\int \log\Big(\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)\Big)\rho_K(dW) - \frac{1}{2}\sum_{i=1}^n \int X_i^T(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)^{-1}X_i\, \rho_K(d\boldsymbol{W})$$

where both integrals can be computed thanks to Monte-Carlo sampling approximations:

$$\int \log\Big(\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)\Big)\rho_K(d\boldsymbol{W}) \approx \sum_{\ell=1}^N \log\Big(\det(\boldsymbol{W}^{(\ell)}\boldsymbol{W}^{(\ell)T} + \sigma^2\boldsymbol{I}_d)\Big)$$

and

$$\int X_i^T(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)^{-1}X_i\, \rho_K(d\boldsymbol{W}) \approx \sum_{\ell=1}^N X_i^T(\boldsymbol{W}^{(\ell)}\boldsymbol{W}^{(\ell)T} + \sigma^2\boldsymbol{I}_d)^{-1}X_i$$

where $\boldsymbol{W}^{(1)}, ..., \boldsymbol{W}^{(N)}$ are N i.i.d. data sampled from ρ_K.

The inverse matrix $(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}_d)^{-1}$ can be derived thanks to classical inversion algorithms. For instance, it is possible to do so in $\mathcal{O}(Kd^2)$ operations instead of the classical $\mathcal{O}(d^3)$ inversion procedure thanks to Sherman-Morrison formula: for any matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and vectors $u, v \in \mathbb{R}^d$ such that $\boldsymbol{A} + uv^T$ is invertible,

$$(\boldsymbol{A} + uv^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}uv^T\boldsymbol{A}^{-1}}{1 + v^T\boldsymbol{A}^{-1}u}.$$

We write

$$\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d = \sigma^2\boldsymbol{I}_d + \sum_{j=1}^{K} W_j W_j^T = \left(\sigma^2\boldsymbol{I}_d + \sum_{j=1}^{K-1} W_j W_j^T\right) + W_K W_K^T$$

and iterate K times Sherman-Morrison formula. The first time, we apply it to $\boldsymbol{A} = \sigma^2\boldsymbol{I}_d + \sum_{j=1}^{K-1} W_j W_j^T$ and $u = v = \boldsymbol{W}_K$, then to $\boldsymbol{A} = \sigma^2\boldsymbol{I}_d + \sum_{j=1}^{K-2} W_j W_j^T$ and $u = v = W_{K-1}$, and so on. We finally obtain $(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d)^{-1} = \boldsymbol{M}_K$ where:

$$\begin{cases} \boldsymbol{M}_0 = \sigma^2\boldsymbol{I}_d \\ \forall j = 1, ..., K, \quad \boldsymbol{M}_j = \boldsymbol{M}_{j-1} - \frac{1}{1+W_j^T Z_j} Z_j Z_j^T \text{ with } Z_j = \boldsymbol{M}_{j-1} W_j. \end{cases}$$

In order to compute the maximum value ELBO(K) of the ELBO associated with rank K, one can use a stochastic gradient descent on $(\mu_1, \boldsymbol{\Sigma}_1, ..., \mu_K, \boldsymbol{\Sigma}_K)$ that will converge to a local maximum and will give the variational estimator for rank K. Then, maximizing ELBO(K) over desired values of K leads to the optimal number of principal components and to the associated optimal variational approximation.

6.5.6 Results in matrix norm for probabilistic PCA.

To prove Corollaries 6.4.2 and 6.4.3, we need the two lemmas presented behind. We introduce some notations first. We refer the interested reader to Forth et al. (2014) for more details.

Notations : Let us call $\mathcal{S}_d^+$ the set of $d \times d$ symmetric positive semi-definite matrices, and $\mathcal{X}_M = \left\{\boldsymbol{A} \in \mathcal{S}_d^+ / \|\boldsymbol{A}\|_2 \leq M\right\}$. We define the vectorization of Matrix $\boldsymbol{A} \in \mathbb{R}^{p\times q}$ with columns $X_1, ..., X_q$:

$$\text{Vec}(\boldsymbol{A}) = (\boldsymbol{A}_1^T, ..., \boldsymbol{A}_q^T)^T \in \mathbb{R}^{p\times q}.$$

We define the Frobenius inner product of two matrices $\boldsymbol{A} \in \mathbb{R}^{p\times q}$ and $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{p\times q}$, that is the sum of componentwise products:

$$\boldsymbol{A} \cdot \widetilde{\boldsymbol{A}} = \text{Vec}(\boldsymbol{A})^T \text{Vec}(\widetilde{\boldsymbol{A}}).$$

Notice that $\|\boldsymbol{A}\|_F^2 = \boldsymbol{A} \cdot \boldsymbol{A} = \text{Vec}(\boldsymbol{A})^T \text{Vec}(\boldsymbol{A})$.

We also introduce the Kronecker and Box products of two matrices $\boldsymbol{A} \in \mathbb{R}^{p_1\times q_1}$ and $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{p_2\times q_2}$ which are respectively the matrices $\boldsymbol{A} \otimes \widetilde{\boldsymbol{A}} \in \mathbb{R}^{p_1 p_2 \times q_1 q_2}$ and $\boldsymbol{A} \boxtimes \widetilde{\boldsymbol{A}} \in \mathbb{R}^{p_1 p_2 \times q_1 q_2}$ such that their coefficients are defined as:

$$(\boldsymbol{A} \otimes \widetilde{\boldsymbol{A}})_{p_2(i-1)+j, q_2(k-1)+l} = \boldsymbol{A}_{i,k} \widetilde{\boldsymbol{A}}_{j,l},$$

$$(\boldsymbol{A} \boxtimes \widetilde{\boldsymbol{A}})_{p_2(i-1)+j,q_1(k-1)+l} = \boldsymbol{A}_{i,l}\widetilde{\boldsymbol{A}}_{j,k}$$

for any integers i, j, k, l such that $1 \leq i \leq p_1$, $1 \leq j \leq q_1$, $1 \leq k \leq p_2$, $1 \leq l \leq q_2$.

We have the following properties for any matrix $\boldsymbol{P}$:

$$(\boldsymbol{A} \otimes \widetilde{\boldsymbol{A}})\text{Vec}(\boldsymbol{P}) = \text{Vec}(\widetilde{\boldsymbol{A}}\boldsymbol{P}\boldsymbol{A}^T),$$
$$(\boldsymbol{A} \boxtimes \widetilde{\boldsymbol{A}})\text{Vec}(\boldsymbol{P}) = \text{Vec}(\widetilde{\boldsymbol{A}}\boldsymbol{P}^T\boldsymbol{A}^T).$$

We also define the gradient $\nabla f(\boldsymbol{A}) \in \mathbb{R}^{p\times q}$ and the Hessian $\nabla^2 f(\boldsymbol{A}) \in \mathbb{R}^{pq\times pq}$ of a differentiable function $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ at matrix $\boldsymbol{A}$:

$$(\nabla f(\boldsymbol{A}))_{p_2(i-1)+j,q_2(k-1)+l} = \frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}_{i,j}},$$
$$(\nabla^2 f(\boldsymbol{A}))_{p_2(j-1)+i,p_2(l-1)+k} = \frac{\partial^2 f(\boldsymbol{A})}{\partial \boldsymbol{A}_{i,j}\partial \boldsymbol{A}_{k,l}}$$

for any integers i, j, k, l such that $1 \leq i, k \leq p$, $1 \leq j, l \leq q$ where ∂f is the partial derivative of f.

We say that a differentiable function $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ is s-strongly convex in $\mathcal{S} \subset \mathbb{R}^{pq\times pq}$ with respect to the norm $\|.\|$ as soon as one of the two following equivalent properties is satisfied:

$$f(\boldsymbol{A}) \geq f(\widetilde{\boldsymbol{A}}) + \nabla f(\boldsymbol{A}) \cdot (\boldsymbol{A} - \widetilde{\boldsymbol{A}}) + \frac{s}{2}\|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|^2$$

or

$$\text{Vec}(\boldsymbol{P})^T \nabla^2 f(\boldsymbol{A})\text{Vec}(\boldsymbol{P}) \geq s\|\boldsymbol{P}\|^2$$

for any matrix $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{S}$ and any symmetric matrix $\boldsymbol{P} \in \mathbb{R}^{pq\times pq}$.

Lemma 6.5.2. *Then, function $f : \boldsymbol{A} \to -\log\Big(\det(\boldsymbol{A} + M\boldsymbol{I}_d)\Big)$ is $1/(M+\sigma^2)^2$ strongly convex in $\mathcal{X}_M$ with respect to the Frobenius norm.*

Proof. The proof follows the same steps than the proof of Theorem 3.1 in Moridomi et al. (2018).

The Hessian of function f at any symmetric matrix in $\boldsymbol{A} \in \mathcal{X}_M$ is given by (see Forth et al. (2014)):

$$\nabla^2 f(\boldsymbol{A}) = \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big)^T \boxtimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} = (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \boxtimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}.$$

Then, we have for any $\boldsymbol{A} \in \mathcal{X}_M$ and any symmetric matrix $\boldsymbol{P} \in \mathbb{R}^{pq\times pq}$:

$$\begin{aligned}
\text{Vec}(\boldsymbol{P})^T \nabla^2 f(\boldsymbol{A})\text{Vec}(\boldsymbol{P}) &= \text{Vec}(\boldsymbol{P})^T \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \boxtimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big)\text{Vec}(\boldsymbol{P}) \\
&= \text{Vec}(\boldsymbol{P})^T \text{Vec}\Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\boldsymbol{P}^T(\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big) \\
&= \text{Vec}(\boldsymbol{P})^T \text{Vec}\Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\boldsymbol{P}(\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}\Big)
\end{aligned}$$

$$= \mathrm{Vec}(\boldsymbol{P})^T \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \otimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \Big) \mathrm{Vec}(\boldsymbol{P}).$$

Note that the eigenvalues of a Kronecker product $\boldsymbol{A} \otimes \boldsymbol{P}$ are the products of an eigenvalue of $\boldsymbol{A}$ and an eigenvalue of $\boldsymbol{P}$, and the eigenvalues of $\boldsymbol{P}^{-1}$ are the inverse of the eigenvalues of $\boldsymbol{P}$. Moreover, the maximum eigenvalue of $\boldsymbol{A} + M\boldsymbol{I}_d$ is $\|\boldsymbol{A}\|_2 + \sigma^2$, so the minimum eigenvalue of $(\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \otimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1}$ is equal to $(\|\boldsymbol{A}\|_2 + \sigma^2)^{-2}$. Hence, for any matrix $\boldsymbol{A} \in \mathcal{X}_M$, we get:

$$\begin{aligned}\mathrm{Vec}(\boldsymbol{P})^T \Big((\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \otimes (\boldsymbol{A} + M\boldsymbol{I}_d)^{-1} \Big) \mathrm{Vec}(\boldsymbol{P})^T &\geq (\|\boldsymbol{A}\|_2 + \sigma^2)^{-2} \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}(\boldsymbol{P}) \\ &\geq \frac{1}{(M+\sigma^2)^2} \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}(\boldsymbol{P}),\end{aligned}$$

and we conclude using the definition of the strong convexity and $\|\boldsymbol{P}\|_F^2 = \mathrm{Vec}(\boldsymbol{P})^T \mathrm{Vec}(\boldsymbol{P})$.

□

Lemma 6.5.3. *For any $\alpha \in (0,1)$ and any matrices $\boldsymbol{W} \in \mathbb{R}^{d \times K_1}$ and $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{d \times K_2}$, as soon as the spectral norms of $\boldsymbol{W}\boldsymbol{W}^T$ and $\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^T$ are bounded by a constant B^2, then:*

$$D_\alpha(P_{\boldsymbol{W}}, P_{\widetilde{\boldsymbol{W}}}) \geq \frac{\alpha}{16(B^2+\sigma^2)^2} \big\| \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^T - \boldsymbol{W}\boldsymbol{W}^T \big\|_F^2.$$

Proof. We recall that function $f : \boldsymbol{A} \to -\log\Big(\det(\boldsymbol{A} + M\boldsymbol{I}_d)\Big)$ is $1/(M+\sigma^2)^2$ strongly convex in $\mathcal{X}_M$ with respect to the Fobrenius norm according to Lemma 6.5.2. Hence, for any matrices $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$ in $\mathcal{X}_M$, we have:

$$\begin{aligned}-\log\Big(\det((1-\alpha)\boldsymbol{A} + \alpha\widetilde{\boldsymbol{A}})\Big) \leq -(1-\alpha)\log\Big(\det(\boldsymbol{A})\Big) - \alpha\log\Big(\det(\widetilde{\boldsymbol{A}})\Big) \\ - \frac{1}{2}\alpha(1-\alpha)\frac{1}{4M^2}\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_F^2.\end{aligned}$$

We rearrange terms:

$$\log\left(\frac{\det\Big((1-\alpha)\boldsymbol{A} + \alpha\widetilde{\boldsymbol{A}}\Big)}{\det(\boldsymbol{A})^{1-\alpha}\det(\widetilde{\boldsymbol{A}})^{\alpha}}\right) \geq \frac{\alpha(1-\alpha)}{8M^2}\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_F^2.$$

Now, we use the fact that:

$$D_\alpha\Big(\mathcal{N}(0,\boldsymbol{A}), \mathcal{N}(0,\widetilde{\boldsymbol{A}})\Big) = \frac{1}{2(1-\alpha)}\log\left(\frac{\det\Big((1-\alpha)\boldsymbol{A} + \alpha\widetilde{\boldsymbol{A}}\Big)}{\det(\boldsymbol{A})^{1-\alpha}\det(\widetilde{\boldsymbol{A}})^{\alpha}}\right)$$

to get for any matrices $\boldsymbol{W} \in \mathbb{R}^{d \times K_1}$ and $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{d \times K_2}$ such that $\|\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^T + \sigma^2\boldsymbol{I}_d\|_2 \leq M$ and $\|\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d\|_F \leq M$:

$$D_\alpha\Big(P_{\boldsymbol{W}}, P_{\widetilde{\boldsymbol{W}}}\Big) \geq \frac{\alpha}{16M^2}\|\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^T + \sigma^2\boldsymbol{I}_d - \boldsymbol{W}\boldsymbol{W}^T - \sigma^2\boldsymbol{I}_d\|_F^2.$$

Moreover, for any matrix $\boldsymbol{W} \in \mathbb{R}^{d \times K}$ such that the spectral norm of $\boldsymbol{W}\boldsymbol{W}^T$ is bounded by B^2, we have $\|\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_d\|_2 \leq B^2 + \sigma^2$. We conclude using the previous inequality for $M = B^2 + \sigma^2$.

□

Now, let us go back to the proof of Corollary 6.4.2.

Proof. We assume that there exists a true model $\mathcal{M}_{K_0}$ such that $P^0 = P_{\boldsymbol{W}_0}$ with $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K_0}$ and such that the spectral norm of $\boldsymbol{W}_0$ is bounded by B (hence the coefficients of $\boldsymbol{W}_0$ are also bounded). As clip_B is a projection onto a closed convex set with respect to the Frobenius norm, we have for any matrix $\boldsymbol{W} \in \mathbb{R}^{d\times \widehat{K}}$:

$$\left\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T) - \mathrm{clip}_B(\boldsymbol{W}_0\boldsymbol{W}_0^T)\right\|_F \leq \left\|\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}_0\boldsymbol{W}_0^T\right\|_F$$

and as the coefficients of $\boldsymbol{W}_0\boldsymbol{W}_0^T$ are bounded by B^2:

$$\left\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0\boldsymbol{W}_0^T\right\|_F = \left\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T) - \mathrm{clip}_B(\boldsymbol{W}_0\boldsymbol{W}_0^T)\right\|_F.$$

According to Lemma 6.5.3, we get for any matrix $\boldsymbol{W} \in \mathbb{R}^{d\times \widehat{K}}$:

$$\left\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0\boldsymbol{W}_0^T\right\|_F^2 \leq \frac{16(B^2+\sigma^2)^2}{\alpha} D_\alpha\Big(P_{\boldsymbol{W}_1}, P_{\boldsymbol{W}_2}\Big).$$

Thus:

$$\mathbb{E}\left[\int \left\|\mathrm{clip}_B(\boldsymbol{W}\boldsymbol{W}^T) - \boldsymbol{W}_0\boldsymbol{W}_0^T\right\|_F^2 \widetilde{\pi}^{\widehat{K}}_{n,\alpha}(dW|X_1^n)\right] \leq \frac{16(B^2+\sigma^2)^2}{\alpha}\mathbb{E}\left[\int D_\alpha(P_W, P_{\boldsymbol{W}_0})\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(dW|X_1^n)\right]$$

and we use Theorem 6.4.1:

$$\mathbb{E}\left[\int D_\alpha(P_{\boldsymbol{W}}, P_{\boldsymbol{W}_0})\widetilde{\pi}^{\widehat{K}}_{n,\alpha}(dW|X_1^n)\right] = \mathcal{O}\left(\frac{dK_0\log(dn)}{n}\right).$$

which ends the proof.

□

We can obtain Corollary 6.4.3 using a simple convexity argument.

Part III

Theoretical bounds for online variational inference algorithms

Chapter 7

A Generalization Bound for Online Variational Inference

Bayesian inference provides an attractive online-learning framework to analyze sequential data, and offers generalization guarantees which hold even under model mismatch and with adversaries. Unfortunately, exact Bayesian inference is rarely feasible in practice and approximation methods are usually employed, but do such methods preserve the generalization properties of Bayesian inference? In this chapter, we show that this is indeed the case for some variational inference (VI) algorithms. We propose new online, tempered VI algorithms and derive their generalization bounds. Our theoretical result relies on the convexity of the variational objective, but we argue that our result should hold more generally and present empirical evidence in support of this. Our work in this chapter presents theoretical justifications in favor of online algorithms that rely on approximate Bayesian methods.

7.1 Introduction

Bayesian methods, such as Kalman Filtering (Kalman, 1960), Hidden Markov Model (Baum and Petrie, 1966) and Particle Filtering (Doucet and Johansen, 2009), are popular methods to analyze sequential data. The posterior distribution provides a natural representation of the past information and can be updated sequentially using the Bayes rule whenever new data is available. Generalizations of Bayesian inference, such as those that *temper* the likelihood, offer good generalization guarantees (Banerjee, 2006; Audibert, 2009; Gerchinovitz, 2013). Such bounds hold even when the model is misspecified or when an adversary manipulates the stream of data. These generalization bounds are in fact very similar and sometimes even identical to the ones obtained by online learning methods commonly used in the optimization community (Cesa-Bianchi and Lugosi, 2006). The Bayesian principle offers a new perspective which can be used to advance online-learning methods used in areas such as convex optimization, machine learning, reinforcement learning, continual learning, and lifelong learning.

Unfortunately, exact Bayesian inference is computationally challenging in cases where

the normalizing constant of the posterior distribution is a high-dimensional integral. Approximation methods, such as variational inference (VI) (Jordan et al., 1999) and expectation propagation (Minka, 2001), can dramatically reduce the computation cost and enable application of the Bayesian principle to large-scale problems. Despite concerns about their approximation error, these methods have extensively been applied to many machine-learning problems where they show satisfactory performance in practice (Blei and Lafferty, 2006; Hoffman et al., 2013; Kingma and Welling, 2013).

The practical success of such approximation methods points to the gap between the theory and practice. A few recent works have established generalization bounds of the approximation methods such as variational inference, but these are restricted to the batch or offline setting (Alquier and Ridgway, 2017; Bhattacharya et al., 2018; Zhang and Gao, 2017). Extending such results to the online setting, without making strong assumption about the model mismatch and adversaries, is the main focus of this chapter.

We propose online version of variational inference with tempered likelihoods, and derive new generalization bound, which has very similar form to the bound of exact Bayesian inference. Unlike existing proof techniques, our proof extend to the case when approximations are used instead of the exact Bayesian update. Our derivation relies on the convexity of the variational objective. This covers a few important cases, but can be limiting. We argue that the generalization bound is likely to hold more generally, and present empirical evidence in support of these arguments. Our work takes a step towards establishing the generalization properties of online approximate Bayesian methods.

7.1.1 Related works

Variational inference is extremely popular in statistics and machine learning, yet its theoretical properties are not investigated until recently. Generalization bounds for generalized versions of variational approximations are derived in Alquier et al. (2016); Cottet and Alquier (2018). Similarly, Bernstein-von Mises' theorems for variational approximations in parametric models are proved in Wang and Blei (2018), while concentration of the posterior in general models is studied in Alquier and Ridgway (2017); Sheth and Khardon (2017); Bhattacharya et al. (2018); Zhang and Gao (2017); Chérief-Abdellatif and Alquier (2018); Chérief-Abdellatif (2019a); Jaiswal et al. (2019b). These works show that variational approximations does enjoy the same consistency properties as the posterior distribution under general conditions. All of these results however only apply to the batch setting and their extension to the online setting is not straightforward.

It is known that the Bayesian approach leads to good online predictions for a stream of data; see Banerjee (2006), and Cesa-Bianchi and Lugosi (2006); Audibert (2009); Gerchinovitz (2013) for generalized posteriors in machine learning. However, there are only a few attempts to study the online properties of variational inference, and the proofs used in Cesa-Bianchi and Lugosi (2006) cannot easily be extended to online variational inference.

Generalization bounds for online approximations of the posterior are studied in Guhaniyogi et al. (2013), but the algorithms analyzed there are different from the ones used in practice and the feasibility of these algorithms is not proven. Recently Nguyen et al. (2017a)

give some results, but the order of magnitude of the bounds are not explicitly written and in many contexts it is not clear that the bound will even be small enough to ensure consistency. Even though stochastic/online versions of variational inference are known to give good results in practice (Sato, 2001; Hoffman et al., 2010; Wang et al., 2011; Hoffman et al., 2013; Khan and Lin, 2017; Nguyen et al., 2017b; Khan and Nielson, 2018; Osawa et al., 2019; Zeno et al., 2018), existing works have not been able to derive theoretical results confirming their generalization properties. Our results fill this gap between theory and practice for some types of variational approximations obtained with specific types of online algorithms.

7.2 Generalization Properties of Bayesian Inference for Online Learning

Given a stream of data, the goal of online learning is to learn to make good decisions, estimations, or predictions on future data examples. The quality of such decisions is defined with a loss function $\ell(\mathcal{D}_t, \widehat{\theta}_t)$, denoted by $\ell_t(\widehat{\theta}_t)$ for brevity, where $\mathcal{D}_t$ is the data at time t and $\widehat{\theta}_t$ is a quantity computed using the past data, i.e., $\mathcal{D}_{1:(t-1)} := \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{t-1}\}$. This definition of the loss includes popular supervised and unsupervised learning methods. For example, in maximum-likelihood training of a parameterized model p_θ, $\widehat{\theta}_t$ is the parameter estimate and the loss is $\ell_t(\theta) := -\log p_\theta(\mathcal{D}_t)$. Similarly, for a classification task with input-output pair $\mathcal{D}_t := (X_t, Y_t)$, the loss could be the hinge loss $\ell_t(\theta) = (1 - Y_t f_\theta(X_t))_+$ with a classifier f_θ. In the whole paper, we assume that $\theta \mapsto \ell_t(\theta)$ is convex. By using losses ℓ_t until time t, our ultimate goal is to find a θ_t which is as close as possible to the minimizer θ^* of the generalization error $\mathcal{E}_*(\theta) = \mathbb{E}_{\mathcal{D}\sim P_*}[\ell(\mathcal{D}, \theta)]$ where P_* is the true distribution of the data. We would want to do this without many strong assumptions such as assuming the data stream to be i.i.d., or the absence of adversaries.

Since $\mathcal{E}_*$ is unavailable at time t, to ensure the quality of $\widehat{\theta}_t$, online-learning algorithms aim at minimizing the cumulative error $\sum_{i=1}^t \ell_i(\widehat{\theta}_t)$ until time t. Many algorithms are known with bounds on the *regret* of the decision $\widehat{\theta}_t$, that is the gap in the cumulative error and the minimal cumulative error that could have been reached with a *fixed* parameter:

$$\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{T} \ell_t(\theta). \tag{7.1}$$

Bounds on this quantity are known as *regret* bounds, e.g., see Cesa-Bianchi and Lugosi (2006); Bubeck (2011); Shalev-Shwartz (2012); Hazan (2016). Fortunately, bounding the regret also leads to upper bounds on the generalization gap, e.g., by using the average $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^T \widehat{\theta}_t$ we can bound the gap $\mathcal{E}_*(\bar{\theta}_T) - \mathcal{E}_*(\theta^*)$. Due to such properties, regret bounds are useful to study generalization properties of an online algorithm. Moreover, the bound holds with very little assumptions on the data and is valid when the data is not i.i.d. and even when it is corrupted by an adversary.

For online learning, Bayesian inference algorithms have good generalization properties, e.g., the following *tempered* posterior distribution introduced by Vovk (1990); Littlestone

and Warmuth (1994) has a controlled regret:

$$p_t^\eta(\theta) := \frac{1}{\mathcal{Z}_t^\eta}\pi(\theta)e^{-\eta\sum_{i=1}^{t-1}\ell_t(\theta)} \tag{7.2}$$

where $\eta > 0$ is a learning rate, π is a prior distribution, and $\mathcal{Z}_t^\eta$ is the normalizing constant of the posterior distribution. Each loss ℓ_t here can be interpreted as the log-likelihood of a data example $\mathcal{D}_t$. When the loss is indeed equal to $-\log p_\theta(\mathcal{D})$ and $\eta = 1$, the above algorithm is equivalent to Bayesian inference whose generalization properties are usually established under the assumption of no model mismatch (e.g., see Ghosal and Van der Vaart (2017)). The tempered version $\eta < 1$ can be shown to generalize well even when the model is misspecified (Grünwald et al., 2017) or when an adversary manipulates the stream of data. Such tempered versions have also been studied in depth in the machine-learning literature by using the PAC-Bayesian bounds (Shawe-Taylor and Williamson, 1997b; McAllester, 1999; Catoni, 2007; Seldin and Tishby, 2010; Suzuki, 2012; Seldin et al., 2011; Cuong et al., 2013; Germain et al., 2016; Catoni and Giulini, 2017; Guedj, 2019; Tsuzuku et al., 2019).

Algorithm 9 Tempered Bayesian Inference, a.k.a Exponentially Weighted Aggregation

Require: Learning rate $\eta > 0$, prior $\pi(\theta)$, $p_1^\eta \leftarrow \pi$.
for $t = 1, ...,$ **do**
1. $\widehat{\theta}_t \leftarrow \mathbb{E}_{\theta\sim p_t^\eta}(\theta)$,
2. Observe $\mathcal{D}_t$ to suffer a loss $\ell_t(\widehat{\theta}_t)$.
3. Update $p_{t+1}^\eta(\theta) \propto p_t^\eta(\theta)\exp\left[-\eta\ell_t(\theta)\right]$.
end for

In the online-learning literature, the regret bound of this algorithm has been studied extensively under a variety of names, e.g., algorithms such as multiplicative update, weighted majority algorithm, exponentially weighted aggregation (EWA) are all specific cases of tempered Bayesian inference. Algorithm 9 shows a pseudo-code for EWA which performs tempered Bayesian inference in an online fashion (Step 3 implements Equation (7.2)). Below, we state a theorem which shows an example of regret bound[1], proved in Theorem 4.6 in Audibert (2009) for the algorithm shown in Algorithm 9.

Theorem 7.2.1. *Assuming that the loss is bounded, i.e.,* $0 \leq \ell_t(\theta) \leq B, \forall\mathcal{D}_t, \theta$, *the cumulative regret has the following upper bound when* $\widehat{\theta}_t = \mathbb{E}_{\theta\sim p_t^\eta}[\theta]$ *is the posterior mean:*

$$\sum_{t=1}^{T}\ell_t(\widehat{\theta}_t) \leq \inf_{p\in\mathcal{S}}\left\{\mathbb{E}_{\theta\sim p}\left[\sum_{t=1}^{T}\ell_t(\theta)\right] + \frac{\eta B^2 T}{8} + \frac{\mathcal{K}(p,\pi)}{\eta}\right\} \tag{7.3}$$

where $\mathcal{S}$ *is the set of all probability distributions over* Θ *and* $\mathcal{K}$ *is the Küllback-Leibler (KL) divergence.*

[1] In online-learning literature such results are usually stated for finite decision space, e.g., see similar results for EWA in Cesa-Bianchi and Lugosi (2006). The result above holds for a more general continuous setting but under a bounded loss.

A proof is given in Appendix 7.7.5 for the sake of completeness.

The above regret bound is useful to derive explicit bounds in expectation on the generalization error $\mathcal{E}_*$ of an estimator that is defined as the average decision $\bar{\theta}_T := \sum_t \widehat{\theta}_t / T$. For example, we can show that, when a classical prior mass condition[2] on the prior is satisfied and when $\mathcal{D}_t$ are actually independent and identically distributed from P_*, the generalization error has the following bound:

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*} \left[\mathcal{E}_*(\bar{\theta}_T) \right] \leq \mathcal{E}_*(\theta^*) + B \sqrt{\frac{d}{2T} \log\left(\frac{T}{d}\right)} \tag{7.4}$$

for some well-chosen $\eta \sim \sqrt{d/T}$ and $d > 0$ is a complexity measure of the parameter space (often the dimension). This bound shows that when $\mathcal{D}_t$ are i.i.d. from P^* then Bayesian inference achieves generalization error at a rate $\sqrt{d/T}$. An exact statement and a complete proof are given in Theorem 7.7.3 Subsection 7.7.3 in the appendix. The proof is based on a technique called *online-to-batch* analysis. Similar bounds can be derived even for the cases when the model is misspecified and an adversary is present.

The regret bound derived in Theorem 7.2.1 assumes that p_t^η is computed exactly, which is extremely challenging and many a times infeasible. The difficulty arises due to the computation of $\mathcal{Z}_t^\eta$ which is a high-dimensional integral when the space of θ is large. Approximate Bayesian inference methods approximate the integral by finding an approximation of p_t^η in a restricted family of distributions $\mathcal{F} = \{q_\mu, \mu \in \mathcal{M}\}$, e.g., Gaussian distribution with μ being the mean and variance. Our focus in this chapter is to derive bounds similar to Theorem 7.2.1 for approximate Bayesian inference methods.

Unfortunately, deriving similar bounds as Theorem 7.2.1 for approximate inference is not possible using existing proof techniques. This is because these techniques do not work when p_t^η and $\mathcal{S}$ in (7.3) are replaced by q_{μ_t} and $\mathcal{M}$ respectively. As shown in Appendix 7.7.5, these proofs rely on cancellation of many terms in a telescoping sum. This cancellation does not take place when an approximation is used instead, and the error accumulates making the regret bound practically useless. In this chapter, we solve this problem using a different proof for tempered, online variational inference algorithms discussed in the next section.

7.3 Online Variational Inference

In this section, we introduce approximate Bayesian inference methods that can obtain tractable approximations in an online fashion. The methods available in the approximate inference literature are not always suitable for our purpose. Therefore, we present modifications of those methods that lead to feasible online variants of the Bayesian update shown in (7.2). To simplify the notation, we will denote the expectation of the loss under an approximation $q_\mu(\theta)$ by $\bar{L}_t(\mu) := \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)]$.

[2]The exact condition is that the prior $\pi(\theta)$ has mass bigger than ϵ^d on an ϵ-ball around θ^* for some d.

7.3.1 Sequential Variational Approximation

An advantage of variational inference is that it can be directly written as a constrained optimization version of Bayesian inference. To see this we first note that the posterior given in (7.2) can be obtained by solving the following optimization problem (Dai et al., 2016):

$$p_{t+1}^{\eta}(\theta) = \arg\min_{p \in \mathcal{S}} \left\{ \mathbb{E}_{\theta \sim p} \left[\sum_{i=1}^{t} \ell_i(\theta) \right] + \frac{\mathcal{K}(p, \pi)}{\eta} \right\}$$

We can obtain an approximation by simply restricting the set $\mathcal{S}$:

$$q_{\mu_t} := \arg\min_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[\sum_{i=1}^{t-1} \ell_i(\theta) \right] + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\} \tag{7.5}$$

where the set $\mathcal{M}$ is the set of parameters for the set $\mathcal{F} := \{q_\mu, \mu \in \mathcal{M}\}$. The above approximation therefore is a variational approximation of the exact Bayesian inference.

Unfortunately, the update (7.5) may not be feasible in practice. The Bayesian update of (7.2) takes a convenient form where update of p_{t+1}^{η} can be written in terms of p_t^{η}; see line 3 in Algorithm 9. For update (7.5), this is not possible in most cases, i.e., we cannot express the optimization problem for $q_{\mu_{t+1}}$ in terms of q_{μ_t}. Typically, one need to store all the past data examples $\mathcal{D}_i$ and recompute their gradients, and then run the optimizer until it converges. This can be very expensive, especially for large t.

We propose a sequential version which solves these problems by using an approximation. We follow the ideas used in online gradient algorithms, e.g., such as those used in Shalev-Shwartz (2012), and replace $\mathbb{E}_{\theta \sim q_\mu}[\ell_i(\theta)] = \bar{L}_i(\mu) \approx \mu^T \nabla_\mu \bar{L}_i(\mu_i)$. This leads to

$$\mu_{t+1} = \arg\min_{\mu \in \mathcal{M}} \left[\sum_{i=1}^{t} \mu^T \nabla_\mu \bar{L}_i(\mu_i) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right]. \tag{7.6}$$

Note that the gradients in the approximation are computed at the past μ_i, rather than the current one μ_t. This results in an algorithm summarized in Algorithm 10 which we call sequential variational approximation (SVA). When computing the gradient of the KL divergence term is feasible, this algorithm can be cheaply performed.

7.3.2 Streaming Variational Bayes

An alternative approach is to remove the term $\mathcal{K}(q_\mu, \pi)$ since π is already included in q_{μ_t}:

$$\mu_{t+1} = \arg\min_{\mu \in \mathcal{M}} \left[\mu^T \nabla_\mu \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\eta} \right]. \tag{7.7}$$

This step, contained in Algorithm 10, is tractable whenever computing the gradient of the KL term is feasible, e.g., when the expectation parameterization is used. This type of update has been proposed in many recent works, e.g., Nguyen et al. (2017a), Zeno et al. (2018). These updates can be seen as a special case of Broderick et al. (2013). Due to this connection, we call this algorithm streaming variational Bayes (SVB).

Algorithm 10 Online Variational Inference

Require: Learning rate $\eta > 0$, a prior $\pi(\theta) \in \mathcal{F}$, $q_{\mu_1} \leftarrow \pi$.

for $t = 1, ...,$ **do**

1. $\widehat{\theta}_t \leftarrow \mathbb{E}_{\theta \sim q_{\mu_t}}[\theta]$,

2. Observe $\mathcal{D}_t$ to suffer a loss $\ell_t(\widehat{\theta}_t)$.

3. Update depending on the type of algorithm.

a) For SVA, solve (7.6).

b) For NGVI, solve (7.8).

c) For SVB, solve (7.7).

end for

7.3.3 Natural Gradient Variational Inference

The algorithm described in the previous sections are closely related to existing natural-gradient variational inference (NGVI) algorithm (Sato, 2001; Hoffman et al., 2013; Khan and Lin, 2017). These algorithms are typically applied for *stochastic* learning but can be easily modified for online setting. We will consider the method of Khan and Lin (2017) because it applies to the most general setting (other methods require strong *conjugacy* assumptions on the loss $\ell_t(\theta)$ and prior $\pi(\theta)$). The NGVI algorithm is typically applied to obtain exponential-family approximations, but as we will show the updates are similar to our SVA algorithm which also reveals a more general way of implementing these algorithms in the online setting.

The advantage of using NGVI for online learning is that it obtains closed-form updates for $q_{\mu_{t+1}}$ which can be expressed in terms of q_{μ_t}. This is done by exploiting the expectation parameterization[3] of the exponential family. Throughout this section, we denote the expectation parameter by μ and natural parameterization of the exponential family by λ. Khan and Lin (2017) propose the following update[4] in the expectation-parameter space:

$$\min_{\mu \in \mathcal{M}} \left[\mu^T \nabla_\mu \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} + \frac{\mathcal{K}(q_\mu, q_{\mu_t})}{\alpha} \right], \tag{7.8}$$

where $\alpha > 0$ is a step size. The difference from (7.6) is that now the linear term does not contain a sum over all past examples i, rather only the current one. Instead, we add another KL divergence term which contains the past information in the previous approximation q_{μ_t}. Therefore, NGVI algorithm, summarized in Algorithm 10, employs

[3]Expectation parameters are expectations of the sufficient statistics, e.g., Gaussian approximation has two expectation parameters: mean vector and correlation matrix respectively.

[4]The exact update proposed in Khan and Lin (2017) is written differently but can be shown to be equivalent to (7.8). This can be done by using their Lemma 1 and setting $1/\alpha := 1/\beta - 1/\eta$ where β is the step-size used in their paper. We use this form since it makes it easier to establish connections to SVA.

a different way to add the past information, but as we show next, it results in a very similar update as SVA. In the appendix, we provide a closed-form solution to (7.8).

7.3.4 Example: Mean-Field Gaussian VI

We now give a concrete example of the algorithms introduced in this section. We will use the mean-field Gaussian VI where $\mathcal{F}$ is the class of all Gaussian approximations with diagonal covariance matrix. We denote the mean vector of the Gaussian by $m = (m_1, \ldots, m_d)^T$ and the diagonal of the covariance matrix by $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)^T$. To derive the updates for SVA and SVB, we used $\mu = \{m, \sigma\}$ while for NGVI we used the expectation parameters $\mu = \{m, m^2 + \sigma^2\}$. (Here, and until (7.10) below, the squares and multiplications on vectors are to be understood componentwise). We also assume the prior $\pi(\theta)$ to be a Gaussian with mean 0 and variance $s^2 I_d$ where I_d is the identity $d \times d$ matrix.

Denoting the gradients $\bar{g}_{m_t} := \frac{\partial \bar{L}_t}{\partial m}$ and $\bar{g}_{\sigma_t} := \frac{\partial \bar{L}_t}{\partial \sigma}$, we give the update for each method below (here $h(x) := \sqrt{1 + x^2} - x$, applied componentwise for vector inputs):

$$\begin{aligned}
\text{SVA: } & m_{t+1} \leftarrow m_t - \eta s^2 \bar{g}_{m_t}, \qquad g_{t+1} \leftarrow g_t + \bar{g}_{\sigma_t}, \\
& \sigma_{t+1} \leftarrow h\left(\tfrac{1}{2}\eta s g_{t+1}\right) s, \qquad (7.9) \\
\text{SVB: } & m_{t+1} \leftarrow m_t - \eta \sigma_t^2 \bar{g}_{m_t}, \\
& \sigma_{t+1} \leftarrow \sigma_t h\left(\tfrac{1}{2}\eta \sigma_t \bar{g}_{\sigma_t}\right). \qquad (7.10)
\end{aligned}$$

7.4 Generalization Bounds for Online VI

In this section, we present regret bounds for online VI algorithms discussed in the previous section. Our bounds take similar form to the one presented in Theorem 7.2.1, and can be used to obtain generalization bounds similar to (7.4). Our proofs require convexity of $\bar{L}_t(\mu) := \mathbb{E}_{q_\mu}[\ell_t(\theta)]$ with respect to μ, which is a strong assumption. Due to this we are able to derive bounds for SVA and SVB. We expect our bound to hold for NGVI too, due to its similarity to SVA. Specifically, all of our results use the following minimal assumption.

Assumption 7.4.1. *$\bar{L}_t$ is L-Lipschitz and convex.*

Some results require the following stronger assumption.

Assumption 7.4.2. *$\bar{L}_t$ is H-strongly convex where $H > 0$, i.e., for any two $\mu, \mu' \in \mathcal{M}$, the following holds:*

$$\bar{L}_t(\mu') - \bar{L}_t(\mu) \geq (\mu' - \mu)^T \nabla \bar{L}_t(\mu) + \frac{H}{2}\|\mu' - \mu\|^2.$$

Finally, some results also require strong convexity for KL.

Assumption 7.4.3. *The KL divergence $\mu \mapsto \mathcal{K}(q_\mu, q_{\mu_1})$ is α-strongly convex.*

All of these assumption depend heavily on the parametrization of $\{q_\mu, \mu \in M\}$. For some parameterization, these assumptions do hold although such cases are limited. For example, for Gaussian approximations and convex ℓ, the assumptions are satisfied, as pointed out by Challis and Barber (2013). This result has recently been extended by Domke (2019) to more generals *location-scale* family. We give a formal statement below.

Proposition 1 (Theorem 1 in Domke (2019))**.** *Assuming that q_μ belongs to a location-scale family $\mathcal{F} = \{q_{m,C}\}$ where m is a d-length vector and C is a $d \times d$ matrix with $q_{m,C}(\theta) = [\det(C)]^{-1/2}\psi(C^{-1/2}(\theta - m))$ for some fixed density ψ, then $\bar{L}_t$ is convex. Moreover when each $\theta \mapsto \ell_t(\theta)$ is H-strongly convex and ψ is the density of a centered random variable with identity variance matrix, then Assumption 7.4.2 is also satisfied.*

The results for Gaussian approximation can be obtained as a special case.

Proposition 2. *Assume that $\theta \mapsto \ell_t(\theta)$ is L'-Lipschitz. Assume that we use the Gaussian approximation family $\mathcal{F} = \left\{q_{m,C} = \mathcal{N}(m, C^TC), (m, C) \in M\right\}$, $M \subset \mathbb{R}^d \times UT(d)$ where $UT(d)$ is the set of full-rank upper triangular $d \times d$ real matrices. Then $\bar{L}_t$ is L-Lipschitz with $L = \sqrt{2}L'$.*

Finally, we remind the formula for the KL divergence between two Gaussian distributions. Let $q_{m,C} = \mathcal{N}(m, C^TC)$ for any $(m, C) \in \mathbb{R}^d \times UT(d)$. Then

$$\mathcal{K}(q_{m,C}, q_{\bar{m},\bar{C}}) = \frac{1}{2}\left((m-\bar{m})^T\bar{C}^T\bar{C}(m-\bar{m}) + \mathrm{tr}[(\bar{C}^T\bar{C})^{-1}(C^TC)] + \log\left(\frac{\det(\bar{C}^T\bar{C})}{\det(C^TC)}\right) - d\right)$$

is known to be strongly convex on $\mathbb{R}^d \times \mathcal{M}_C$ where $\mathcal{M}_C$ is a closed bounded subset of $UT(d)$. Thus, Assumption 7.4.3 is satisfied with a Gaussian prior and a Gaussian approximation family.

We are now ready to state our regret bounds for SVA and SVB.

7.4.1 Bounds for SVA

Theorem 7.4.1. *Under Assumptions 7.4.1 and 7.4.3, SVA has the following regret bound:*

$$\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) \leq \inf_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu}\left[\sum_{t=1}^{T} \ell_t(\theta)\right] + \frac{\eta L^2 T}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}. \tag{7.11}$$

The above bound is almost identical to the bound given in Theorem 7.2.1 where we can replace p by q_μ, $\mathcal{S}$ by $\mathcal{M}$, the bound B by the Lipschitz constant L, and factor of 8 by the strong convexity parameter α. However, our proof of Theorem 7.4.1 is completely different from the one for Theorem 7.2.1. It relies on arguments from online convex optimization that can be found in Shalev-Shwartz (2012); Hazan (2016). A detailed proof is given in Appendix 7.7.5.

Similar to the Bayesian update case discussed in Section 7.2, using the online-to-batch analysis detailed in Appendix 7.7.3, we can show that the average $\bar{\theta}_T = (1/T)\sum_{t=1}^T \widehat{\theta}_t$ satisfies

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\mu\in\mathcal{M}}\left\{\mathbb{E}_{\theta\sim q_\mu}[\mathcal{E}_*(\theta)] + \frac{\eta L^2}{\alpha} + \frac{\mathcal{K}(q_\mu,\pi)}{\eta T}\right\}. \tag{7.12}$$

As an example consider the mean-field Gaussian approximation and assume that for any $\mathcal{D}$, $\ell(\mathcal{D},\cdot)$ is $L/2$-Lipschitz (note that these are the assumptions of Proposition 2 ensuring that Assumption 7.4.1 is satisfied). Then $\mathbb{E}_{\theta\sim q_\mu}[\mathcal{E}_*(\theta)] = \mathcal{E}_*(m) + \|\sigma\|L/2$. Therefore, given the expression of the KL-divergence between Gaussian distributions, taking a vector σ with $\sigma_j = L\eta/(\alpha\sqrt{d})$, $\eta = (1/L)\sqrt{\alpha d\log(T/d)/T}$, and considering only the regret with respect to bounded means m leads to

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{m\in[-\bar{M},\bar{M}]^d}\mathcal{E}_*(m) + (1+o(1))\frac{2L}{\alpha}\sqrt{\frac{d\log(dT)}{T}}.$$

This again is very similar to the generalization error shown in (7.4).

7.4.2 Bounds for SVB

Similarly to the SVA case, we can derive a regret bound, however our proof only applies to the Gaussian case. For this case, we require a dynamic learning η_t. We use a different learning rate for each element of θ_j which we denote by $\eta_{t,j}$. The result also works for a bounded parameter space $\mathcal{M} = \mathcal{M}_m \times \mathcal{M}_\sigma$ that will imply a projection step in addition to the update in (7.10):

$$\begin{aligned}\text{SVB: } m_{t+1} &\leftarrow \Pi_{\mathcal{M}_m}\left[m_t - \eta\sigma_t^2\bar{g}_{m_t}\right],\\ \sigma_{t+1} &\leftarrow \Pi_{\mathcal{M}_\sigma}\left[\sigma_t h\left(\tfrac{1}{2}\eta\sigma_t\bar{g}_{\sigma_t}\right)\right].\end{aligned}$$

where $\Pi_{\mathcal{M}_m}$ and $\Pi_{\mathcal{M}_\sigma}$ denote the orthogonal projection on $\mathcal{M}_m$ and $\mathcal{M}_\sigma$ respectively. The following theorem states the result.

Theorem 7.4.2. *We consider the mean-field Gaussian family $q_\mu = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ and $\mathcal{M} = \mathcal{M}_m \times \mathcal{M}_\sigma$ where $\mathcal{M}_m$ and $\mathcal{M}_\sigma$ are closed, bounded, convex subsets of $\mathbb{R}^d$ and $\mathbb{R}^d_+$ respectively, and $0 \in \mathcal{M}_\sigma$. Define $D^2 = \sup\{\|m-m'\|_2^2 + \|\sigma\|^2, m, m' \in \mathcal{M}_m, \sigma\in\mathcal{M}_\sigma\}$. Then, under Assumption 7.4.1, with the choice $\eta_{t,j} = \frac{D\sqrt{2}}{L}\frac{1}{\sqrt{t\sigma_{t,j}^2}}$ we get:*

$$\sum_{t=1}^T \ell_t(\widehat{\theta}_t) \leq \inf_{\theta\in\mathcal{M}_m}\sum_{t=1}^T \ell_t(\theta) + DL\sqrt{2T}. \tag{7.13}$$

Under Assumptions 7.4.1 and 7.4.2, the choice $\eta_t = 2/Ht\sigma_t^2$ leads to:

$$\sum_{t=1}^T \ell_t(\widehat{\theta}_t) \leq \inf_{\theta\in\mathcal{M}_m}\sum_{t=1}^T \ell_t(\theta) + \frac{L^2(1+\log T)}{H}. \tag{7.14}$$

Here again the results are similar to the Bayesian inference case but now expressed in terms of the parameters μ instead of expectations.

A similar bound on the generalization error can also be proved. Define $\bar{\theta}_T = (1/T)\sum_{t=1}^{T}\widehat{\theta}_t$. Here, the online-to-batch analysis directly leads to

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\theta\in\mathcal{M}_m} \mathcal{E}_*(\theta) + \frac{DL\sqrt{2}}{\sqrt{T}}$$

in the convex case and

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\theta\in\mathcal{M}_m} \mathcal{E}_*(\theta) + \frac{L^2(1+\log T)}{HT}$$

in the strongly convex case.

Note the in the online optimization setting studied in Shalev-Shwartz (2012), it is usual to optimize on Euclidean balls. Here, $M_m = \{m \in \mathbb{R}^d : \|m\| \leq \bar{M}\}$ and $M_\sigma = \{\sigma \in \mathbb{R}^d_+ : \|\sigma\| \leq \bar{S}\}$ leads to $D = 4\bar{M}^2 + \bar{S}^2$ leads to dimension-free bounds.

On the other hand, the choice $M_m = [-\bar{M}, \bar{M}]^d$ and $M_\sigma = [0, \bar{S}]^d$ implies $D^2 = d(4\bar{M}^2 + \bar{S}^2)$, and so the bound in the convex case is

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \inf_{\theta\in\mathcal{M}_m} \mathcal{E}_*(\theta) + \frac{L\sqrt{2d(4\bar{M}^2 + \bar{S}^2)}}{\sqrt{T}}$$

and its dependence in d is the same as in the bound on SVA.

7.4.3 Generalization

We expect our bounds to hold for NGVI as well. When expectation parameterization is used, the assumptions are satisfied only in very limited models. This is because the results of Propositions 1 and 2 do not directly apply to expectation parameterization. However, the NGVI update shown in (7.8) can be applied in other parameterization as well, in which case some of our result can be extended to NGVI too.

7.5 Experiments

In this section, we conduct experiments on real and simulated datasets, in classification and linear/nonlinear regression. The objective is twofold: check the convergence of SVA/SVB, with and without the convexity assumption on $\bar{L}_t$, and compare SVA, NGVI and SVB.

7.5.1 Experimental setup

We compare the empirical performance of the algorithms we present in this chapter through classification and regression tasks on several toy and real-world datasets. We

also include the classical online gradient descent and the online gradient descent on the expected loss as benchmarks. Please refer to Appendix 7.7.2 for more details on these algorithms. In the following, OGA will stand for the classical online gradient descent while OGA-EL for the OGA on the expected loss (Algorithm 10). We recall that SVA, NGVI and SVB respectively refer to the sequential variational approximation (7.6), natural gradient variational inference (7.8) and streaming variational Bayes (7.7).

Binary classification We consider first a classification problem. At each round t the learner receives a data point $x_t \in \mathbb{R}^d$ and predicts its label $y_t \in \{-1, +1\}$ using $\langle x_t, \theta_t \rangle$. The adversary reveals the true value y_t, then the learner suffers the loss $\ell_t(\theta_t) = (1 - y_t \theta_t^T x_t)_+$, where $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise.

Regression At each round t, the learner receives a set of features $x_t \in \mathbb{R}^d$ and predicts $y_t \in \mathbb{R}$ using $\langle x_t, \theta_t \rangle$. Then the adversary reveals the true value y_t and the learner suffers the loss $\ell_t(\theta_t) = (y_t - f_{\theta_t}(x_t))^2$. We will consider both the linear case when the predictions are linear $f_\theta(x_t) = \theta^T x_t$ and the nonlinear case where the predictions are outputs of a one-hidden-layer neural network with a ReLU activation. The first case of linear predictions leads to a convex loss with respect to θ, while the latter leads to a nonconvex loss.

Variational family For both tasks, we use a Gaussian mean-field variational family $\mathcal{F} = \{q_\mu = \mathcal{N}(m, \mathrm{diag}(\sigma^2)) / \mu = (m, \sigma) \in M_m \times M_\sigma\}$, $M_m = [-20, 20]^d$ and $M_\sigma = [0, 1]^d$.

Datasets We consider here six different datasets: one toy and three real datasets for classification, and one real world dataset for both linear and nonlinear regression. The three real world datasets used for the binary classification problem are the popular Breast Cancer, the Pima Indians and the Forest Cover Type datasets, while those used for regression are the Boston Housing and the California Housing datasets respectively for the convex and the nonconvex case. All come from the UCI machine learning repository. Note that in some databases, the data are ordered according to some criterion such as the date or the label. In order to avoid any effect linked to this, we randomly permuted the observations.

The toy dataset is as follows: we sample $n = 10^4$ points y_t according to a Bernoulli distribution $\mathcal{B}e(2/3)$. Then

$$x_t|(y_t = 1) \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}\right) \text{ and } x_t|(y_t = 0) \sim \mathcal{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

Dataset	T	d	Dataset	T	d
Toy classification	10000	2	Cover Type	581012	54
Breast cancer	569	30	Boston Housing	506	13
Pima Indians	768	8	California Housing	20640	9

7.5.2 Experimental results

For each task and each dataset, we plot the evolution of the average cumulative loss $\sum_{i=1}^t \ell_i(\theta_i)/t$ as a function of the step $t = 1, ..., T$, where T is the number of instances of the dataset and θ_i is the decision made by the learner at step i. We compare this

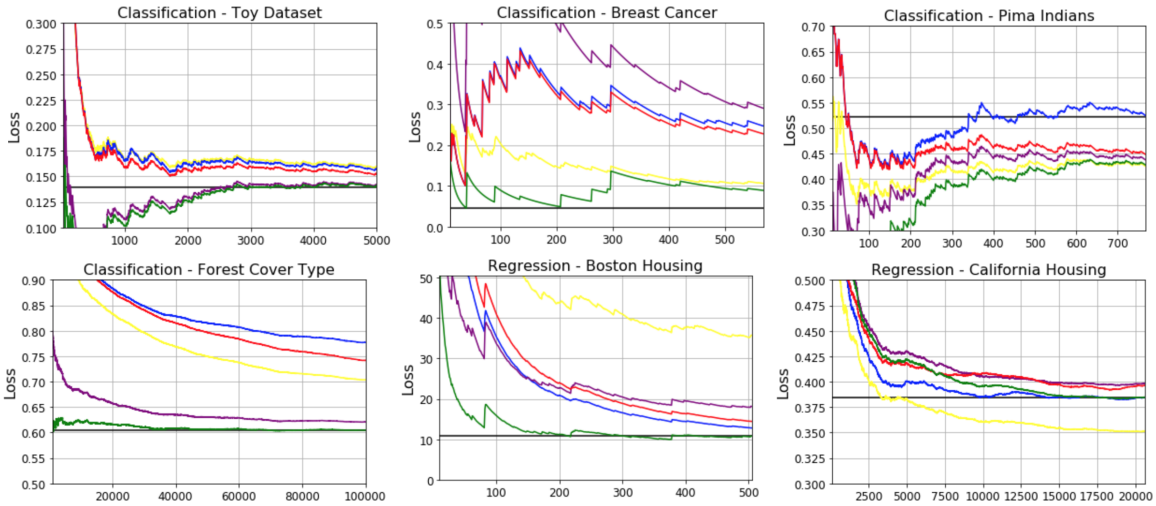


Figure 7.1: Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green) for the convex hinge loss and the squared loss functions. The black line shows the average total cumulative loss in hindsight. We see that in most cases NGVI outperforms the other algorithms. The last plot (California Housing dataset) shows the consistency of our algorithms for a nonconvex loss $\bar{L}_t$.

quantity to the best average total cumulative loss in hindsight $\inf_{\theta \in M_m} \frac{1}{T} \sum_{t=1}^{T} \ell_t(\theta)$ which is represented by a straight black horizontal line in Figure 7.1.

Parameters setting We initialize all means to 0 and all values of the variance to 1. For simplicity, the values of the learning rates are set to $\eta = 1/\sqrt{T}$ for OGA, OGA-EL and SVA while $\eta_t = 1/\sigma_t^2\sqrt{t}$ for SVB and $\eta_t = 1$ for NGVI respectively. It is possible to optimize the values of the step sizes. Nevertheless, we draw attention to the fact that a simple cross validation technique would not be valid here as it would require to know the whole dataset before selecting the step size, which is not possible in an online setting, and using such a strategy at each step t using the past data would change the learning rate of OGA, OGA-EL and SVA at each step.

Conclusions The results are reported in Figure 7.1 that shows the consistency of our algorithms. The goal of our simulations is to observe the empirical performance of our algorithms in practice, and to see if it is possible to go further than the convexity assumption that is required in Section 7.4. Looking at the plots, the two main findings of our experiments are the following:

- the generalization properties of online variational inference seem to go beyond the convex assumption we stated in the previous theoretical parts.
- even though SVA and SVB exhibit good performances, NGVI is the best method in practice as it converges faster on all the datasets.

7.6 Conclusion

In this chapter, we derive the first generalization bounds for some online variational inference algorithms. Our proof techniques applies to cases where existing methods do

not work. By using existing variational methods, we proposed a few online methods for variational inference. We provided generalization bounds for the SVA algorithm, and related them to the NGVI methods. We also derived a bound for a special case of SVB. We provided numerical results to establish consistency of our results. We observed that NGVI outperforms all the other methods, and that the theoretical convexity assumption is not needed in practice.

We believe that it is possible to extend our proof techniques to NGVI case. Currently, our proofs strongly rely on the convexity of $\mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$ with respect to μ. This analysis cannot directly be used for the parameterization of Khan and Lin (2017). However, it can be applied to a general formulation where our assumptions hold. We believe that generalization bounds for NGVI is possible to derive and will pursue this direction in the future.

7.7 Proofs and additional results

7.7.1 Closed-form solutions for NGVI

The expectation parameterization of NGVI enables closed-form solution. This is because the gradient of the KL diverence with respect to expectation parameter is available in closed-form (see Eq. 10 in Khan and Nielson (2018)). The closed update is given in Eq. (50) in Khan and Lin (2017) using which we obtain the following update:

$$\lambda_{t+1} = (1-\beta)\lambda_t + \beta\lambda_1 - \eta\beta\nabla_\mu \bar{L}_t(\mu_t), \tag{7.15}$$

where $1/\beta := 1/\alpha + 1/\eta$. Given λ_{t+1}, we can get $\mu_{t+1} = \nabla_\lambda A(\lambda_{t+1})$ where A is the log-partition function of the exponential family.

Now we show that this closed-form update is similar to SVA. By using induction similar to Lemma 4 in Khan and Lin (2017), we can write the update in terms of all past gradients:

$$\lambda_{t+1} = \lambda_1 - \eta\sum_{i=1}^{t} w_i \nabla_\mu \bar{L}_i(\mu_i) \tag{7.16}$$

where $w_i := \beta\prod(1-\beta)^{i-2}$. This can be compared to the SVA update in the expectation parameterization where applying the gradient to (7.6) gives us the following update similar to (7.15) but where $w_i = 1$ for all i:

$$\lambda_{t+1} = \lambda_1 - \eta\sum_{i=1}^{t} \nabla_\mu \bar{L}_i(\mu_i) \tag{7.17}$$

Therefore, SVA takes a gradient step assuming that all gradients are equally important, which is similar to the Bayesian update (7.2) where all loss ℓ_i are treated equally. In contrast, in NGVI, the past gradients are discounted using β and ultimately forgotten. Weighting past gradients makes sense when we do not want the current mistakes to affect the future. However, the choice of step-size is crucial to know the rate at which the past gradients should be discounted.

NGVI is typically applied using expectation parameterization, but the formulation (7.8) is more general although could be computationally difficult. The theoretical results in the paper further assume that $\bar{L}_i$ is convex in μ. Still, in our experiments, NGVI gives good performance in an online setting compared to many other algorithms.

7.7.2 Online gradient algorithm on the expected loss (OGA-EL)

It is possible to directly use the online gradient algorithm (OGA) on the expected loss $\mathbb{E}_{\theta\sim q_\mu}[\ell_t(\theta)]$, see Algorithm 11.

Algorithm 11 OGA-EL

Require: Learning rate $\eta > 0$, a prior $\pi(\theta) \in \mathcal{F}$, $q_{\mu_1} \leftarrow \pi$.
for $t = 1, ...,$ **do**
1. $\widehat{\theta}_t \leftarrow \mathbb{E}_{\theta\sim q_{\mu_t}}[\theta]$,
2. Observe $\mathcal{D}_t$ to suffer a loss $\ell_t(\widehat{\theta}_t)$.
3. Update $\mu_{t+1} = \mu_t - \eta\nabla\bar{L}_t(\mu_t)$.
end for

Note first that from Shalev-Shwartz (2012) step (iii) is actually equivalent to

$$\mu_{t+1} = \arg\min_{\mu\in M}\left[\sum_{i=1}^{t}\mu^T\nabla\bar{L}_i(\mu_i) + \frac{\|\mu-\mu_1\|^2}{\eta}\right],$$

which means that we replaced the Küllback-Leibler divergence by the Euclidean norm in SVA.

Also, when $\mu = (m, \sigma) \in \mathbb{R}^d \times (\mathbb{R}_+)^d$ and $q_\mu = \mathcal{N}(m, \mathrm{diag}(\sigma))$, then Algorithm 11 becomes

$$m_{t+1} = m_t - \eta s^2\frac{\partial\bar{L}_t}{\partial m}(m_t, \sigma_t),$$
$$\sigma_{t+1} = \sigma_t - \eta s^2\frac{\partial\bar{L}_t}{\partial\sigma}(m_t, \sigma_t).$$

We have regret bounds for this method, similar to the one for EWA:

Theorem 7.7.1. *Under Assumption 7.4.1, Algorithm 11 leads to:*

$$\sum_{t=1}^{T}\ell_t(\widehat{\theta}_t) \le \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}\ell_t(\theta)\right] + \eta L^2T + \frac{\|\mu-\mu_1\|^2}{\eta}\right\},$$

and moreover, under Assumptions 7.4.3 and 7.4.1, Algorithm 11 leads to:

$$\sum_{t=1}^{T}\ell_t(\widehat{\theta}_t) \le \inf_{\mu\in M}\left\{\mathbb{E}_{\theta\sim q_\mu}\left[\sum_{t=1}^{T}\ell_t(\theta)\right] + \eta L^2T + \frac{\alpha\mathcal{K}(q_\mu, \pi)}{2\eta}\right\}.$$

The proof of this result is given below with the other proofs of the paper.

7.7.3 Online-to-batch conversion

Many times in the paper, we derived generalization error bounds from regret bounds, using the online-to-batch conversion. We here give a formal statement for this result, note that this result is essentially Theorem 5.1 in Shalev-Shwartz (2012). We also provide a proof for the sake of completeness.

Theorem 7.7.2. *Assume that $\mathcal{D}_1, \ldots, \mathcal{D}_T$ are i.i.d from P_*. Assume we use an online algorithm on the data that produce a sequence of parameters $\widehat{\theta}_1, \ldots, \widehat{\theta}_T$. That is, $\widehat{\theta}_t = \widehat{\theta}(\mathcal{D}_1, \ldots, \mathcal{D}_{t-1})$. Define the estimator*

$$\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \widehat{\theta}_t.$$

Then

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\bar{\theta}_T)] \leq \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t)\right].$$

Proof. We have:

$$\mathcal{E}_*(\bar{\theta}_T) = \mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell(\mathcal{D}, \bar{\theta}_T)\right] = \mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell\left(\mathcal{D}, \frac{1}{T}\sum_{t=1}^{T}\widehat{\theta}_t\right)\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell\left(\mathcal{D}, \widehat{\theta}_t\right)\right]$$

by Jensen's inequality. The key is that as $\widehat{\theta}_t = \widehat{\theta}_t(\mathcal{D}_1, \ldots, \mathcal{D}_{t-1})$ does not depend on $\mathcal{D}_t$, we can rewrite:

$$\mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell\left(\mathcal{D}, \widehat{\theta}_t\right)\right] = \mathbb{E}_{\mathcal{D}_t\sim P_*}\left[\ell\left(\mathcal{D}_t, \widehat{\theta}_t\right)\right] = \mathbb{E}_{\mathcal{D}_t\sim P_*}\left[\ell_t(\widehat{\theta}_t)\right]$$

and so we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\mathcal{E}_*(\bar{\theta}_T)\right] &\leq \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}\sim P_*}\left[\ell_t(\widehat{\theta}_t)\right]\right\} \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\ell_t(\widehat{\theta}_t)\right] = \mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t(\widehat{\theta}_t)\right].\end{aligned}$$

□

As an application, we state an exact version of (7.4) and prove it from Theorem 7.2.1 and Theorem 7.7.2.

Theorem 7.7.3. *Assume that the loss ℓ is bounded by B as in Theorem 7.2.1 and that $\mathcal{D}_1, \ldots, \mathcal{D}_T$ are i.i.d from P_*. Assume that there is some $d > 0$ such that*

$$r(\varepsilon) \leq d\log(1/\varepsilon)$$

where $r(\varepsilon) = \log[1/\pi(B(\theta^*, \varepsilon))]$ and $B(\theta^*, \varepsilon) = \{\theta \in \Theta : \mathcal{E}(\theta) - \mathcal{E}(\theta^*) \leq \varepsilon\}$. Use on this data the EWA strategy with $\eta = (1/2\sqrt{2}B)\sqrt{(d/T)\log(d/T)}$, then

$$\mathbb{E}_{\mathcal{D}_{1:T}\sim P_*}[\mathcal{E}_*(\widehat{\theta}_T)] \leq \mathcal{E}_*(\theta^*) + B\sqrt{\frac{d}{2T}\log\left(\frac{T}{d}\right)} + \frac{d}{T}.$$

Note that the prior mass condition is classical in the PAC-Bayesian literature and in the frequentist analysis of Bayesian estimators, see e.g Catoni (2007); Rousseau (2016); Bhattacharya et al. (2016); Ghosal and Van der Vaart (2017). The estimator $\bar{\theta}_T$ averaging the decisions $\widehat{\theta}_t$ was first introduced by Catoni (2004) as the "double mixture rule".

Proof. Define p_ε as π restricted to $B(\theta^*, \varepsilon)$ and note that

$$\mathcal{K}(p_\varepsilon, \pi) = -\log \pi\left(B(\theta^*, \varepsilon)\right) = r(\varepsilon) \leq d \log(1/\varepsilon).$$

From Theorem 7.2.1, for any ε,

$$\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) \leq \mathbb{E}_{\theta \sim p_\varepsilon}\left[\sum_{t=1}^{T} \ell_t(\theta)\right] + \frac{\eta B^2 T}{8} + \frac{d \log(1/\varepsilon)}{\eta}.$$

From Theorem 7.7.2,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\widehat{\theta}_T)] &= \mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t)\right] \\
&\leq \mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}\left\{\mathbb{E}_{\theta \sim p_\varepsilon}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(\theta)\right]\right\} + \frac{\eta B^2}{8} + \frac{d \log(1/\varepsilon)}{T\eta} \\
&= \mathbb{E}_{\theta \sim p_\varepsilon}\left[\mathcal{E}_*(\theta)\right] + \frac{\eta B^2}{8} + \frac{d \log(1/\varepsilon)}{T\eta} \\
&\leq \mathcal{E}_*(\theta^*) + \varepsilon + \frac{\eta B^2}{8} + \frac{d \log(1/\varepsilon)}{T\eta}
\end{aligned}$$

where the last inequality comes from the definition of p_ε. Taking $\varepsilon = d/T$ gives:

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\widehat{\theta}_T)] \leq \mathcal{E}_*(\theta^*) + \frac{d}{T} + \frac{\eta B^2}{8} + \frac{d \log(T/d)}{T\eta}.$$

Finally, substitute its value to η to get

$$\mathbb{E}_{\mathcal{D}_{1:T} \sim P_*}[\mathcal{E}_*(\widehat{\theta}_T)] \leq \mathcal{E}_*(\theta^*) + B\sqrt{\frac{d}{2T}\log\left(\frac{T}{d}\right)} + \frac{d}{T}.$$

□

7.7.4 A tool for the proofs

We remind the following classical lemma. We refer the reader for example to Catoni (2007) for a proof of this result, where it is stated as Lemma 1.1.3 (page 16).

Lemma 7.7.4. *Let $h : \Theta \to \mathbb{R}$ be a bounded measurable function and $\pi \in \mathcal{S}(\Theta)$. Then*

$$\sup_{p \in \mathcal{S}(\Theta)} \left\{\mathbb{E}_{\theta \sim p}[h(\theta)] - \mathcal{K}(p, \pi)\right\} = \log \mathbb{E}_{\theta \sim \pi}[\exp(h(\theta))]$$

and the supremum is actually reached for

$$p(\theta) \propto \exp[h(\theta)]\pi(\theta).$$

This lemma will actually turn out to be a fundamental tool for some of the proofs.

7.7.5 Proofs

Proof of Theorem 7.2.1. Note that this proof is classical and is reminded here for the sake of completeness. We have first:

$$\begin{aligned}\exp\left[-\eta\ell_t(\widehat{\theta}_t)\right] &= \exp\left[-\eta\ell_t(\mathbb{E}_{\theta\sim p_t^\eta}(\theta))\right] \\ &\geq \exp\left[-\eta\mathbb{E}_{\theta\sim p_t^\eta}(\ell_t(\theta))\right] \\ &\geq \mathbb{E}_{\theta\sim p_t^\eta}\left\{\exp\left[-\eta\ell_t(\theta) - \frac{\eta^2 B^2}{8}\right]\right\}\end{aligned}$$

where we used respectively Jensen and Hoeffding's inequality. So

$$\ell_t(\widehat{\theta}_t) \leq \frac{\eta B^2}{8} - \frac{1}{\eta}\log \mathbb{E}_{\theta\sim p_t^\eta}\exp\left[-\eta\ell_t(\theta)\right]. \tag{7.18}$$

Remind that by definition,

$$p_t^\eta(\theta) = \frac{\exp\left(-\eta\sum_{i=1}^{t-1}\ell_i(\theta)\right)\pi(\theta)}{N_t}$$

where N_t is the normalisation constant given by

$$N_t = \mathbb{E}_{\theta\sim\pi}\left[\exp\left(-\eta\sum_{i=1}^{t-1}\ell_i(\theta)\right)\right].$$

But note that then

$$\log \mathbb{E}_{\theta\sim p_t^\eta}\exp\left[-\eta\ell_t(\theta)\right] = \log\left(\frac{N_{t+1}}{N_t}\right).$$

We plug this into (7.18) and sum for $t = 1, \ldots, T$. We obtain

$$\begin{aligned}\sum_{t=1}^T \ell_t(\widehat{\theta}_t) &\leq \frac{\eta B^2 T}{8} - \frac{1}{\eta}\sum_{t=1}^T \log\left(\frac{N_{t+1}}{N_t}\right) \\ &= \frac{\eta B^2 T}{8} - \frac{1}{\eta}\log\left(\frac{N_{T+1}}{N_1}\right) \\ &= \frac{\eta B^2 T}{8} - \frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\pi}\left[\exp\left(-\eta\sum_{t=1}^T\ell_t(\theta)\right)\right]\right).\end{aligned}$$

Lemma 7.7.4 leads to

$$\sum_{t=1}^T \ell_t(\widehat{\theta}_t) \leq \frac{\eta B^2 T}{8} + \inf_{p\in\mathcal{S}(\Theta)}\left\{\mathbb{E}_{\theta\sim p}\left[\sum_{t=1}^T\ell_t(\theta)\right] + \frac{\mathcal{K}(p,\pi)}{\eta}\right\}.$$

□

Proof of Proposition 2. Let $\varphi_{m,C}(\cdot)$ denote the p.d.f of the Gaussian distribution with mean m and variance matrix C. Let $(m_1, C_1), (m_2, C_2) \in M$,

$$|\bar{L}_t(m_1, C_1) - \bar{L}_t(m_2, C_2)| = \left|\int \ell_t(\theta)\varphi_{m_1,C_1}(\theta)\mathrm{d}\theta - \int \ell_t(\theta)\varphi_{m_2,C_2}(\theta)\mathrm{d}\theta\right|$$

$$\leq \int |\ell_t(m_1 + C_1 u) - \ell_t(m_2 + C_2 u)| \, \varphi_{0,I_d}(u) \mathrm{d}u$$
$$\leq L' \|m_1 - m_2\| + L' \int \|(C_1 - C_2)u\| \varphi_{0,I_d}(u) \mathrm{d}u.$$

For any $C = (C_{i,j}) \in UT(d)$, we have

$$\int \|Cu\| \varphi_{0,I_d}(u) \mathrm{d}u \leq \sqrt{\int \|Cu\|^2 \varphi_{0,I_d}(u) \mathrm{d}u}$$
$$= \sqrt{\int \sum_{i=1}^{d} \left(\sum_{j=1}^{d} C_{i,j} u_j \right)^2 \varphi_{0,I_d}(u) \mathrm{d}u} = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{d} C_{i,j}^2}$$

which leads to

$$|\bar{L}_t(m_1, C_1) - \bar{L}_t(m_2, C_2)| \leq L' \|m_1 - m_2\| + L' \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{d} (C_1 - C_2)_{i,j}^2}$$
$$\leq \sqrt{2} L' \|(m_1, C_1) - (m_2, C_2)\|.$$

This ends the proof. □

Proof of Theorem 7.4.1. First, Assumption 7.4.1 ensures that the $\bar{L}_t$'s are convex. By definition of the subgradient of a convex function,

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) - \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] &= \sum_{t=1}^{T} \ell_t \left(\mathbb{E}_{\theta \sim q_{\mu_t}}(\theta) \right) - \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] - \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \\
&= \sum_{t=1}^{T} \bar{L}_t(\mu_t) - \sum_{t=1}^{t} \bar{L}_t(\mu) \\
&\leq \sum_{t=1}^{T} \mu_t^T \nabla \bar{L}_t(\mu_t) - \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t). \qquad (7.19)
\end{aligned}$$

Then, following the general proof scheme detailed in Chapter 2 in Shalev-Shwartz (2012), we prove by recursion on T that for any $\mu \in \mathcal{M}$,

$$\sum_{t=1}^{T} \mu_t^T \nabla \bar{L}_t(\mu_t) - \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) \leq \sum_{t=1}^{T} \mu_t^T \nabla \bar{L}_t(\mu_t) - \sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \qquad (7.20)$$

which is exactly equivalent to

$$\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta}. \qquad (7.21)$$

Indeed, for $T = 0$, (7.21) just states that $\mathcal{K}(q_\mu, \pi) \geq 0$ which is a well-known property of KL. Assume that (7.21) holds for some integer $T - 1$. We then have, for all $\mu \in M$,

$$\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) = \sum_{t=1}^{T-1} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) + \mu_{T+1}^T \nabla \bar{L}_T(\mu_T)$$

$$\leq \sum_{t=1}^{T-1} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} + \mu_{T+1}^T \nabla \bar{L}_T(\mu_T)$$

as (7.21) holds for $T-1$. Apply this to $\mu = \mu_{T+1}$ to get

$$\begin{aligned}
\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &\leq \sum_{t=1}^{T} \mu_{T+1}^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(p_{\mu_{T+1}}, \pi)}{\eta} \\
&= \min_{m \in \mathcal{M}} \left[\sum_{t=1}^{T} m^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(p_m, \pi)}{\eta} \right], \text{ by definition of } \mu_{T+1} \\
&\leq \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta}
\end{aligned}$$

for all $\mu \in \mathcal{M}$. Thus, (7.21) holds for T. Thus, by recursion, (7.21) and (7.20) hold for all $T \in \mathbb{N}$.

The last step is to prove that for any $t \in \mathbb{N}$,

$$\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \frac{\eta L^2}{\alpha}. \tag{7.22}$$

Indeed,

$$\begin{aligned}
\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &= (\mu_t - \mu_{t+1})^T \nabla \bar{L}_t(\mu_t) \\
&\leq \|\mu_t - \mu_{t+1}\| \|\nabla \bar{L}_t(\mu_t)\| \text{ by Cauchy-Schwarz} \\
&\leq L \|\mu_t - \mu_{t+1}\| \qquad (7.23)
\end{aligned}$$

as $\bar{L}_t$ is L Lipschitz (Assumption 7.4.1). Define

$$G_t(\mu) = \sum_{i=1}^{t-1} \mu^T \nabla \bar{L}_i(\mu_i) + \frac{\mathcal{K}(q_\mu, \pi)}{\eta}.$$

Note that from Assumption 7.4.3, $\mu \mapsto \mathcal{K}(q_\mu, \pi)/\eta$ is α/η-strongly convex. As the sum of a linear function and an α/η-strongly convex function, G_t is α/η-strongly convex. So, for any (μ, μ'),

$$G_t(\mu') - G_t(\mu) \geq (\mu' - \mu)^T \nabla G_t(\mu) + \frac{\alpha \|\mu' - \mu\|^2}{2\eta}.$$

As a special case, using the fact that μ_t is a minimizer of G_t, we have

$$G_t(\mu_{t+1}) - G_t(\mu_t) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

In the same way,

$$G_{t+1}(\mu_t) - G_{t+1}(\mu_{t+1}) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Summing the two previous inequalities gives

$$\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \geq \frac{\alpha \|\mu_{t+1} - \mu_t\|^2}{\eta},$$

and so, combined with, this gives:

$$\|\mu_{t+1} - \mu_t\| \leq \sqrt{\frac{\eta}{\alpha}\left[\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t)\right]}.$$

Combining this inequality with (7.23) leads to (7.22).

Plugging (7.19), (7.20) and (7.22) together gives

$$\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) - \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \leq \frac{\eta T L^2}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta},$$

that is the statement of the theorem. □

Proof of Theorem 7.4.2. We prove this theorem from scratch and use the main techniques outlined in Hazan (2016). As previously, the idea is to study differences $\bar{L}_t(\mu_t) - \bar{L}_t(\mu)$. However, in this case, we have, for any $\mu = (m, \sigma)$, using Jensen's inequality,

$$\bar{L}_t(m, \sigma) = \mathbb{E}_{\theta \sim q_{m,\sigma}}[\ell_t(\theta)] \geq \ell_t(m) = \bar{L}_t(m, 0).$$

So, we can assume from the beginning that $\mu = (m, 0)$.

Convex case:

First, we assume that each function $\bar{L}_t$ is convex, for all $m = (m_1, ..., m_d) \in \mathcal{M}_m$ and $\mu = (m, 0)$:

$$\bar{L}_t(\mu_t) - \bar{L}_t(\mu) \leq \nabla \bar{L}_t(\mu_t)^T(\mu_t - \mu) = \sum_{j=1}^{d} \left[\frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j) + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j}\right].$$

Using the update formulas 7.10:

$$(m_{t+1,j} - m_j)^2 = (m_{t,j} - m_j)^2 + \eta_{t,j}^2 \sigma_{t,j}^4 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2 - 2\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j)$$

and

$$\sigma_{t+1,j}^2 = \sigma_{t,j}^2 + \frac{\eta_{t,j}^2 \sigma_{t,j}^4}{2} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_{t,j}, \sigma_{t,j})^2 - \eta_{t,j}\sigma_{t,j}^2 \sqrt{1 + \left(\frac{\eta_{t,j}\sigma_{t,j} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2}\right)^2} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j}.$$

Rearranging the terms, we get:

$$\frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)(m_{t,j} - m_j) = \frac{(m_{t,j} - m_j)^2 - (m_{t+1,j} - m_j)^2}{2\eta_{t,j}\sigma_{t,j}^2} + \frac{\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2}{2}$$

and

$$\frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)\sigma_{t,j} = \frac{\sigma_{t,j}^2 - \sigma_{t+1,j}^2}{\eta_{t,j}\sigma_{t,j}^2 \sqrt{1 + \left(\frac{\eta_{t,j}\sigma_{t,j} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2}\right)^2}} + \frac{\eta_{t,j}\sigma_{t,j}^2 \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)^2}{2\sqrt{1 + \left(\frac{\eta_{t,j}\sigma_{t,j} \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)}{2}\right)^2}}.$$

We also use the boundedness of the gradients: for any $(m,\sigma)\in\mathcal{M}$, at any date t,

$$\sum_{j=1}^{d}\left[\frac{\partial\bar{L}_t}{\partial m_j}(m,\sigma)^2+\frac{\partial\bar{L}_t}{\partial\sigma_j}(m,\sigma)^2\right]\leq L^2.$$

We upper bound the inverse of the square root by 1, the gradient by L and we sum over time:

$$\begin{aligned}
\sum_{t=1}^{T}\bar{L}_t(\mu_t)-\bar{L}_t(\mu)&\leq\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{(m_{t,j}-m_j)^2}{2}\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2}-\frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right]\\
&+\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2\\
&+\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\sigma_{t,j}^2}{2}\left[\frac{2}{\eta_{t,j}\sigma_{t,j}^2}-\frac{2}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right]\\
&+\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2\\
&=\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{(m_{t,j}-m_j)^2}{2}\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2}-\frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right]\\
&+\sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\sigma_{t,j}^2}{2}\left[\frac{2}{\eta_{t,j}\sigma_{t,j}^2}-\frac{2}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right]\\
&+\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\sum_{j=1}^{d}\left[\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2+\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2\right]\\
&\leq\sum_{j=1}^{d}\sum_{t=1}^{T}\left[(m_{t,j}-m_j)^2+\sigma_{t,j}^2\right]\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2}-\frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}\right]\\
&+\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\sum_{j=1}^{d}\left[\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2+\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2\right].
\end{aligned}$$

The key point in the following is that the difference

$$\frac{1}{\eta_{t,j}\sigma_{t,j}^2}-\frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2}$$

does not depend on j on account of the formula $\eta_{t,j}=K/(\sqrt{t}\sigma_{t,j}^2)>0$. We also recall that

$$\sum_{j=1}^{d}(m_{t,j}-m_j)^2+\sigma_{t,j}^2\leq D^2.$$

Moreover,

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\leq 2\sqrt{T},$$

so setting $\eta_{t,j}=\frac{K}{\sqrt{t}\sigma_{t,j}^2}>0$ with $K=\frac{D\sqrt{2}}{L}$, we finally have:

$$\sum_{t=1}^{T}\bar{L}_t(\mu_t)-\bar{L}_t(\mu)\leq\frac{1}{K}\sum_{t=1}^{T}(\sqrt{t}-\sqrt{t-1})\sum_{j=1}^{d}[(m_{t,j}-m_j)^2+\sigma_{t,j}^2]+\sum_{t=1}^{T}\frac{K}{\sqrt{t}}L^2$$

$$
\begin{aligned}
&\leq \frac{D^2}{K}\sum_{t=1}^{T}(\sqrt{t}-\sqrt{t-1}) + \frac{KL^2}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\
&= \left(\frac{D^2}{K}+\frac{KL^2}{2}\right)\sqrt{T} \\
&= DL\sqrt{2T},
\end{aligned}
$$

where K is chosen so that it minimizes the bound.

Strongly convex case:

Now, we assume that each function $\bar{L}_t$ is H-strongly convex, for all $m \in \mathcal{M}_m$ and $\mu = (m, 0)$:

$$
\begin{aligned}
\bar{L}_t(\mu_t) - \bar{L}_t(\mu) &\leq \nabla\bar{L}_t(\mu_t)^T(\mu_t-\mu) - \frac{H}{2}\|\mu_t-\mu\|^2 \\
&= \sum_{j=1}^{d}\left[\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)(m_{t,j}-m_j) + \frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)\sigma_{t,j} - \frac{H}{2}(m_{t,j}-m_j)^2 - \frac{H}{2}\sigma_{t,j}^2\right].
\end{aligned}
$$

Again,

$$
\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)(m_{t,j}-m_j) = \frac{(m_{t,j}-m_j)^2-(m_{t+1,j}-m_j)^2}{2\eta_{t,j}\sigma_{t,j}^2} + \frac{\eta_{t,j}\sigma_{t,j}^2\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2}{2}
$$

and

$$
\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)\sigma_{t,j} = \frac{\sigma_{t,j}^2-\sigma_{t+1,j}^2}{\eta_{t,j}\sigma_{t,j}^2\sqrt{1+\left(\frac{\eta_{t,j}\sigma_{t,j}\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)}{2}\right)^2}} + \frac{\eta_{t,j}\sigma_{t,j}^2\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2}{2\sqrt{1+\left(\frac{\eta_{t,j}\sigma_{t,j}\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)}{2}\right)^2}},
$$

and then as previously with $\eta_{t,j} = \frac{2}{Ht\sigma_{t,j}^2}$:

$$
\begin{aligned}
\sum_{t=1}^{T}\bar{L}_t(\mu_t) - \bar{L}_t(\mu) &\leq \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{(m_{t,j}-m_j)^2}{2}\left[\frac{1}{\eta_{t,j}\sigma_{t,j}^2} - \frac{1}{\eta_{t-1,j}\sigma_{t-1,j}^2} - H\right] \\
&\quad + \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\frac{\partial\bar{L}_t}{\partial m_j}(m_t,\sigma_t)^2 \\
&\quad + \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\sigma_{t,j}^2}{2}\left[\frac{2}{\eta_{t,j}\sigma_{t,j}^2} - \frac{2}{\eta_{t-1,j}\sigma_{t-1,j}^2} - H\right] \\
&\quad + \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{\eta_{t,j}\sigma_{t,j}^2}{2}\frac{\partial\bar{L}_t}{\partial\sigma_j}(m_t,\sigma_t)^2 \\
&\leq \sum_{j=1}^{d}\sum_{t=1}^{T}\frac{(m_{t,j}-m_j)^2}{2}\left[\frac{tH}{2} - \frac{(t-1)H}{2} - H\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{j=1}^{d} \sum_{t=1}^{T} \frac{\sigma_{t,j}^2}{2} \left[tH - (t-1)H - H \right] \\
&+ \sum_{t=1}^{T} \frac{1}{Ht} \sum_{j=1}^{d} \left[\frac{\partial \bar{L}_t}{\partial m_j}(m_t, \sigma_t)^2 + \frac{\partial \bar{L}_t}{\partial \sigma_j}(m_t, \sigma_t)^2 \right] \\
&\leq \sum_{j=1}^{d} \sum_{t=1}^{T} \frac{(m_{t,j} - m_j)^2}{2} \left[\frac{H}{2} - H \right] + 0 + \sum_{t=1}^{T} \frac{L^2}{Ht} \\
&\leq \frac{L^2}{H}(1 + \log(T)),
\end{aligned}
$$

which ends the proof. □

Proof of Theorem 7.7.1. The proof is exactly the same as for Theorem 7.4.1. As previously, we first prove by recursion on T that

$$
\forall \mu \in \mathcal{M},\ \sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\|\mu - \mu_1\|^2}{\eta}. \tag{7.24}
$$

It is obvious that it holds for $T = 0$. Assume now that (7.24) holds for some integer $T - 1$. Then for all $\mu \in M$,

$$
\begin{aligned}
\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &= \sum_{t=1}^{T-1} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) + \mu_{T+1}^T \nabla \bar{L}_T(\mu_T) \\
&\leq \sum_{t=1}^{T-1} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\|\mu - \mu_1\|^2}{\eta} + \mu_{T+1}^T \nabla \bar{L}_T(\mu_T)
\end{aligned}
$$

as (7.24) holds for $T - 1$. Apply this again to $\mu = \mu_{T+1}$:

$$
\begin{aligned}
\sum_{t=1}^{T} \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &\leq \sum_{t=1}^{T} \mu_{T+1}^T \nabla \bar{L}_t(\mu_t) + \frac{\|\mu - \mu_1\|^2}{\eta} \\
&= \min_{m \in \mathcal{M}} \left[\sum_{t=1}^{T} m^T \nabla \bar{L}_t(\mu_t) + \frac{\|\mu - \mu_1\|^2}{\eta} \right], \text{ by definition of } \mu_{T+1} \\
&\leq \sum_{t=1}^{T} \mu^T \nabla \bar{L}_t(\mu_t) + \frac{\|\mu - \mu_1\|^2}{\eta}
\end{aligned}
$$

for all $\mu \in \mathcal{M}$. Thus, (7.24) holds for T, and thus for integers.

We prove now that for any $t \in \mathbb{N}$,

$$
\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) \leq \eta L^2. \tag{7.25}
$$

Indeed,

$$
\begin{aligned}
\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t) &= (\mu_t - \mu_{t+1})^T \nabla \bar{L}_t(\mu_t) \\
&\leq \|\mu_t - \mu_{t+1}\| \|\nabla \bar{L}_t(\mu_t)\| \\
&\leq L \|\mu_t - \mu_{t+1}\| \qquad (7.26)
\end{aligned}
$$

as previously. Define

$$G_t(\mu) = \sum_{i=1}^{t-1} \mu^T \nabla \bar{L}_t(\mu_i) + \frac{\|\mu - \mu_1\|^2}{\eta}.$$

Obviously, G_t is $1/\eta$-strongly convex: for any (μ, μ'),

$$G_t(\mu') - G_t(\mu) \geq (\mu' - \mu)^T \nabla G_t(\mu) + \frac{\|\mu' - \mu\|^2}{2\eta}.$$

In particular, μ_t is a minimizer of G_t:

$$G_t(\mu_{t+1}) - G_t(\mu_t) \geq \frac{\|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Similarly,

$$G_{t+1}(\mu_t) - G_{t+1}(\mu_{t+1}) \geq \frac{\|\mu_{t+1} - \mu_t\|^2}{2\eta}.$$

Hence:

$$\bar{L}_t(\mu_t) - \bar{L}_t(\mu_{t+1}) \geq \frac{\|\mu_{t+1} - \mu_t\|^2}{\eta},$$

and then

$$\|\mu_{t+1} - \mu_t\| \leq \sqrt{\eta \left[\mu_t^T \nabla \bar{L}_t(\mu_t) - \mu_{t+1}^T \nabla \bar{L}_t(\mu_t)\right]}$$

which combined with (7.26) leads to (7.25).

Finally, as for Theorem 7.4.1:

$$\sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) - \sum_{t=1}^{T} \mathbb{E}_{\theta \sim q_\mu}[\ell_t(\theta)] \leq \eta T L^2 + \frac{\|\mu - \mu_1\|^2}{\eta},$$

which ends the proof. □

Part IV

Robustness to misspecification via Maximum Mean Discrepancy

Chapter 8

Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence

Many works in statistics aim at designing a universal estimation procedure. This question is of major interest, in particular because it leads to robust estimators, a very hot topic in statistics and machine learning. In this chapter, we tackle the problem of universal estimation using a minimum distance estimator presented in Briol et al. (2019) based on the Maximum Mean Discrepancy. We show that the estimator is robust to both dependence and to the presence of outliers in the dataset. We also highlight the connections that may exist with minimum distance estimators using L_2-distance. Finally, we provide a theoretical study of the stochastic gradient descent algorithm used to compute the estimator, and we support our findings with numerical simulations.

8.1 Introduction

One of the main challenges in statistics is the design of a *universal* estimation procedure. Given data, a universal procedure is an algorithm that provides an estimator of the generating distribution which is simultaneously statistically optimal when the true distribution belongs to the model, and robust otherwise. Typically, a universal estimator is consistent for any model, with minimax-optimal or fast rates of convergence and is robust to small departures from the model assumptions (Bickel, 1976) such as sparse instead of dense effects or non-Gaussian errors in high dimensional linear regression. Unfortunately, most statistical procedures are based upon strong assumptions on the model or on the corresponding parameter set, and very famous estimation methods such as maximum likelihood estimation (MLE), method of moments or Bayesian posterior inference may fail even on simple problems when such assumptions do not hold. For instance, even though MLE is consistent and asymptotically normal with optimal rates of convergence in parametric estimation under suitable regularity assumptions (Le Cam, 1970; Van der

Vaart, 2000) and in nonparametric estimation under entropy conditions, this method behaves poorly in case of misspecification when the true generating distribution of the data does not belong to the chosen model.

Let us investigate a simple example presented in Birgé (2006) that illustrates the non-universal characteristic of MLE. We observe a collection of n independent and identically distributed (i.i.d) random variables $X_1, ..., X_n$ that are distributed according to some mixture distribution $P_n^0 = (1-2n^{-1})\mathcal{U}([0,1/10]) + 2n^{-1}\mathcal{U}([1/10,9/10])$ where $\mathcal{U}([a,b])$ is the uniform distribution between a and b. We consider the parametric model of independent uniform distributions $\mathcal{U}([0,\theta])$, $0 \leq \theta < 1$, and we choose the squared Hellinger distance $h^2(\cdot,\cdot)$ as the risk measure. Here the maximum likelihood is the maximum of the observations $X_{(n)} := \max(X_1, ..., X_n)$, and $\mathcal{U}([0,1/10])$ is a good approximation of the generating distribution P_n^0 as $h^2(P_n^0, \mathcal{U}([0,1/10])) < 5/4n$ for $n \geq 4$. Hence, one would expect that $\mathbb{E}[h^2(P_n^0, \mathcal{U}([0, X_{(n)}]))]$ goes to 0 as $n \to +\infty$, which is actually not the case. We do not even have consistency: $\mathbb{E}[h^2(P_n^0, \mathcal{U}([0, X_{(n)}]))] > 0.38$. Hence, the MLE is not robust to this small deviation from the parametric assumption. Other problems can arise for the MLE: for instance, the quadratic risk can be much bigger than the minimax risk, and the performance of the MLE may be too sensitive to the choice of the family of densities used in the model, see respectively Birgé (2006) and Baraud and Birgé (2016). The same happens in Bayesian statistics: the regular posterior distribution is not always robust to model misspecification. Indeed, authors of Barron et al. (1999); Grünwald et al. (2017) show pathologic cases where the posterior does not concentrate to the true distribution.

Universal estimation is all the more important since it provides a generic approach to tackle the more and more popular problem of robustness to outliers under the i.i.d assumption, although definitions and goals involved in robust statistics are quite different from the universal estimation perspective. Hüber introduced a framework that models situations where a small fraction ϵ of data is contaminated, and he assumes that the true generated distribution can be written $(1-\epsilon)P_{\theta^0} + \epsilon Q$ where Q is the contaminating distribution and ϵ is the proportion of corrupted observations (Hüber, 1964). The goal when using this approach is to estimate the true parameter θ^0 given a misspecified model $\{P_\theta / \theta \in \Theta\}$ with $\theta^0 \in \Theta$. A procedure is then said to be robust in this case if it leads to a good estimation of the true parameter θ^0. More generally, when a procedure is able to provide a good estimate of the generating distribution of i.i.d data when a small proportion of them is corrupted, whatever the values of these outliers, then such an estimator is considered as robust.

8.1.1 Related work

Several authors attempted to design a general universal estimation method. Sture Holm (Bickel, 1976) suggested that Minimum Distance Estimators (MDE) were the most natural procedures being robust to misspecification. Motivated by Wolfowitz (1957); Parr and Schucany (1980), MDE consists in minimizing some probability distance d between the empirical distribution and a distribution in the model. The MDE $\widehat{\theta}_n$ is defined by:

$$d(\widehat{P}_n, P_{\widehat{\theta}_n}) = \inf_{\theta \in \Theta} d(\widehat{P}_n, P_\theta)$$

where $\widehat{P}_n$ is the empirical measure and Θ the parameter set associated to the model. If the minimum does not exist, then one can consider a ε-approximate solution. In fact, this minimum distance estimator is used in many usual procedures. Indeed, the generalized method of moments (Hansen, 1982) is actually defined as minimizing the weighted Euclidean distance between moments of $\widehat{P}_n$ and P_θ while the MLE minimizes the KL divergence. When the distance d is wisely chosen, among others, it must be bounded, then MDE can be robust and consistent. A typical choice of the metric is the Total Variation (TV) distance (Yatracos, 1985; Devroye and Lugosi, 2001). Yatracos (1985) showed that under the i.i.d assumption, the minimum distance estimator based on the TV metric is uniformly consistent in TV distance and is robust to misspecification without any assumption on the parameter set, with a rate of convergence depending on the Kolmogorov entropy of the space of measures. A few decades later, Devroye and Lugosi studied in details the skeleton estimate, a variant of the estimator of (Yatracos, 1985) that is based on the TV-distance restricted to the so-called Yatracos sets, see (Devroye and Lugosi, 2001). Unfortunately, the skeleton estimate and the original Yatracos estimate are not computationally tractable.

In Baraud and Birgé (2016) and Baraud et al. (2017), Baraud, Birgé and Sart introduced in the independent framework the so-called ρ-estimators, a universal method that retains some appealing properties of the MLE such as efficiency under some regularity assumptions, while being robust to Hellinger deviations. ρ-estimation is inspired from T-estimation (Birgé, 2006), itself inspired from earlier works of Le Cam (1973, 1975) and Birgé (1983), and goes beyond the classical compactness assumption used in T-estimation. In compact models, ρ-estimators can be seen as variants of T-estimators also based on robust tests, but they can be extended to noncompact models such as linear regression with fixed or random design with various error distributions. As T-estimators, they enjoy robustness properties, but involve other metric dimensions which lead to optimal rates of convergence with respect to the Hellinger distance even in cases where T-estimators can not be defined. Moreover, note that when the sample size is large enough, ρ-estimation recovers the usual MLE in density estimation when the model is parametric, well-specified and regular enough. Hence, ρ-estimation can be seen as a robust version of the MLE, but once again, such a strategy is intractable.

More recently, Briol et al. (2019) showed that using the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) as a minimum distance estimator leads to both robust and tractable estimation in the i.i.d case. MMD, a metric based on embeddings of probability measures into a reproducing kernel Hilbert space, has been applied successfully in a wide range of problems such as kernel Bayesian inference (Song and Gretton, 2011), approximate Bayesian computation (Park et al., 2016), two-sample (Gretton et al., 2012) and goodness-of-fit testing (Jitkrittum et al., 2017), and MMD GANs (Dziugaite et al., 2015; Li et al., 2015) and autoencoders (Zhao et al., 2017), to name a few prominent examples. Such minimum MMD-based estimators are proved to be consistent, asymptotically normal and robust to model misspecification. The trade-off between the statistical efficiency and the robustness is made through the choice of the kernel. The authors investigated the geometry induced by the MMD on the finite-dimensional parameter space and introduced a (natural) gradient descent algorithm for efficient computation of the estimator. This algorithm is inspired from the stochastic gradient descent (SGD) used in the context of

MMD GANs where the usual discriminator is replaced with a two-sample test based on MMD (Dziugaite et al., 2015). These results were extended in the Bayesian framework by Chérief-Abdellatif and Alquier (2020).

8.1.2 Contributions

In this chapter, we further investigate universality properties of minimum distance estimation based on MMD distance (Briol et al., 2019). Inspired by the related literature, our contributions in this chapter are the following:

- We go beyond the classical i.i.d framework. Indeed, we prove that the the estimator is robust to dependence between observations. To do so, we introduce a new dependence coefficient expressed as a covariance in some reproducing kernel Hilbert space, and which is very simple to use in practice.
- We show that our oracle inequalities imply robust estimation under the i.i.d assumption in the Hüber contamination model and in the case of adversarial contamination.
- We also highlight the connection between our MMD estimator and minimum distance estimation using L_2-distance for radial kernels.
- We propose a theoretical analysis of the SGD algorithm used to compute this estimator in Briol et al. (2019) and Dziugaite et al. (2015) for some finite dimensional models. We provide numerical simulation to illustrate our theoretical results.

The first result of this chapter is a generalization bound in the non-i.i.d setting. It states that under a very general dependent assumption, the generalization error with respect to the MMD distance decreases in $n^{-1/2}$ as $n \to +\infty$. This result extends the inequalities in Briol et al. (2019) that are only available in the i.i.d framework, and is obtained using dependence concepts for stochastic processes. Since the seminal work of Rosenblatt (1956), many mixing conditions, that is, restrictions on the dependence bewteen observations, were defined. These conditions lead to limit theorems (LLN, CLT) useful to analyze the asymptotic behavior of time series (Doukhan, 1994). Nevertheless, checking mixing assumptions is difficult in practice and many classes of processes that are of interest in statistics such as elementary Markov chains are sometimes not mixing. More recently, Doukhan and Louhichi (1999) proposed a new weak dependence condition for time series that is built on covariance-based coefficients which are much easier to compute than mixing ones, and that is more general than mixing as it stands for most relevant classes of processes. We introduce in this chapter a new dependence coefficient in the wake of Doukhan and Louhichi (1999) which can be expressed as a covariance in some reproducing kernel Hilbert space associated with MMD, which can be easily computed in many situations and which may be related to usual mixing coefficients such as the popular β-mixing one. We show that a weak assumption on this new dependence coefficient can relax the i.i.d assumption of Briol et al. (2019) and can lead to valid generalization bounds even in the dependent setting.

Also, we provide inequalities in L_2-distance. Previous attempts of designing a universal estimator lead to bounds in TV or Hellinger distances (Baraud and Birgé, 2016;

Devroye and Lugosi, 2001), but state that the quadratic loss is to be avoided as a minimum distance estimator, in particular because this metric exclude distributions for which no density is available. We show here that for radial kernels, the MMD distance is a good approximation of the L_2-metric when densities exist, and thus can be seen as a "universalized" and robustified version of the L_2-distance-based minimum distance estimator. We introduce conditions on the kernel leading to valid generalization bounds in quadratic loss. Moreover, we show how our results can be used in the context of robust estimation with contamination, and how they can provide statistically optimal robust Gaussian mean estimation with respect to the Euclidean distance.

Regarding computational issues, we provide a Stochastic Gradient Descent algorithm as in Briol et al. (2019); Dziugaite et al. (2015) involving a U-statistic approximation of the expectation in the formula of the MMD distance. We theoretically analyze this algorithm in parametric estimation using a convex parameter set. We also perform numerical simulations that illustrate the efficiency of our method, especially by testing the behavior of the algorithm in the presence of outliers.

The rest of this chapter is organized as follows. Section 8.2 defines the MMD-based minimum distance estimator and our new dependence coefficient based on the kernel mean embedding. Section 8.3 provides nonasymptotic bounds in the dependent and misspecified framework, with results in robust parametric estimation and the connection with density estimation using quadratic loss. Section 8.4 illustrates the efficiency of our method in several different frameworks. We finally present an SGD algorithm with theoretical convergence guarantees in Section 8.5 and we perform numerical simulations in Section 8.6. Section 8.7 is dedicated to the proofs.

8.2 Background and definitions

In this section, we introduce first some notations and present the statistical setting of the paper in Section 2.1. Then we remind in Section 2.2 some theory on reproducing kernel Hilbert spaces (RKHS) and we define both the maximum mean discrepancy (MMD) and our minimum distance estimator based on the MMD. Finally, we introduce in Section 2.3 a new dependence coefficient expressed as a covariance in a RKHS.

8.2.1 Statistical setting

We shall consider a dependent setting throughout the paper. We observe in a measurable space $\left(\mathbb{X}, \mathcal{X}\right)$ a collection of n random variables X_1,...,X_n generated from a stationary process. This implies that the X_i's are identically distributed, we will let P^0 denote their marginal distribution. Note that this include as an example the case where the X_i's are i.i.d with generating distribution P^0. We introduce a statistical model $\{P_\theta / \theta \in \Theta\}$ indexed by a parameter space Θ.

8.2.2 Maximum Mean Discrepancy

We consider a positive definite kernel function k, i.e a symmetric function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that for any integer $n \geq 1$, for any $x_1, ..., x_n \in \mathbb{X}$ and for any $c_1, ..., c_n \in \mathbb{R}$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

We then consider the reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ associated with the kernel k which satisfies the reproducing property $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$ for any function $f \in \mathcal{H}_k$ and any $x \in \mathbb{X}$. From now on, we assume that the kernel is bounded by some positive constant, that will be assumed to be 1 without loss of generality.

Now we introduce the notion of *kernel mean embedding*, a Hilbert space embedding of a probability measure that can be viewed as a generalization of the original feature map used in support vector machines and other kernel methods. The basic idea is to map measures into the RKHS $\mathcal{H}_k$, enabling to apply all various kernel methods to the underlying measures. Given a probability measure P, we define the mean embedding $\mu_P \in \mathcal{H}_k$ as:

$$\mu_P(\cdot) := \mathbb{E}_{X \sim P}[k(X, \cdot)] \in \mathcal{H}_k.$$

All the applications and the theoretical properties of those embeddings have been well studied (Muandet et al., 2017). In particular, the mean embedding μ_P satisfies the relationship $\mathbb{E}_{X \sim P}[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}_k}$ for any function $f \in \mathcal{H}_k$, and induces a semi-metric [1] on measures called maximum mean discrepancy and defined for two measures P and Q as follows:

$$\mathbb{D}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$$

or alternatively

$$\mathbb{D}_k^2(P, Q) = \mathbb{E}_{X, X' \sim P}[k(X, X')] - 2\mathbb{E}_{X \sim P, Y \sim Q}[k(X, Y)] + \mathbb{E}_{Y, Y' \sim Q}[k(Y, Y')].$$

A kernel k is said to be characteristic if $P \mapsto \mu_P$ is injective. This ensures that $\mathbb{D}_k$ is a metric, and not only a semi-metric. Subsection 3.3.1 of the thorough survey Muandet et al. (2017) provides a wide range of conditions ensuring that k is characteristic. They also provide many examples of characteristic kernels, see their Table 3.1. Among others, the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\gamma^2)$ and the Laplace kernel $k(x, y) = \exp(-\|x - y\|/\gamma)$, that we will use in all of our applications, are known to be characteristic. From now, we will assume that k is characteristic.

Note that there are many applications of the kernel mean embedding and MMD in statistics such as two-sample testing (Gretton et al., 2012), change-point detection (Arlot et al., 2012), detection (Lerasle et al., 2019), we refer the reader to Liu et al. (2019) for a thorough introduction to the applications of kernels and MMD to computational biology.

Here, we will focus on estimation of parameters based on MMD. This principle was used to train generative networks (Dziugaite et al., 2015; Li et al., 2015), it's only recently

[1] This means that $P \to \|\mu_P\|_{\mathcal{H}_k}$ satisfies the requirements of a norm besides $\|\mu_P - \mu_Q\|_{\mathcal{H}_k} = 0$ only for $\mu_P = \mu_Q$.

that it was studied as a general principle for estimation (Briol et al., 2019). Following these papers we define the MMD estimator $\widehat{\theta}_n$ such that:

$$\mathbb{D}_k(P_{\widehat{\theta}_n}, \widehat{P}_n) = \inf_{\theta \in \Theta} \mathbb{D}_k(P_\theta, \widehat{P}_n)$$

where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ is the empirical measure, i.e.:

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \left\{ \mathbb{E}_{X,X' \sim P_\theta}[k(X, X')] - \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{X \sim P_\theta}[k(X, X_i)] \right\}.$$

It could be that there is no minimizer, see the discussion in Theorem 1 page 9 in Briol et al. (2019). In this case, we can use an approximate minimizer. More precisely, for any $\varepsilon > 0$ we can always find a $\widehat{\theta}_{n,\varepsilon}$ such that:

$$\mathbb{D}_k(P_{\widehat{\theta}_{n,\varepsilon}}, \widehat{P}_n) \leq \inf_{\theta \in \Theta} \mathbb{D}_k(P_\theta, \widehat{P}_n) + \varepsilon.$$

In what follows, we will consider the case where the minimizer exists (that is, $\varepsilon = 0$) but when this is not the case, everything can be easily extended by considering $\widehat{\theta}_{n,1/n}$.

8.2.3 Covariances in RKHS

In this subsection, we introduce and discuss a new dependence coefficient based on the kernel mean embedding. This coefficient allows to go beyond the i.i.d case and to show that the MMD estimator of Briol et al. (2019) is actually robust to dependence. Please refer to Briol et al. (2019) for results in the i.i.d case.

Definition 8.2.1. *We define, for any $t \in \mathbb{N}$,*

$$\varrho_t = \left| \mathbb{E} \left\langle k(X_t, \cdot) - \mu_{P^0}, k(X_0, \cdot) - \mu_{P^0} \right\rangle_{\mathcal{H}_k} \right|.$$

In the i.i.d case, note that $\varrho_t = 0$ for any $t \geq 1$. In general, the following assumption will ensure the consistency of our estimator:

Assumption 8.2.1. *There is a $\Sigma < +\infty$ such that, for any n, $\sum_{t=1}^n \varrho_t \leq \Sigma$.*

Our mean embedding dependence coefficient may be seen as a covariance expressed in the RKHS $\mathcal{H}_k$. We shall see throughout the paper that the kernel mean embedding coefficient ϱ_t can be easily computed in many situations, and that it is closely related to the widely used mixing coefficients. In particular, we will show in Section 4.2 that our coefficient ϱ_t is upper-bounded by the celebrated β-mixing coefficient for radial kernels in the case of a strictly stationary time series. More importantly, we exhibit in Section 4.3 an example of a non-mixing process such that $\sum_{t=1}^{+\infty} \beta_t = \sum_{t=1}^{+\infty} \alpha_t = +\infty$ but for which ϱ_t is exponentially decaying and hence Assumption 8.2.1 still holds, which means that ϱ_t is a more general and weaker dependence coefficient than usual mixing coefficients, due to its covariance structure. Hence, Assumption 8.2.1 may be referred to as a weak dependence condition in the wake of the concept of weak dependence introduced in Doukhan and Louhichi (1999). Using a Hoeffding-like inequality due to Rio (2017b), we will show in the next section that under Assumption 8.2.1, we can obtain a nonasymptotic generalization bound of the same order than in the i.i.d case.

8.3 Nonasymptotic bounds in the dependent, misspecified case

In this section, we provide nonasymptotic generalization bounds in MMD distance for the minimum MMD estimator. In particular, we show in Subsection 8.3.1 that under a weak dependence assumption, it is robust to both dependence and misspecification, and is consistent at the same $n^{-1/2}$ rate than in the i.i.d case. In particular, we give explicit bounds in the Hüber contamination model and in a more general adversarial setting in Subsection 8.3.2. Finally, we connect in Subsection 8.3.3 our MMD-based estimator to an L_2-based one when densities exist. We discuss conditions on the kernel and provide oracle inequalities in L_2-distance.

8.3.1 Estimation with respect to the MMD distance

First, we begin with a theorem that gives an upper bound on the generalization error, i.e the expectation of $\mathbb{D}_k(P_{\widehat{\theta}_n}, P^0)$. The rate of convergence of this error is of order $n^{-1/2}$ independently of the dimensions d and the property of the kernel.

Theorem 8.3.1. *We have:*

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + 2\sqrt{\frac{1+2\sum_{t=1}^n \varrho_t}{n}}.$$

As a consequence, under Assumption 8.2.1:

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + 2\sqrt{\frac{1+2\Sigma}{n}}.$$

We remind that all the proofs are deferred to Section 8.7. It is also possible to provide a result that holds with large probability as in Briol et al. (2019) and in Dziugaite et al. (2015). Naturally, it requires stronger assumptions, and the conditions on the dependence become more intricate in this case. Here, we use a condition introduced in Louhichi (1998) for generic metric spaces that we adapt to the kernel embedding and to stationarity:

Assumption 8.3.1. *Assume that there is a family $(\gamma_\ell)_\ell$ of nonnegative numbers such that, for any integer n, for any $\ell \in \{1,\dots,n-1\}$ and any function $g : \mathcal{H}_k^\ell \to \mathbb{R}$ such that*

$$|g(a_1,\dots,a_\ell) - g(b_1,\dots,b_\ell)| \leq \sum_{i=1}^{\ell} \|a_i - b_i\|_{\mathcal{H}_k},$$

we have, $\left|\mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}},\dots,\mu_{\delta_{X_n}})|X_1,\dots,X_\ell] - \mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}},\dots,\mu_{\delta_{X_n}})]\right| \leq \gamma_1 + \cdots + \gamma_{n+\ell-1}$, almost surely. Assume that there is a $\Gamma = \sum_{\ell\geq 1}\gamma_\ell < \infty$.

Again, note that in the case of independence, we can take all the $\gamma_{i,j} = 0$ and hence $\Gamma = 0$ in addition to $\Sigma = 0$. We can now state our result in probability:

Theorem 8.3.2. *Assume that Assumptions 8.2.1 and 8.3.1 are satisfied. Then, for any* $\delta \in (0,1)$,

$$\mathbb{P}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right) \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + 2\frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right] \geq 1-\delta.$$

Assumption 8.3.1 is fundamental to obtain a result in probability. Indeed, the rate of convergence in Theorem 8.3.2 is characterized by some concentration inequality upper bounding the MMD distance between the empirical and the true distribution as done in Briol et al. (2019). Nevertheless, the proof of this inequality in Briol et al. (2019) is based on a Hoeffding-type inequality known as McDiarmid's inequality (McDiarmid, 1989) that is only valid for independent variables, which makes this inequality not applicable in our dependent setting. Hence we use a version of McDiarmid's inequality for time series obtained by Rio (2013) which is available under a polynomial decay assumption on some mixing dependence coefficients $(\gamma_{i,j})_{1\leq i<j}$. This decay assumption is expressed here in the RKHS $\mathcal{H}_k$ of Kernel k as Assumption 8.3.1.

Remark 8.3.1 (The i.i.d case)**.** *Note that when the* X_i*'s are i.i.d, Assumptions 8.2.1 and 8.3.1 are always satisfied with* $\Sigma = \Gamma = 0$ *and thus Theorem 8.3.1 gives simply*

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + \frac{2}{\sqrt{n}}$$

while Theorem 8.3.2 gives

$$\mathbb{P}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right) \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + 2\frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right] \geq 1-\delta.$$

8.3.2 Robust parametric estimation

Contamination models

As explained in the introduction, when all observations but a small proportion of them are sampled independently from a generating distribution P_{θ^0} ($\theta^0 \in \Theta$), robust parametric estimation consists in finding estimators being both rate optimal and resistant to outliers. Two among the most popular frameworks for studying robust estimation are the so-called Hüber's contamination model and the adversarial contamination model.

Hüber's contamination model is as follows. We observe a collection of random variables $X_1, ..., X_n$. We consider a contamination rate $\epsilon \in (0, 1/2)$, latent i.i.d random variables $Z_1, ..., Z_n \sim \text{Ber}(\epsilon)$ and some noise distribution Q, such that the distribution of X_i given $Z_i = 0$ is P_{θ^0}, and that the distribution of X_i given $Z_i = 1$ is Q. Hence, the observations X_i's are independent and sampled from the mixture $P^0 = (1-\epsilon)P_{\theta^0} + \epsilon Q$.

The adversarial model is more general. Contrary to Hüber's contamination where outliers were all sampled from the contaminating distribution, we do not make any particular

assumption on the outliers here. Hence, we shall adopt slightly different notations. We assume that $X_1, \dots, X_n$ are identically distributed from P_{θ^0} for some $\theta^0 \in \Theta$. However, the statistician only observes $\widetilde{X_1}, \dots, \widetilde{X_n}$ where $\widetilde{X_i}$ can be any arbitrary value for $i \in \mathcal{O}$, where $\mathcal{O}$ is an arbitrary set subject to the constraint $\mathcal{O} \leq \epsilon n$, and $\widetilde{X_i} = X_i$ for $i \notin \mathcal{O}$. The estimators are built based on these observations $\widetilde{X_1}, \dots, \widetilde{X_n}$.

Literature

One hot research trend in robust statistics is focused on the search of both statistically optimal and computationally tractable procedures for the Gaussian mean estimation problem $\{P_\theta = \mathcal{N}(\theta, I_d)/\theta \in \mathbb{R}^d\}$ in the presence of outliers under the i.i.d assumption, which remains a major challenge. Usual robust estimators such as the coordinatewise median and the geometric median are known to be suboptimal in this case, and there is a need to look at more complex estimators such as Tukey's median that achieves the minimax optimal rate of convergence $\max(\frac{d}{n}, \epsilon^2)$ with respect to the squared Euclidean distance, where d is the dimension, n is the sample size and ϵ is the proportion of corrupted data. Unfortunately, computation of Tukey's median is not tractable and even approximate algorithms lead to an $\mathcal{O}(n^d)$ complexity (Chan, 2004; Amenta et al., 2000). This has led to the rise of the recent studies in robust statistics which adress how to build robust and optimal statistical procedures, in the wake of the works of Tukey (1975) and Hüber (1964), but that are also computationally efficient.

This research area started with two seminal works presenting two procedures for the normal mean estimation problem: the *iterative filtering* (Diakonikolas et al., 2016) and the *dimension halving* (Lai et al., 2016). These algorithms are based upon the idea of using higher moments in order to obtain a good robust moment estimation, and are minimax optimal up to a poly-logarithmic factor in polynomial time in the adversarial contamination model. This idea was then used in several other problems in robust statistics, for instance in sparse functionals estimation (Du et al., 2017), clustering (Kothari et al., 2018), mixtures of spherical Gaussians learning (Diakonikolas et al., 2018b), and robust linear regression (Diakonikolas et al., 2018c). In Hüber's contamination model, a recent paper of Collier and Dalalyan (2017) achieves the minimax rate without any extra factor in the $\epsilon = \mathcal{O}(\min(d^{-1/2}, n^{-1/4}))$ regime with an improved overall complexity. Meanwhile, Gao et al. (2019) offers a different perspective on robust estimation and connects the robust normal mean estimation problem with Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Biau et al., 2018), what enables computing robust estimators using efficient tools developed for training GANs. Hence, the authors compute depth-like estimators that retain the same appealing robustness properties than Tukey's median and that can be trained using stochastic gradient descent (SGD) algorithms that were originally designed for GANs.

Another popular approach for the more general problem of mean estimation under the i.i.d assumption in the presence of outliers is the study of finite-sample sub-Gaussian deviation bounds. Indeed, designing estimators achieving sub-Gaussian performance under minimal assumptions ensures robustness to outliers that are inevitably present when the generating distribution is heavy-tailed. In the univariate case, some estimators present a sub-Gaussian behavior for all distributions under first and second order moments. A

simple but powerful strategy, the Median-of-Means (MOM), dates back to Nemirovski and Yudin (1983); Jerrum et al. (1986); Alon et al. (2008). This method consists in randomly splitting the data into several equal-size blocks, then computing the empirical mean within each block, and finally taking the median of them. Most MOM-based procedures lead to estimators that are simultaneously statistically optimal (Lugosi and Mendelson, 2016; Devroye et al., 2016; Lecué et al., 2018; Lerasle et al., 2019; Chinot et al., 2019) and computationally efficient (Hopkins, 2019; Cherapanamjeri et al., 2019; Depersin and Lecué, 2019). Moreover, this approach be easily extended to the multivariate case (Minsker, 2015; Hsu and Sabato, 2016). An important advantage is that the MOM estimator has good performance even for distributions with infinite variance. An elegant alternative to the MOM strategy is due to Catoni, whose estimator is based on PAC-Bayesian truncation in order to mitigate heavy tails (Catoni, 2012). It has the same performance guarantees than the MOM method but with sharper and near-optimal constants. In Catoni and Giulini (2017), Catoni and Giulini proposed a very simple and trivial-to-compute multidimensional extension of Catoni's M-estimator defined as an empirical average of the data, with the observations with large norm shrunk towards zero, and that still satisfies a sub-Gaussian concentration using PAC-Bayes inequalities. The influence function of Catoni and Giulini has been widely used since then, see Giulini (2017, 2018); Holland (2019a,b). We refer the reader to the excellent review of Lugosi and Mendelson (2019) for more details on those mean estimation procedures.

Robust MMD estimation

In this section, we show the properties of our MMD-based estimator in robust parametric estimation with outliers, both in Hüber's and in the adversarial contamination model. Our bounds are obtained by working directly in the RKHS rather than in the parameter space, and going back and forth between the two spaces.

First we consider Hüber's contamination model (Hüber, 1964). The objective is to estimate P_{θ^0} by observing contaminated random variables X_1, ..., X_n with actual distribution is $P^0 = (1-\alpha)P_{\theta^0} + \alpha Q$ for some Q, and some $0 \leq \alpha \leq \epsilon$. We state the key following lemma:

Lemma 8.3.3. *We have:* $|\mathbb{D}_k(P_{\widehat{\theta}_n}, P^0) - \mathbb{D}_k(P_{\widehat{\theta}_n}, P_{\theta^0})| \leq 2\epsilon$.

As a consequence of Lemma 8.3.3 and Theorem 8.3.1, we have the following result.

Corollary 8.3.4. *Assume that* $X_1, \ldots, X_n$ *are identically distributed from* $P^0 = (1-\alpha)P_{\theta^0} + \alpha Q$ *for some* $\theta^0 \in \Theta$, *some* Q, *with* $0 \leq \alpha \leq \epsilon$. *Then:*

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P_{\theta^0}\right)\right] \leq 2\epsilon + 2\sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}}.$$

If moreover we assume that Assumptions 8.2.1 and 8.3.1 are satisfied, then for any $\delta \in (0,1)$,

$$\mathbb{P}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P_{\theta^0}\right) \leq 2\left(\epsilon + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)\right] \geq 1 - \delta.$$

We obtain a rate $\max(1/\sqrt{n}, \epsilon)$ (in MMD distance) that can be expressed in the same way than the minimax rate (in the Euclidean distance) when estimating a Gaussian mean. When $\epsilon \lesssim 1/\sqrt{n}$, then we recover the minimax rate of convergence without contamination, and when $1/\sqrt{n} \lesssim \epsilon$, then the rate is dominated by the contamination ratio ϵ. Hence, the maximum number of outliers which can be tolerated without breaking down the minimax rate is $n\epsilon \asymp \sqrt{n}$.

This result can also be extended to the adversarial contamination setting, where no assumption is made on the outliers.

Proposition 3. *Assume that $X_1, \ldots, X_n$ are identically distributed from from $P^0 = P_{\theta^0}$ for some $\theta^0 \in \Theta$. However, the statistician only observes $\widetilde{X_1}, \ldots, \widetilde{X_n}$ where $\widetilde{X_i}$ can be any arbitrary value for $i \in \mathcal{O}$, $\mathcal{O}$ is any arbitrary set subject to the constraint $\mathcal{O} \leq \epsilon n$, and $\widetilde{X_i} = X_i$ for $i \notin \mathcal{O}$ and builds the estimator $\widetilde{\theta}_n$ based on these observations:*

$$\mathbb{D}_k\left(P_{\widetilde{\theta}_n}, \frac{1}{n}\sum_{i=1}^n \delta_{\widetilde{X}_i}\right) = \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \frac{1}{n}\sum_{i=1}^n \delta_{\widetilde{X}_i}\right).$$

Then:

$$\mathbb{D}_k\left(P_{\widetilde{\theta}_n}, P_{\theta^0}\right) \leq 4\epsilon + 2\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P_{\theta^0}\right).$$

Thus

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widetilde{\theta}_n}, P_{\theta^0}\right)\right] \leq 4\epsilon + 4\sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}}$$

and, under Assumptions 8.2.1 and 8.3.1, for any $\delta \in (0,1)$,

$$\mathbb{P}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P_{\theta^0}\right) \leq 4\left(\epsilon + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)\right] \geq 1 - \delta.$$

One can see that the rate of convergence we obtain without making any assumption on the outliers is exactly the same than in Hüber's contamination model. The only different thing is that the constant in the right hand side of the inequality is tighter in Hüber's contamination model.

8.3.3 Density estimation with quadratic loss

To obtain universal oracle inequalities in statistics is a goal that has been studied by many authors, and that leads to the question: what distance should be used between probability measures? Results with the total variation (TV) norm were obtained by Yatracos (1985), and more recently, the monograph Devroye and Lugosi (2001) gives a complete overview of estimation with TV. While the Kullback-Leibler distance seems natural as it leads to maximum likelihood estimation, it leads to many problems, including a very strong sensitivity to misspecification, see for example the discussion in Baraud and Birgé (2016). There, the authors derive universal inequalities for the Hellinger distance. The Wasserstein distance became recenly extremely popular, partly due to its ability to take into account the geometry of the space $\mathcal{X}$, see Peyré (2019). Some attempts to obtain

universal estimation with the Wasserstein distance can be found in Bernton et al. (2017); Tat Lee et al. (2018). Note that in Devroye and Lugosi (2001) and in Baraud and Birgé (2016), it is argued that the L_2-distance between densities is not a good distance. Among others:

- it is not universal, as some probability measures don't have densities, while some others have densities that are not in L_2,
- it depends on the choice of the reference measure (usually taken as the Lebesgue measure).

Regarding the first objection, we argue here that for reasonable kernels, the MMD distance is an approximation of the L_2-distance that is defined for any probability distribution. Thus, from the oracle inequalities on the MMD distance it is possible to derive oracle inequalities on the L_2-distance. On the other hand, the MMD distance remains well-defined for any probability measure, even those without density. To this regard, estimation with respect to the MMD distance can be seen as a universal approximation of density estimation in L_2.

Regarding the second objection, our estimator does not depend on any reference measure, but it depends on the choice of the kernel. However, we believe that this can actually be an attractive property – indeed, the popularity of the Wasserstein distance is due to the fact that it takes into account the geometry of $\mathcal{X}$ through the choice of a distance on $\mathcal{X}$. But the same argument hold for MMD-based estimation, the geometry of $\mathcal{X}$ being taken into account through the choice of the kernel. For example, $\mathbb{D}_k(\delta_x, \delta_y) \to 0$ when $x \to y$, a property shared with the Wasserstein distance, and which does not hold for the Hellinger nor the TV distance.

Let us come back to the link with the L_2 distance. Consider for example the Gaussian kernel $k_\gamma(x, y) = \exp(-\|x - y\|^2/\gamma^2)$. On the one hand, when $\gamma \to +\infty$, $k_\gamma(x, y) \simeq 1$ for any (x, y), this leads to $\mathbb{D}_{k_\gamma}(P, Q) \simeq 0$ for any P and Q. This case is not very useful as it does not "see" the difference between any probability distributions. On the other hand, when $\gamma \to 0$, assume that P and Q have densities p and q with respect to the Lebesgue measure respectively. Under suitable regularity and integrability assumptions on p and q, we have

$$\mathbb{E}_{X\sim P, Y\sim Q}[k_\gamma(X, Y)] \sim \pi^{\frac{d}{2}}\gamma^d \int p(x)q(x)\mathrm{d}x$$

when $\gamma \to 0$, and thus

$$\mathbb{D}_{k_\gamma}(P, Q) \sim \pi^{\frac{d}{4}}\gamma^{\frac{d}{2}}\|p - q\|_{L_2}.$$

Considering γ small enough, but not zero, will then allow to give a sense to L_2 estimation even for densities that are not in L_2. Of course, when p and q are not regular, these approximations can become wrong. Thus, the regularity of the model is something important in the link between MMD and L_2 distance. In order to make this discussion more formal, let us first introduce a measure of the distortion between the MMD and the L_2 distance.

Assumption 8.3.2. *Assume that* $\mathbb{X} = \mathbb{R}^d$ *and that* $k(x, y) = k_\gamma(\|x - y\|)$ *where* $k_\gamma(h) = k_1(h/\gamma)$ *and* $k_1 : \mathbb{R}_+ \to [0, 1]$.

When Assumption 8.3.2 is satisfied, we will use the notation $\mathbb{D}_{k_\gamma} = \mathbb{D}_k$.

Assumption 8.3.3. *For any $\theta \in \Theta$, P_θ has density $p_\theta \in L_2$ w.r.t the Lebesgue measure.*

Definition 8.3.1. *When Assumptions 8.3.2 and 8.3.3 are satisfied, we define:*

$$\mathcal{L}(\gamma) = \inf_{(\theta,\theta')\in\Theta} \frac{\mathbb{D}_{k_\gamma}(P_\theta, P_{\theta'})}{\|p_\theta - p_{\theta'}\|_{L^2}}, \text{ and}$$
$$\mathcal{U}(\gamma) = \sup_{(\theta,\theta')\in\Theta} \frac{\mathbb{D}_{k_\gamma}(P_\theta, P_{\theta'})}{\|p_\theta - p_{\theta'}\|_{L^2}}.$$

Assumption 8.3.4. *The true distribution P^0 has density $p^0 \in L_2$ w.r.t the Lebesgue measure.*

When all these assumptions are satisfied, the model and the true distribution are well-behaved and it possible to derive from MMD estimation a bound for L_2 estimation. The tightness of the bound will depend on the ratio $\mathcal{U}(\gamma)/\mathcal{L}(\gamma)$.

Theorem 8.3.5. *Under Assumptions 8.3.2, 8.3.3 and 8.3.4,*

$$\mathbb{E}\left[\|p_{\widehat{\theta}_n} - p^0\|_{L^2}\right] \leq \left(1 + \frac{2\mathcal{U}(\gamma)}{\mathcal{L}(\gamma)}\right) \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + \frac{2}{\mathcal{L}(\gamma)} \sqrt{\frac{1 + 2\sum_{t=1}^n \left(1 - \frac{t}{n}\right) \varrho_t}{n}}.$$

If moreover we assume that Assumptions 8.2.1 and 8.3.1 are satisfied,

$$\mathbb{P}\left[\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq \left(1 + \frac{2\mathcal{U}(\gamma)}{\mathcal{L}(\gamma)}\right) \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + 2\frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log(1/\delta)}}{\mathcal{L}(\gamma)\sqrt{n}}\right] \geq 1-\delta.$$

We end this subsection by a proposition that allows to upper bound $1/\mathcal{L}(\gamma)$ and $\mathcal{U}(\gamma)/\mathcal{L}(\gamma)$. We remind the definition of the Fourier transform of a function f:

$$\mathcal{F}[f](t) = \int f(x) \exp(-2i\pi \langle t, x\rangle) \mathrm{d}x.$$

Assumption 8.3.5. *The kernel K_1 is such that $\mathcal{F}[K_1](t) = \mu(t)$ for some function μ with*

1. *$D > \mu(t) > 0$ for any t,*
2. *$\int \mu(t)\mathrm{d}t = C$ for some $C > 0$,*
3. *there is an a such that $\|t\| \leq a \Rightarrow \mu(t) \geq b^2 > 0$.*

Example 8.3.1. *All these conditions are satisfied by the Gaussian kernel: $k_1(h) = \exp(-\|h\|^2)$ and $k_\gamma(h) = \exp(-\|h\|^2/\gamma^2)$. Then $k_1 \leq 1$ as required. Moreover: $\mathcal{F}[k_1](t) = \pi^{d/2}\exp(-\|t\|^2/4)$ and so we have 1) and 2) with $C = D = (2\pi)^{d/2}$ and 3) with $a = 1$ and $b = 1/\mathrm{e}$.*

Proposition 4. *Under 1. and 2. in Assumption 8.3.5,*

$$\mathcal{U}(\gamma) \leq D^{1/2}\gamma^{d/2}.$$

Under 3. in Assumption 8.3.5,

$$\mathcal{L}(\gamma) \geq b\gamma^{d/2}\mathcal{A}\left(\frac{a}{\gamma}\right)$$

where

$$\mathcal{A}(\xi) := \inf_{(\theta,\theta')\in\Theta^2} \sqrt{\frac{\int_{\|t\|\leq\xi} |\mathcal{F}[p_\theta - p_{\theta'}](t)|^2 \,\mathrm{d}t}{\int_{\mathbb{R}^d} |\mathcal{F}[p_\theta - p_{\theta'}](t)|^2 \,\mathrm{d}t}}.$$

Let us now discuss the consequences of Theorem 8.3.5 and Proposition 4 for L_2 density estimation, as well as the role of $\mathcal{A}$ and γ. For the sake of simplicity we stick to the bound in expectation in the discussion, but the the same comments apply to the bound in probability. First, pluging Proposition 4 into Theorem 8.3.5 gives:

$$\mathbb{E}\left(\|p_{\widehat{\theta}_n} - p^0\|_{L^2}\right) \leq \left(1 + \frac{2\sqrt{D}}{b\mathcal{A}\left(\frac{a}{\gamma}\right)}\right) \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + \frac{2}{b\gamma^{\frac{d}{2}}\mathcal{A}\left(\frac{a}{\gamma}\right)} \sqrt{\frac{1 + 2\sum_{t=1}^n \left(1 - \frac{t}{n}\right)\varrho_t}{n}} \tag{8.1}$$

and, in the well-specified case,

$$\mathbb{E}\left(\|p_{\widehat{\theta}_n} - p^0\|_{L^2}\right) \leq \frac{2}{b\gamma^{\frac{d}{2}}\mathcal{A}\left(\frac{a}{\gamma}\right)} \sqrt{\frac{1 + 2\sum_{t=1}^n \left(1 - \frac{t}{n}\right)\varrho_t}{n}}. \tag{8.2}$$

It is clear that the optimization with respect to γ might change the way the bound in (8.1) depends on d, but will not affect the way it depends on n. The first examples in Section 8.4 will clearly illustrate this fact. So, in small dimensions, one can always take $\gamma = 1$ to obtain the bound:

$$\mathbb{E}\left(\|p_{\widehat{\theta}_n} - p^0\|_{L^2}\right) \leq \left(1 + \frac{2\sqrt{D}}{b\mathcal{A}(a)}\right) \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + \frac{2}{b\mathcal{A}(a)} \sqrt{\frac{1 + 2\sum_{t=1}^n \left(1 - \frac{t}{n}\right)\varrho_t}{n}}.$$

However, in large dimension, the dependence on d matters. Note that the function $\mathcal{A}$ satisfies:

- $\mathcal{A}(0) = 0$,
- $\mathcal{A}$ is nondecreasing,
- $\mathcal{A}(\infty) = 1$.

For a fixed ξ, $\mathcal{A}(\xi)$ is the ratio between the energy in low frequencies and the whole energy of $p_\theta - p_{\theta'}$ which might exhibit different behaviors depending on the smoothness of $\theta \mapsto p_\theta$.

When $p_\theta - p_{\theta'}$ has enough energy in its low frequencies for any θ and θ', then one can expect $\mathcal{A}(\xi) \sim \xi^{d/2}$ for $\xi \to 0$. In this case, the function $\gamma^{d/2}\mathcal{A}(a/\gamma)$ would satisfy $\lim_{\gamma\to 0}\gamma^{d/2}\mathcal{A}(a/\gamma) = 0$ and $\lim_{\gamma\to+\infty}\gamma^{d/2}\mathcal{A}(a/\gamma) = a^{d/2}$, which means that even though the function might have a global minimum somewhere in between 0 and ∞, taking γ as large as possible cannot really hurt in (8.2) – even though it will make the first term in the right-hand side of (8.1) explode, which means that taking γ too large is unsafe in case of misspecification.

For nonsmooth models, one can however have $\mathcal{A}(\xi)/\xi^{d/2} \to 0$ for $\xi \to 0$. In this case, both $\gamma \to 0$ and $\gamma \to \infty$ will make the r.h.s explode both in (8.2) and (8.1). In this case, a careful optimization w.r.t γ is in order.

8.4 Examples

8.4.1 Independent observations

In this subsection, we focus on i.i.d observations. That is, $\varrho_t = 0$ for any $t \geq 1$. Moreover, we will only use the Gaussian kernel $k_\gamma(x, y) = \exp(-\|x - y\|^2/\gamma^2)$. Note that for this kernel, Proposition 4 gives $\mathcal{U}(\gamma) \leq (2\pi)^{1/4}\gamma^{1/2}$. We will use this bound in some situations, however, in the Gaussian model, we will see that it is possible to derive the explicit dependence between the L_2 norm and the MMD norm, thus avoiding Proposition 4. We assume that Assumption 8.3.4 is satisfied in this whole section.

Estimation of the mean in a Gaussian model

Here, $\mathbb{X} = \mathbb{R}^d$ and we are interested in the estimation of the mean in a Gaussian model. For the sake of simplicity, we assume that the variance is known. In this case, the proof of Proposition 5 below will show that we have explicit formulas, for any (θ, θ'), for $\|P_\theta - P_{\theta'}\|_{\mathcal{H}_k}$ and for $\|p_\theta - p_{\theta'}\|_{L^2}$, both as functions of $\|\theta - \theta'\|$. In particular, this leads to exact formulas

$$\mathcal{L}(\gamma) = \frac{4\sigma^2}{4\sigma^2+\gamma^2}\left(\frac{4\pi\sigma^2\gamma^2}{4\sigma^2+\gamma^2}\right)^{d/2} \text{ and } \mathcal{U}(\gamma) = \left(\frac{4\pi\sigma^2\gamma^2}{4\sigma^2+\gamma^2}\right)^{d/2}.$$

The complete proof is postponed to Section 8.7.

Proposition 5. *Assume that $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ for $\theta \in \Theta = \mathbb{R}^d$. Then, for any $\delta > 0$,*

$$\mathbb{P}\left[\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq \left(1 + \frac{4\sigma^2+\gamma^2}{2\sigma^2}\right)\inf_{\theta\in\Theta}\|p_\theta - p^0\|_{L^2} + \frac{4\sigma^2+\gamma^2}{2\sigma^2}\left(\frac{4\sigma^2+\gamma^2}{4\pi\sigma^2\gamma^2}\right)^{d/2}\frac{2+2\sqrt{2\log(1/\delta)}}{\sqrt{n}}\right] \geq 1-\delta. \quad (8.3)$$

Moreover, the second term in the upper bound is minimized for $\gamma^2 = 2d\sigma^2$ *which leads to*

$$\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq (3+d) \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + \frac{\mathrm{e}(d+2)}{(4\pi\sigma^2)^{\frac{d}{2}}} \frac{2 + 2\sqrt{2\log(1/\delta)}}{\sqrt{n}}, \tag{8.4}$$

still with probability $1-\delta$*. Finally, assume that we are in an adversarial contamination model where a proportion at most* ϵ *of the observations is contaminated, then, with probability* $1-\delta$*,*

$$\|\widetilde{\theta}_n - \theta_0\|^2 \leq -2\sigma^2(d+2)\log\left[1 - 8\mathrm{e}\left(\epsilon + \frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2\right]. \tag{8.5}$$

Note that when ϵ is small and n is large,

$$\begin{aligned}\|\widetilde{\theta}_n - \theta_0\|^2 &\leq -2\sigma^2(d+2)\log\left[1 - 16\mathrm{e}\left(\epsilon^2 + \frac{\left(1+\sqrt{2\log\left(\frac{1}{\delta}\right)}\right)^2}{n}\right)\right] \\ &\sim 32\mathrm{e}\sigma^2(d+2)\left(\epsilon^2 + \frac{\left(1+\sqrt{2\log\left(\frac{1}{\delta}\right)}\right)^2}{n}\right).\end{aligned}$$

According to Theorems 2.1 and 2.2 in Chen et al. (2018), the minimax rate with respect d, ε and n is $\varepsilon^2 + d/n$. Hence, we obtain a convergence rate $d\varepsilon^2 + d/n$ that achieves a quadratic dependence in ε, contrary to most popular robust estimators such as Median-of-Means which dependence in ε is linear. Note that the rate of convergence we obtain is the one achieved by the geometric median.

Cauchy model

Here, $\mathbb{X} = \mathbb{R}$ and $P_\theta = \mathcal{C}(\theta, 1)$ where $\mathcal{C}(\theta, s)$ has density $1/[\pi s(1+(x-\theta)^2/s^2)]$. This time, we use the generic upper bound $\mathcal{U}(\gamma) \leq (2\pi)^{1/4}\gamma^{1/2}$ and prove a lower bound $\mathcal{L}(\gamma) \geq 1/3$ for $\gamma = 2$ thanks to Proposition 4. We obtain the following result.

Proposition 6. *Assume that* $P_\theta = \mathcal{C}(\theta, 1)$ *for* $\theta \in \Theta = \mathbb{R}$*. Then, taking* $\gamma = 2$ *leads to, for any* $\delta > 0$*,*

$$\mathbb{P}\left[\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq 14 \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + \frac{6 + 6\sqrt{2\log(1/\delta)}}{\sqrt{n}}\right] \geq 1 - \delta.$$

Moreover, assume that we are in an adversarial contamination model where a proportion at most ϵ *of the observations is contaminated, then, with probability* $1-\delta$*,*

$$(\widetilde{\theta}_n - \theta_0)^2 \leq 4\left(1 - \frac{1}{1 - 96\pi\left(\epsilon^2 + \frac{2+4\log(1/\delta)}{n}\right)}\right).$$

Note that

$$(\widetilde{\theta}_n - \theta_0)^2 \leq 4\left(1 - \frac{1}{1 - 96\pi\left(\epsilon^2 + \frac{2+4\log(1/\delta)}{n}\right)}\right) \sim 384\pi\left(\epsilon^2 + \frac{2+4\log(1/\delta)}{n}\right).$$

Again, we achieve the optimal quadratic dependence in ε.

Uniform model

Here, $\mathbb{X} = \mathbb{R}$ and $P_\theta = \mathcal{U}[\theta - 1/2, \theta + 1/2]$.

Proposition 7. *Assume that $P_\theta = \mathcal{U}[\theta - 1/2, \theta + 1/2]$, $\theta \in \Theta = \mathbb{R}$ Then, taking $\gamma = 2$ leads to, for any $\delta > 0$,*

$$\mathbb{P}\left[\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq 23.4 \inf_{\theta\in\Theta} \|p_\theta - p^0\|_{L^2} + \frac{10 + 10\sqrt{2\log(1/\delta)}}{\sqrt{n}}\right] \geq 1 - \delta.$$

Note that the proof shows that in the well specified case $p^0 = p_{\theta_0}$, $\|p_{\widehat{\theta}_n} - p^0\|_{L^2} = \min(1, |\widehat{\theta}_n - \theta_0|)$. Thus, for n large enough to ensure that the bound is smaller than 1, Proposition 7 states that, with probability at least $1 - \delta$,

$$|\widehat{\theta}_n - \theta_0| \leq \frac{10 + 10\sqrt{2\log(1/\delta)}}{\sqrt{n}}.$$

Note that in this model, the moment estimator reaches the rate $1/\sqrt{n}$ but the MLE reaches the rate $1/n$. In practice, we indeed observe in the simulations that for $\gamma \sim 1$, the MMD estimator is "as bad" as the moment estimator. However, on the contrary to MLE and moment estimators, it is highly robust to the presence of outliers.

Moreover, for $\gamma \to 0$, we observe that the MMD estimator becomes as good as the MLE in the nice situation (correct specification, no outliers). We were not able to explain this with our theoretical analysis and leave it to future works.

Estimation with a dictionary

We consider here estimation of the density as a linear combination of given functions in a dictionary. This framework actually appears in various models:

- first, when the dictionary contains densities, this is simply a mixture of known components. In this case, the linear combination is actually a convex combination. This context is for example studied in Dalalyan and Sebbar (2017).
- in nonparametric density estimation, we can use this setting, the dictionary being a basis of L_2. This is for example the point of view in Alquier (2008a); Bunea et al. (2007, 2010).

We will here focus on the first setting, but an extension to the second one is quite straightforward. Let $\{\Phi_1, \ldots, \Phi_d\}$ be a family of probability measures over $\mathbb{X} = \mathbb{R}^D$. For $1 \leq i \leq d$ we remind that

$$\mu_{\Phi_i}(\cdot) = \int k(x, \cdot)\Phi_i(\mathrm{d}x).$$

Define the measure $P_\theta = \mathcal{D}(\theta; \Phi_1, \ldots, \Phi_d) = \sum_{i=1}^d \theta_i \Phi_i$ with respect to the Lebesgue measure, and we define the model $\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$. We could consider $\Theta = \mathbb{R}^d$ in a general framework, but as we only study the mixture case, we assume that $\Theta \subseteq \mathcal{S}_d = \{\theta \in \mathbb{R}_+^d : \sum_{i=1}^d \theta_i = 1\}$. Note that in the fisrt case, most P_θ's are not probability measures, but this is in accordance with our definition of a statistical model. The estimator is then

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \left\| \sum_{\ell=1}^d \theta_\ell \mu_{\Phi_\ell}(\cdot) - \mu_{\widehat{P}_n} \right\|_{\mathcal{H}_k}^2.$$

Assuming that each Φ_i has a density with respect to the Lebesgue measure $\phi_i \in L^2$, each P_θ has a density $p_\theta = \sum_{i=1}^d \theta_i \phi_i$ and we have:

$$\frac{\mathbb{D}_{k_\gamma}(P_\theta, P_{\theta'})}{\|p_\theta - p_{\theta'}\|_{L_2}} = \frac{\sum_{1 \leq i,j \leq n} (\theta_i - \theta_i')(\theta_j - \theta_j') \left\langle \mu_{\Phi_i}, \mu_{\Phi_j} \right\rangle_{\mathcal{H}_k}}{\sum_{1 \leq i,j \leq n} (\theta_i - \theta_i')(\theta_j - \theta_j') \left\langle \mu_{\Phi_i}, \mu_{\Phi_j} \right\rangle_{L_2}}.$$

This immediately leads to the following result.

Proposition 8. *Assume that $P_\theta = \sum_{i=1} \theta_i \Phi_i$ where Φ_i has density $\phi_i \in L_2$, and let p_θ denote the density of P_θ. Define the matrices $G = \left(\left\langle \mu_{\Phi_i}, \mu_{\Phi_j} \right\rangle_{L_2} \right)$ and $G_\gamma = \left\langle \mu_{\Phi_i}, \mu_{\Phi_j} \right\rangle_{\mathcal{H}_{k_\gamma}}$. Letting $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote respectively the smallest and largest eigenvalue of a symmetric matrix, we have:*

$$\mathcal{U}(\gamma) = \frac{\lambda_{\max}(G_\gamma)}{\lambda_{\min}(G)} \text{ and } \mathcal{L}(\gamma) = \frac{\lambda_{\min}(G_\gamma)}{\lambda_{\max}(G)}.$$

Let $C(\cdot) = \lambda_{\max}(\cdot)/\lambda_{\min}(\cdot)$ denote the condition number of a matrix. Then

$$\mathbb{P}\left[\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq \left[1 + 2C(G)C(G_\gamma)\right] \inf_{\theta \in \Theta} \|p_\theta - p^0\|_{L^2} + \frac{2\lambda_{\max}(G)}{\lambda_{\min}(G_\gamma)} \frac{1 + 1\sqrt{2\log(1/\delta)}}{\sqrt{n}} \right] \geq 1 - \delta.$$

8.4.2 β-mixing observations

We now consider non-independent random variables: as in the general framework presented above, $(X_t)_{t \in \mathbb{Z}}$ is a strictly stationary time series, with stationary distribution P^0, and that we observe $X_1, \ldots, X_n$. We will exhibit some condition on the dependence of the X_i's ensuring that we can still estimate P^0 with the MMD method.

There is a very rich literature on limit theorems and exponential inequalities under conditions on various dependence coefficients. Mixing coefficients and their applications are detailed in the monographs Doukhan (1994); Rio (2017a), weak dependence coefficients in Dedecker et al. (2007). In this subsection, we show that our coefficient ϱ_t can

be upper-bounded by the β-mixing coefficients. So for any β-mixing process, the estimation of P^0 using MMD remains possible. We also remind some examples of β-mixing processes. Note that we will show in the next subsection that Theorem 8.3.1 can be successfully applied to non β-mixing processes.

We start by a reminder of the defintion of the β-mixing coefficients, from page 4 (Chapter 1) in Dedecker et al. (2007).

Definition 8.4.1. *Given two sigma algebras $\mathcal{A}$ and $\mathcal{B}$,*

$$\beta(\mathcal{A},\mathcal{B}) = \frac{1}{2} \sup_{\substack{I,\,J\geq 1 \\ U_1,\ldots,U_I \\ V_1,\ldots,V_J}} \sum_{1\leq i\leq I} \sum_{1\leq j\leq J} |\mathbb{P}(U_i\cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j)|$$

where $(U_1,\ldots,U_I)$ is any partition of $\mathcal{A}$ and $V_1,\ldots,V_j$ any partition of $\mathcal{B}$. Put:

$$\beta_t^{(X)} = \beta(\sigma(X_0,X_{-1},\ldots),\sigma(X_t,X_{t+1},\ldots)).$$

Section 1.5 in Doukhan (1994) provides summability conditions on the $\beta_t^{(X)}$ leading to a law of large numbers and to a central limit theorem. Examples are also discussed.

Example 8.4.1. *Assume in this example that (X_t) is an homogeneous Markov chain given by its transition kernel P and $X_0\sim\pi$ where $\pi P=\pi$. Assume that there is a $0<c\leq 1$ and a probability measure Q on $\mathbb{R}^d$ such that, for some integer $r\geq 1$ and for any measurable A, $P^r(x,A)\geq cQ(A)$. Then it is known, see e.g. Theorem 1 page 88 in Doukhan (1994) that*

$$\beta_t^{(X)} \leq 2(1-c)^{\frac{t}{r}-1}.$$

We now compare our ϱ coefficients with the β-mixing coefficients.

Proposition 9. *Assume that $k(x,y)=F(\|x-y\|)$ were $F(a)=\int_a^\infty f(b)\mathrm{d}b$ for some nonnegative continuous function f with $\int_0^\infty f(b)\mathrm{d}b=1$. Then we have*

$$\varrho_t \leq 2\beta(\sigma(X_0),\sigma(X_t)) \leq 2\beta_t.$$

Note that $k(x,y)=\exp(-\|x-y\|/\gamma)$ and $k(x,y)=\exp(-\|x-y\|^2/\gamma^2)$ for example trivially work, respectively with $f(b)=\exp(-b/\gamma)/\gamma$ and $f(b)=2b\exp(-b^2/\gamma^2)/\gamma^2$.

Hidden Markov chains

Assume here that $(Y_t)_{t\in\mathbb{N}}$ is a Markov chain on $\{1,\ldots,d\}$, and that $X_t|(Y_t=i)$ is independent from all the other values $Y_{t'}$ and is drawn in $\mathbb{R}^D$ from a probability measure Φ_i. The Φ_i's are known and $X_1,\ldots,X_n$ are observed but the $(Y_t)_{t\in\mathbb{N}}$ are not observed. Note that this is a dependend extension of the misture model $\mathcal{D}(\theta;\Phi_1,\ldots,\Phi_d)$ discussed above. Indeed, we consider this as a case of misspecification: the statistician uses the mixture model $\mathcal{D}(\theta;\Phi_1,\ldots,\Phi_d)$ with $\Theta=\mathcal{S}_d$, being not aware that the data is actually not independent.

Letting P denote the transition matrix of Y, we assume that there exists $c > 0$ and an integer $r \geq 0$ such that $P^r(i,j) \geq c/d$ for any $(i,j) \in \{1,\dots,d\}^2$. Then we have $\beta_t^{(Y)} \leq 2(1-c)^{t/r-1}$. This also implies that there is a unique π such that $\pi P = \pi$ and we assume that $Y_0 \sim \pi$. Then the distribution P^0 of each X_t is given by $P^0(x) = \sum_{i=1}^d \pi_i \Phi_i(x)$.

Also, note that

$$\varrho_t = \beta(\sigma(X_0), \sigma(X_t)) \leq \beta(\sigma(X_0, Y_0), \sigma(X_t, Y_t)) = \beta(\sigma(Y_0), \sigma(Y_t)) \leq 2(1-c)^{t/r-1}.$$

So, a direct application of Theorem 8.3.1 gives:

$$\mathbb{E}\left[\|G^{-1/2}(\widehat{\theta} - \pi)\|\right] = \mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right)\right] \leq 2\sqrt{\frac{1 + (1-c)^{\frac{1}{r}-1}(3+c)}{n[1-(1-c)^{\frac{1}{r}}]}}.$$

Note that we can add a second layer in the process: assume that an opponent is allowed to replace a fraction ϵ of the X_t, as in Proposition 3. This result in the observation of $\widetilde{X}_t$ such that $\widetilde{X}_t = X_t$ for a proportion $(1-\epsilon)$ of the data, and $\widetilde{X}_t$ can be anything for the remaining ε. For example, the opponent can try fo fool the learner, by drawing from the wrong Φ_i. The MMD estimator $\widetilde{\theta}$ still satisfies, from Proposition 3,

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right)\right] \leq 4\epsilon + 4\sqrt{\frac{1 + (1-c)^{\frac{1}{r}-1}(3+c)}{n[1-(1-c)^{\frac{1}{r}}]}}.$$

8.4.3 Non-mixing processes

In this subsection, we provide an example of non-mixing process, with $\beta_t = 1/4$ and so $\sum_{t=1}^\infty \beta_t = \infty$, such that $\sum_{t=1}^\infty \varrho_t < \infty$. We then provide statistical application.

Examples of non-mixing processes with $\sum_t \varrho_t < \infty$

First, we remind a classical example of non-mixing process, in the sense that $\sum_{t=1}^\infty \beta_t = \infty$. See for example Section 1.5 page 8 in Dedecker et al. (2007) where it is also proven that it is neither α-mixing. The process id defined by $X_{t+1} = X_t/2 + \varepsilon_{t+1}$, where the ε_t are i.i.d $\mathcal{B}e(1/2)$ and $X_0 \sim \mathcal{U}([0,1])$. As for any t, $X_t = f(X_{t+1})$ where f is the measurable function $f(x) = 2x - \lfloor 2x \rfloor$, it is possible to take $I = J = 2$, $V_1 = U_1$ and $V_2 = U_2 = U_1^c$ for some U_1 with $\mathbb{P}(U_1) = 1/2$ in Definition 8.4.1. This leads to $\beta(\sigma(X_0), \sigma(X_t)) \geq 1/4$.

However, the ϱ_t will decay exponentially. This is a consequence of the more general following proposition.

Proposition 10. *Assume that $k(x,y) = F(\|x-y\|)$ where F is an L-Lipschitz function and assume that X_k can be written as $X_k = G_k(X_0, B_k)$ where B_k is independent of X_0 and G_k is L_k-Lipschitz in its first component: $\|G_k(x,b) - G_k(x',b)\| \leq L_k\|x-x'\|$. Then $\varrho_k \leq 2LL_k\mathbb{E}(\|X_0\|)$.*

In the previous example, $G_k(X_0, B_k) = X_0/2^k + B_k$ with $B_k = \varepsilon_k + \varepsilon_{k-1}/2 + \cdots + \varepsilon_1/2^{k-1}$ is indeed independent of X_0. So G_k is L_k-Lipschitz in x with $L_k = 1/2^k$. Moreover, as

$X_t \sim \mathcal{U}([0,1])$ for any t, $\mathbb{E}(|X_0|) = \mathbb{E}(X_0) = 1/2$. So

$$\varrho_k \leq 2LL_k\mathbb{E}(|X_0|) = \frac{L}{2^k}.$$

With a Gaussian kernel $k(x,y) = \exp(-|x-y|^2/\gamma^2)$ one has $L = 2/[\exp(1/\gamma)\gamma]$ and so

$$\varrho_k = \frac{1}{\gamma\exp(1/\gamma)2^k}.$$

Another classical example of non-mixing process is a reversed version of the previous one. We draw $X_0 \sim \mathcal{U}([0,1])$ and simply define $X_{t+1} = f(X_t)$ where we still have $f(x) = 2x - \lfloor 2x \rfloor$. Note that apart from X_0, the process is entirely deterministic, and thus non-mixing. Properties of (generalized versions) of such processes are studied in Section 3.3 page 28 in Dedecker et al. (2007). Still, following step by step the proof of Proposition 10, we can show that if X_0 can be written as $X_0 = G_k'(X_k, B_k)$ where B_k is independent of X_k and G_k' is L_k'-Lipschitz in its first component, then we still have $\varrho_k \leq 2LL_k\mathbb{E}(\|X_0\|)$. Thus, this process also satisfies $\varrho_k \leq L/2^k$.

8.5 Stochastic gradient algorithm for MMD estimation

In this section, we briefly discuss gradient-based algorithms to compute the estimator $\widehat{\theta}_n$ when $\Theta \subset \mathbb{R}^d$. In Subsection 8.5.1 we provide an expression of the gradient of the criterion to be minimized. We briefly provide a special case where this gradient can be computed explicitly. However, in general, this is not the case, but we can provide unbiased estimators of this gradient as soon as we are able to sample from P_θ, in this case the model is often refered to as a *generative model*. Thus it is possible to use a stochastic gradient algorithm when $\{P_\theta, \theta \in \Theta\}$ is a generative model. We describe this algorithm in Subsection 8.5.2, and remind its theoretical properties in Subsection 8.5.3.

Note that the idea to use a stochastic gradient algorithm to compute $\widehat{\theta}_n$ was first used to train a generative neural network by Dziugaite et al. (2015). In Briol et al. (2019) the authors propose to use a stochastic natural gradient algorithm instead. By providing adaptation to the geometry of the problem, the natural gradient will lead to better results but increase the computational burden when the dimension of the problem is large.

8.5.1 Gradient of the MMD distance

We remind that in this whole section, $\Theta \subset \mathbb{R}^d$. To compute $\widehat{\theta}_n$, one must minimize, with respect to $\theta \in \Theta$,

$$\mathbb{D}_k(P_\theta, \widehat{P}_n) = \mathbb{E}_{X,X'\sim P_\theta}[k(X,X')] - \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{X\sim P_\theta}[k(X_i,X)] + \frac{1}{n^2}\sum_{1\leq i,j\leq n} k(X_i,X_j)$$

or, equivalently,

$$\mathrm{Crit}(\theta) = \mathbb{E}_{X,X' \sim P_\theta}[k(X, X')] - \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}_{X \sim P_\theta}[k(X_i, X)].$$

In order to use gradient algorithms or any first order method, a first step is to compute the gradient of this quantity with respect to θ.

Proposition 11. *Assume that, for any x, $\theta \mapsto p_\theta(x)$ is differentiable with respect to θ and that there is a nonnegative function $g(x, x')$ such that, for any $\theta \in \Theta$, $|k(x, x')\nabla_\theta[p_\theta(x)p_\theta(x')]| \leq g(x, x')$ and $\iint g(x, x')\mu(\mathrm{d}x)\mu(\mathrm{d}x') < \infty$. Then*

$$\nabla_\theta \mathrm{Crit}(\theta) = 2\mathbb{E}_{X,X' \sim P_\theta}\left[\left(k(X, X') - \frac{1}{n}\sum_{i=1}^{n} k(X_i, X)\right) \nabla_\theta[\log p_\theta(X)]\right].$$

Note that the gradient of $\mathrm{Crit}(\theta)$ is given by an expectation with respect to P_θ. So, as soon as it is feasible to sample from P_θ, on can provide unbiased estimates of $\nabla_\theta \mathrm{Crit}(\theta)$, and thus implement a stochastic gradient algorithm.

Remark 8.5.1. *It might be that in special cases, we have explicit formulas for the expectations in $\mathrm{Crit}(\theta)$ and its gradient. For example, assume that we are a translation parameter, that is: $p_\theta(x) = f(x - \theta)$ for some density f, and that the kernel k is given by $k(x, x') = K(x - x')$ for some function K. Then*

$$\begin{aligned}\mathrm{Crit}(\theta) &= \iint K(x - x')f(x - \theta)f(x' - \theta)\mu(\mathrm{d}x)\mu(\mathrm{d}x') - \frac{2}{n}\sum_{i=1}^{n}\int K(X_i - x)f(x - \theta)\mu(\mathrm{d}x) \\ &= \iint K(x - x')f(x)f(x')\mu(\mathrm{d}x)\mu(\mathrm{d}x') - \frac{2}{n}\sum_{i=1}^{n}\int K(\theta + x - X_i)f(x)\mu(\mathrm{d}x).\end{aligned}$$

For example, in the case $P_\theta = \mathcal{U}[\theta - 1/2, \theta + 1/2]$ we have

$$\begin{aligned}\mathrm{Crit}(\theta) &= \iint_{[-1/2,1/2]^2} K(x - x')\mathrm{d}x\mathrm{d}x' - \frac{2}{n}\sum_{i=1}^{n}\int_{-1/2}^{1/2} K(\theta + x - X_i)\mathrm{d}x \\ &= \iint_{[-1/2,1/2]^2} K(x - x')\mathrm{d}x\mathrm{d}x' - \frac{2}{n}\sum_{i=1}^{n}\int_{\theta-1/2-X_i}^{\theta+1/2-X_i} K(u)\mathrm{d}u\end{aligned}$$

and thus

$$\nabla_\theta \mathrm{Crit}(\theta) = -\frac{2}{n}\sum_{i=1}^{n}[K(\theta + 1/2 - X_i) - K(\theta - 1/2 - X_i)].$$

So, in this special case, the estimation of the gradient is unnecessary and we can use a gradient algorithm to compute $\widehat{\theta}_n$.

8.5.2 Projected stochastic gradient algorithm for the MMD estimator

From Proposition 11,

$$\nabla_\theta \mathrm{Crit}(\theta) = 2\mathbb{E}_{X,X' \sim P_\theta}\left[\left(k(X, X') - \frac{1}{n}\sum_{i=1}^{n} k(X_i, X)\right) \nabla_\theta[\log p_\theta(X)]\right].$$

So, if we can compute $\nabla[\log p_\theta(x)]$ and if it is feasible to simulate from P_θ, we can easily compute a Monte Carlo estimator of $\nabla_\theta \mathrm{Crit}(\theta)$ and thus use a stochastic gradient descent (SGD). First, simulate $(Y_1, \ldots, Y_M)$ i.i.d from P_θ, then put

$$\widehat{\nabla_\theta \mathrm{Crit}}(\theta) = \frac{2}{M} \sum_{j=1}^{M} \left(\frac{1}{M-1} \sum_{\ell \neq j} k(Y_j, Y_\ell) - \frac{1}{n} \sum_{i=1}^{n} k(X_i, Y_j) \right) \nabla_\theta [\log p_\theta(Y_j)].$$

We now provide the details of a projected stochastic gradient algorithm (PSGA). The projection step is necessary if $\Theta \subsetneq \mathbb{R}^d$. Thus, we assume that $\Theta \subset \mathbb{R}^d$ is a closed and convex subset and let Π_Θ denote the orthogonal projection on Θ.

Algorithm 12 PSGA for MMD

Require: A dataset $(X_1, ..., X_n)$, a model $(P_\theta, \theta \in \Theta \subset \mathbb{R}^d)$ a kernel k, a sequence of steps $(\eta_t)_{t \geq 1}$, an integer M, a stopping time T, an initial point $\theta^{(0)} \in \Theta$.
$X_0 \sim \nu_0$
for $t = 1, \cdots, T$ **do**
draw $(Y_1, \ldots, Y_M)$ i.i.d from $P_{\theta^{(t-1)}}$
$\theta^{(t)} = \Pi_\Theta \left\{ \theta^{(t-1)} - \frac{2\eta_t}{M} \sum_{j=1}^{M} \left[\frac{1}{M-1} \sum_{\ell \neq j} k(Y_j, Y_\ell) - \frac{1}{n} \sum_{i=1}^{n} k(X_i, Y_j) \right] \nabla_{\theta^{(t-1)}} [\log p_{\theta^{(t-1)}}(Y_j)] \right\}$
end for

8.5.3 Theoretical analysis of the algorithm

In its original version, the stochastic gradient algorithm was proposed with a sequence of steps $(\eta)_t$ such that $\eta_t \to 0$ and $\sum_t \eta_t = \infty$. However, Nemirovski et al. (2009) proved that the method can be made more robust by taking a constant step size $\eta_t = \eta$ and by averaging the parameters. The following proposition is actually a direct application of the results of Nemirovski et al. (2009).

Proposition 12. *Under the conditions of Proposition (11) above, and under the assumption that Θ is closed, convex and bounded with $D = \sup_{(\theta,\theta') \in \Theta^2} \|\theta - \theta'\|$, define*

$$\widehat{\theta}_n^{(T)} = \frac{1}{T} \sum_{t=1}^{T} \theta^{(t)}$$

where the $\theta^{(t)}$'s are given by Algorithm 1 above. Assume that, for any $\theta \in \Theta$,

$$\mathbb{E}\left[\|\widehat{\nabla_\theta \mathrm{Crit}}(\theta)\|^2 \right] \leq M^2.$$

Assume that $\mathrm{Crit}(\theta)$ is a convex function of θ. Then the choice $\eta = D/(M\sqrt{T})$ leads to

$$\mathbb{E}\left[\mathrm{Crit}(\widehat{\theta}_n^{(T)}) - \mathrm{Crit}(\widehat{\theta}_n) \right] \leq \frac{DM}{\sqrt{T}}, \tag{8.6}$$

where the expectation $\mathbb{E}$ is taken with respect to drawings of the Y_i's in Algorithm 1. Moreover

$$\mathbb{E}\left[\mathbb{D}_k \left(P_{\widehat{\theta}_n^{(T)}}, P^0 \right) \right] \leq \inf_{\theta \in \Theta} \mathbb{D}_k(P_\theta, P^0) + 3\sqrt{\frac{1 + 2\sum_{t=1}^{n} \varrho_t}{n}} + \sqrt{\frac{DM}{\sqrt{T}}}$$

where the expectation is taken with repect to the sample and to the Y_i's, and the choice $T = n^2$ leads to

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n^{(n^2)}}, P^0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P^0) + \frac{\sqrt{DM} + 3\sqrt{1 + 2\sum_{t=1}^n \varrho_t}}{n}.$$

The restrictive assumption in this proposition is the convexity assumption on the criterion. However, it is satisfied in some of the examples of Section 8.4.

Example 8.5.1. *Let us come back to the "estimation with a dictionary" example of Section 8.4: P_θ is given by its density*

$$p_\theta = \sum_{\ell=1}^d \theta_\ell \Phi_\ell.$$

As mentioned in Section 8.4, if θ is unrestricted ($\Theta = \mathbb{R}^d$) we cannot apply Proposition 12 but there is an explicit formula for $\widehat{\theta}_n$. Now, let us assume that $\Theta = \mathcal{S}_d$ and the Φ_ℓ's are probability densities (this is the mixture of densities case). Then Θ closed, convex and bounded with $D = 1$. Moreover,

$$\widehat{\nabla_\theta \mathrm{Crit}}(\theta) = \frac{2}{M}\sum_{j=1}^M \left[\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j, Y_\ell) - \frac{1}{n}\sum_{i=1}^n k(X_i, Y_j)\right] \nabla_\theta[\log p_\theta(Y_j)].$$

and

$$\nabla_\theta[\log p_\theta(Y_j)] = \begin{pmatrix} \frac{\Phi_1(Y_j)}{\sum_{\ell=1}^d \theta_\ell \Phi_\ell(Y_j)} \\ \vdots \\ \frac{\Phi_d(Y_j)}{\sum_{\ell=1}^d \theta_\ell \Phi_\ell(Y_j)} \end{pmatrix}.$$

Consequently,

$$\begin{aligned}
\left\|\widehat{\nabla_\theta \mathrm{Crit}}(\theta)\right\|^2 &= \sum_{\ell=1}^d \left(\frac{2}{M}\sum_{j=1}^M \left[\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j, Y_\ell) - \frac{1}{n}\sum_{i=1}^n k(X_i, Y_j)\right] \frac{\Phi_\ell(Y_j)}{\sum_{\ell=1}^d \theta_\ell \Phi_\ell(Y_j)}\right)^2 \\
&\leq \sum_{\ell=1}^d \frac{4}{M^2} \sum_{1\leq j,k\leq M} \frac{\Phi_\ell(Y_j)\Phi_\ell(Y_k)}{\left(\sum_{\ell=1}^d \theta_\ell \Phi_\ell(Y_j)\right)\left(\sum_{\ell=1}^d \theta_\ell \Phi_\ell(Y_k)\right)} \\
&= \frac{4}{M^2} \sum_{1\leq j,k\leq M} \frac{\sum_{\ell=1}^d \Phi_\ell(Y_j)\Phi_\ell(Y_k)}{p_\theta(Y_j)p_\theta(Y_k)},
\end{aligned}$$

and then

$$\begin{aligned}
\mathbb{E}\left(\left\|\widehat{\nabla_\theta \mathrm{Crit}}(\theta)\right\|^2\right) &\leq \iint 4\frac{\sum_{\ell=1}^d \Phi_\ell(y)\Phi_\ell(y')}{p_\theta(y)p_\theta(y')} p_\theta(y)p_\theta(y')\mathrm{d}y\mathrm{d}y' \\
&= 4\iint \sum_{\ell=1}^d \Phi_\ell(y)\Phi_\ell(y')\mathrm{d}y\mathrm{d}y' \\
&= 4d.
\end{aligned}$$

Hence Proposition 12 leads to

$$\mathbb{E}\left[\mathrm{Crit}(\widehat{\theta}_n^{(T)}) - \mathrm{Crit}(\widehat{\theta}_n)\right] \leq \sqrt{\frac{4d}{T}} = 2\sqrt{\frac{d}{T}}.$$

8.6 Simulation study

In this section, we test our stochastic gradient algorithm on several synthetic datasets composed of $n = 200$ datapoints that were generated independently using four different distributions: a uni- and multidimensional univariate Gaussian, a uniform, a Cauchy, and a Gaussian mixture. All datasets are corrupted by outliers whose proportion ranges from 0 to 0.20 with a step-size of 0.025 in the experiments. We chose a number of Monte-Carlo samples equal to n and a step-size of $\eta_t = 1/\sqrt{t}$, and we used the Gaussian kernel $k(x, y) = e^{-\|x-y\|_2^2/d}$ where d is the dimension. Each experiment is repeated 100 times.

Gaussian mean estimation: First, we estimate the mean of a Gaussian distribution $\mathcal{N}(\theta, I_d)$ where I_d is the identity matrix of dimension d and where θ is the vector with all components equal to 2. All the outliers are generated using a standard Cauchy distribution $\mathcal{C}(0, 1)$ independently for each component. The MMD gradient descent is compared with the componentwise median (MED) and the maximum likelihood estimator (MLE) which is here the arithmetic mean. The metric considered here is the square root of the mean square error (MSE) over all the 100 repetitions. We can see in the two plots below that our algorithm achieves the smallest error as the proportion of outliers grows, clearly outperforming the MLE and being comparable to the componentwise median, and grows linearly as the ratio of outliers increases.

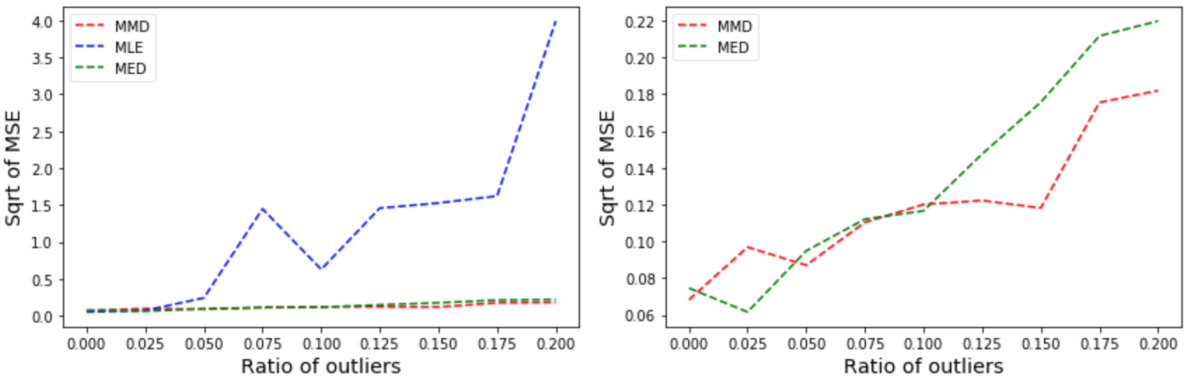


Figure 8.1: Comparison of the square root of the MSE for the MMD estimator, the MLE and the componentwise median (MED) in the robust unidimensional Gaussian mean estimation problem for several values of the proportion of outliers.

Uniform location parameter estimation: Then we estimate the location parameter of a uniform distribution $[\theta - 1/2, \theta + 1/2]$ with $\theta = 1$. Outliers are generated from a Cauchy distribution $\mathcal{C}(0, 1)$ with a location parameter equal to 1. We compare the mean of the variational approximation with the MLE (i.e the average between the largest and the lowest values) and the method of moments estimator (i.e the arithmetic mean). We use again the square root of the MSE over the 200 repetitions as the metric. Figure 3 clearly shows that the MMD estimator is the best estimator and is not affected by a reasonable proportion of outliers, contrary to the method of moments which square root of MSE is increasing linearly with the proportion of outliers and to the MLE that fails as soon as there is one outlier in the data.

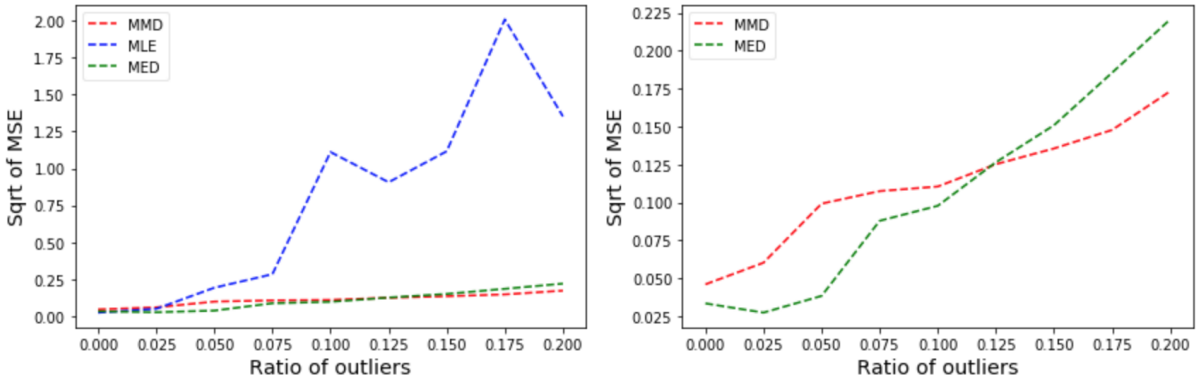


Figure 8.2: Comparison of the square root of the MSE for the MMD estimator, the MLE and the componentwise median (MED) in the robust 20-dimensional Gaussian mean estimation problem for several values of the proportion of outliers.

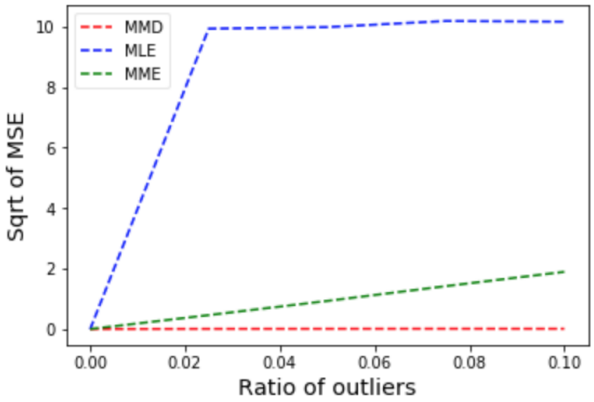


Figure 8.3: Comparison of the square root of the MSE for the MMD estimator, the MLE and the method of moments in the robust estimation of the location parameter of a uniform distribution for several values of the proportion of outliers.

Cauchy estimation: We also estimate the location parameter of a Cauchy $\mathcal{C}(\theta, 1)$ where $\theta = 2$. We corrupt the data using a standard Cauchy distribution, and we multiply this noise by 2. Note that the theoretical mean of a $\mathcal{C}(2, 1)$ is not defined and that its theoretical median is equal to $\theta = 2$. The estimators we will use here to be compared with the MMD procedure are the arithmetic mean and the geometric median. We still consider the square root of the MSE. The plots in Figure 4 show similar results to those obtained for the Gaussian mean estimation problem, with an unstable mean estimator and comparable median and MMD estimators.

Gaussian mixture estimation: In the last experiment, we sample data according to a three component Gaussian mixture $0.3.\mathcal{N}(-3.72, 1) + 0.3.\mathcal{N}(0.11, 1) + 0.4.\mathcal{N}(4.54, 1)$. Here, we use the same approach than in Section 8.4.1. We try to estimate the mixture as a linear combination of mixture in a dictionary composed of all Gaussians of variance 1 and whose means range from -5 to 5 with a stepsize of 0.02. Note that the Gaussian $\mathcal{N}(0.11, 1)$ is not even in the dictionary. The goal is to estimate the weights of each Gaussian in the dictionary. This estimation method is compared to the gold

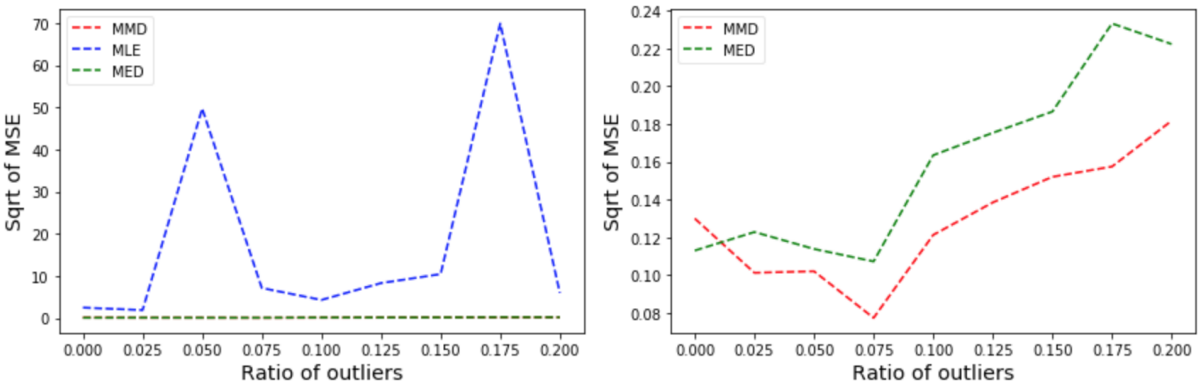


Figure 8.4: Comparison of the square root of the MSE for the MMD estimator, the arithmetic mean and the componentwise median in the robust estimation of the location parameter of a Cauchy distribution for several values of the proportion of outliers.

standard Expectation-Maximization (EM) (Dempster et al., 1977) algorithm and to the tempered Coordinate Ascent Variational Inference (CAVI) algorithm (Chérief-Abdellatif and Alquier, 2018; Blei et al., 2017) that estimate directly the means and the weights of the three-component mixture, using ten random initializations. The experiment is conducted first without any outlier, and then with an outlier equal to 100. Here, the MSE metric is more complicated to define. Indeed, we try to estimate the difference between densities rather than parameters as the parameters that are estimated are not the same for the different methods (weights over the whole dictionary versus weights and means over the three components). First, we sample 10.000 datapoints independently according to the true mixture. Then, we evaluate the square of the difference between the true density p^0 and the estimated density $p_{\widehat{\theta}_n}$ evaluated at each of the 10.000 datapoints, and we finally take the average:

$$\begin{cases} z_1, ..., z_N \overset{i.i.d}{\sim} p^0 \text{ where } N = 10.000, \\ \text{MSE} = \frac{1}{N} \sum_{\ell=1}^{N} \left| p^0(z_\ell) - p_{\widehat{\theta}_n}(z_\ell) \right|. \end{cases}$$

Again, the final metric is the average over 100 repetitions of the experiment. Figures 5, 6 and 7, and Table 1 clearly show that our estimator performs comparably to both the EM and the CAVI algorithms in the well-specified case, while it is the only one that is not sensitive to the outlier and that gives a consistent estimate.

Algorithm	Without the outlier	With the outlier
MMD	**0.0135**	**0.0142**
CAVI	0.0192	0.0314
EM	0.0136	0.0280

Table 8.1: Square root of the MSE for the Gaussian mixture with and without the outlier

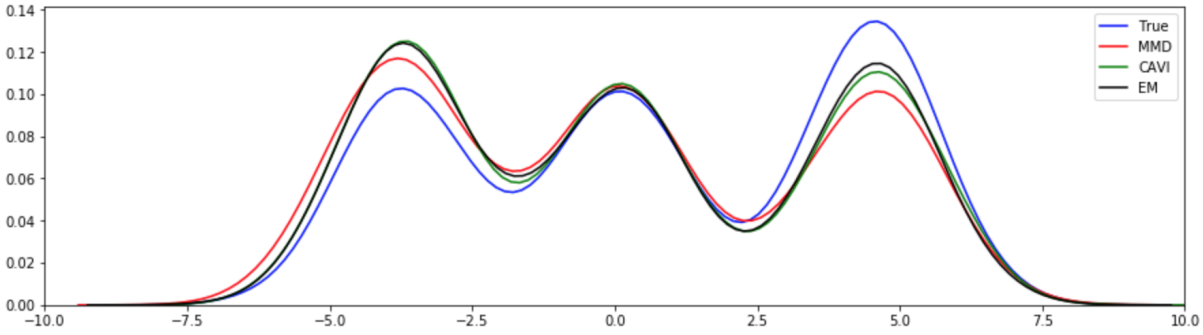


Figure 8.5: Plot of the estimated densities using different methods without outliers. The blue curve represents the true density, the red one the MMD density, the green one the CAVI density and the black one the EM density.

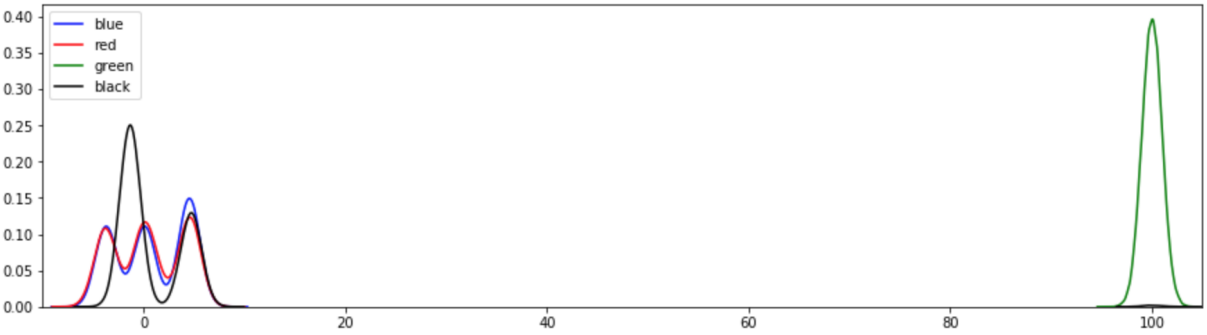


Figure 8.6: Plot of the estimated densities using different methods in presence of 1 outlier at 100. The blue curve represents the true density, the red one the MMD density, the green one the CAVI density and the black one the EM density. The EM estimate has a small component at 100, and CAVI only one component at 100.

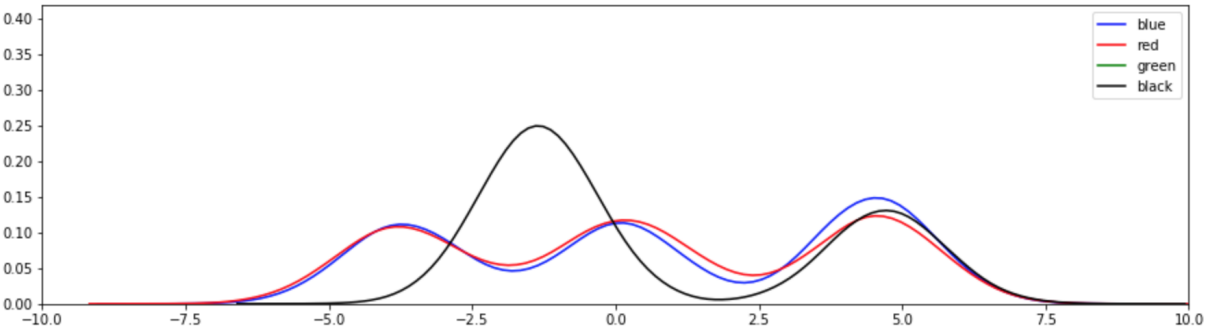


Figure 8.7: Zoom of Figure 6, without the component of EM at 100.

8.7 Proofs

8.7.1 A preliminary lemma: convergence of $\widehat{P}_n$ to P^0 with respect to $\mathbb{D}_k$

Lemma 8.7.1. *We have*

$$\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P^0\right)\right] \leq \frac{1 + 2\sum_{t=1}^{n}\left(1 - \frac{t}{n}\right)\varrho_t}{n}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P^0\right)\right] &= \mathbb{E}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\left[k(X_i,\cdot) - \mu_{P^0}\right]\right\|_{\mathcal{H}_k}^2\right\} \\
&= \frac{1}{n^2}\mathbb{E}\left\{\sum_{i=1}^{n}\|k(X_i,\cdot) - \mu_{P^0}\|_{\mathcal{H}_k}^2 + 2\sum_{1\leq i<j\leq n}\left\langle k(X_i,\cdot) - \mu_{P^0}, k(X_j,\cdot) - \mu_{P^0}\right\rangle_{\mathcal{H}_k}\right\} \\
&\leq \frac{1}{n^2}\left(n + 2\sum_{1\leq i<j\leq n}\varrho_{|i-j|}\right) = \frac{1 + 2\sum_{t=1}^{n}\left(1 - \frac{t}{n}\right)\varrho_t}{n}.
\end{aligned}$$

□

Note that in the i.i.d case, this leads to

$$\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P^0\right)\right] \leq \frac{1}{n}$$

and thus

$$\mathbb{E}\left[\mathbb{D}_k\left(\widehat{P}_n, P^0\right)\right] \leq \sqrt{\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P^0\right)\right]} \leq \frac{1}{\sqrt{n}}.$$

The rate $1/\sqrt{n}$ is known to be minimax in this case: Theorem 1 in Tolstikhin et al. (2017).

8.7.2 Proof of Theorem 8.3.1

Proof. First,

$$\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right) \leq \mathbb{D}_k\left(P_{\widehat{\theta}_n}, \widehat{P}_n\right) + \mathbb{D}_k\left(\widehat{P}_n, P^0\right) \leq \mathbb{D}_k\left(P_\theta, \widehat{P}_n\right) + \mathbb{D}_k\left(\widehat{P}_n, P^0\right)$$

for any $\theta \in \Theta$, by definition of $\widehat{\theta}_n$, and thus, using the triangular inequality again,

$$\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right) \leq \mathbb{D}_k\left(P_\theta, P^0\right) + 2\mathbb{D}_k\left(\widehat{P}_n, P^0\right).$$

Take the expectation on both sides and note that

$$\mathbb{E}\left[\mathbb{D}_k\left(\widehat{P}_n, P^0\right)\right] \leq \sqrt{\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P^0\right)\right]} \leq \sqrt{\frac{1 + 2\sum_{t=1}^{n}\left(1 - \frac{t}{n}\right)\varrho_t}{n}} \leq \sqrt{\frac{1 + 2\sum_{t=1}^{n}\varrho_t}{n}}$$

where the second inequality is given by Lemma 8.7.1. □

8.7.3 Proof of Theorem 8.3.2

We start by reminding the following result from Briol et al. (2019); similar results can be found in Song (2008) or Gretton et al. (2009).

Lemma 8.7.2 (Lemma 1 page 10 (Briol et al., 2019)). *For any $\delta > 0$,*

$$\mathbb{P}\left(\mathbb{D}_k\left(\widehat{P}_n, P^0\right) \leq \frac{1}{\sqrt{n}}\left(1 + \sqrt{\log\left(1/\delta\right)}\right)\right) \geq 1 - \delta.$$

This result (that we won't use here) relies on McDiarmid inequality (McDiarmid, 1989) who proposed a beautiful way to control the difference between a function of the data, $f(X_1, \ldots, X_n)$, and its expectation. The idea relies on writing this function as a martingale, $f(X_1, \ldots, X_n) = M_n$ where M_t, for $t \leq n$, is given by $M_t = \mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_t]$, and controling the martingale increments. It appears that many inequalities can be proven by using this technique, this is discussed in details in Chapter 3 in Boucheron et al. (2012). Using this technique, Rio (2017b) proved a version of McDiarmid's inequality for series satisfying Assumption 8.3.1 (note that the paper is written in French, a more recent paper by the same author (Rio, 2013) in English contains this result and new ones). We start by reminding Rio's result.

Lemma 8.7.3 (Theorem 1 page 906 (Rio, 2017b)). *Assume that $f : \mathcal{H}_k^n \to \mathbb{R}$ satisfies:*

$$\left|f(a_1, \ldots, a_n) - f(a_1', \ldots, a_n')\right| \leq \sum_{i=1}^{n} \|a_i - a_i'\|_{\mathcal{H}_k}.$$

Then, for any $t > 0$,

$$\mathbb{E}\exp\left[tf(\mu_{\delta_{X_1}}, \ldots, \mu_{\delta_{X_1}}) - t\mathbb{E}[f(\mu_{\delta_{X_1}}, \ldots, \mu_{\delta_{X_1}})]\right] \leq \exp\left(\frac{t^2(1+\Gamma)^2 n}{2}\right).$$

This allows us to state our variant of Lemma 8.7.2.

Lemma 8.7.4. *Under Assumptions 8.2.1 and 8.3.1,*

$$\mathbb{P}\left(\mathbb{D}_k\left(\widehat{P}_n, P^0\right) \leq \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right) \geq 1 - \delta.$$

Proof of Lemma 8.7.4. Define

$$f(a_1, \ldots, a_n) = \left\|\sum_{i=1}^{n}(a_i - \mu_{P^0})\right\|_{\mathcal{H}_k}.$$

Under Assumption 8.3.1, the conditions of Lemma 8.7.3 are satisfied, and thus, for any $x > 0$ and any $t > 0$,

$$\mathbb{P}\left(\mathbb{D}_k\left(\widehat{P}_n, P^0\right) - \mathbb{E}[\mathbb{D}_k\left(\widehat{P}_n, P^0\right)] \geq x\right) = \mathbb{P}\left(\frac{f(\mu_{\delta_{X_1}}, \ldots, \mu_{\delta_{X_n}})}{n} - \mathbb{E}[\mathbb{D}_k\left(\widehat{P}_n, P^0\right)] \leq x\right)$$

$$\leq \exp\left(\frac{t^2(1+\Gamma)^2}{2n} - tx\right)$$
$$= \exp\left(-\frac{x^2 n}{2(1+\Gamma)^2}\right)$$

where we chose $t = xn/(1+\Gamma)^2$. Put $x = (1+\Gamma)\sqrt{2\log(1/\delta)/n}$ to get:

$$\mathbb{P}\left(\mathbb{D}_k\left(\widehat{P}_n, P^0\right) \leq \mathbb{E}[\mathbb{D}_k\left(\widehat{P}_n, P^0\right)] + (1+\Gamma)\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}\right) \geq 1-\delta.$$

Plug Theorem 8.7.1 to get the result:

$$\mathbb{P}\left(\mathbb{D}_k\left(\widehat{P}_n, P^0\right) \leq \sqrt{\frac{1+2\Sigma}{n}} + (1+\Gamma)\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}\right) \geq 1-\delta.$$

□

We are now in position to prove Theorem 8.3.2.

Proof of Theorem 8.3.2. With probability $1-\delta$, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{D}_k\left(P_{\widehat{\theta}_n}, P^0\right) &\leq \mathbb{D}_k\left(P_{\widehat{\theta}_n}, \widehat{P}_n\right) + \mathbb{D}_k\left(\widehat{P}_n, P^0\right) \\
&\leq \mathbb{D}_k\left(P_\theta, \widehat{P}_n\right) + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}} \\
&\leq \mathbb{D}_k\left(P_\theta, P^0\right) + \mathbb{D}_k\left(\widehat{P}_n, P^0\right) + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}} \\
&\leq \mathbb{D}_k\left(P_\theta, P^0\right) + 2\frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}
\end{aligned}$$

□

8.7.4 Proof of Lemma 8.3.3 and of Proposition 3

Proof of Lemma 8.3.3. We have

$$\begin{aligned}
\left|\mathbb{D}_k(P_{\widehat{\theta}_n}, P^0) - \mathbb{D}_k(P_{\widehat{\theta}_n}, P_{\theta^0})\right| &\leq \mathbb{D}_k(P, P^0) \\
&= \left\|(1-\epsilon)\mu_{P_{\theta^0}} + \epsilon\mu_Q - \mu_{P_{\theta^0}}\right\|_{\mathcal{H}_k} \\
&= \left\|\epsilon(\mu_Q - \mu_{P_{\theta^0}})\right\|_{\mathcal{H}_k} \\
&\leq 2\epsilon.
\end{aligned}$$

□

Proof of Proposition 3. Let us put

$$\widetilde{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\widetilde{X}_i}.$$

First, note that for any probability measure Q,

$$\begin{aligned}
\left|\mathbb{D}_k(Q,\widetilde{P}_n) - \mathbb{D}_k(Q,\widehat{P}_n)\right| &\leq \mathbb{D}_k(\widehat{P}_n,\widetilde{P}_n) \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n} \Big(k(X_i,\cdot) - k(\widetilde{X}_i,\cdot)\Big) \right\|_{\mathcal{H}_k} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\| k(X_i,\cdot) - k(\widetilde{X}_i,\cdot) \right\|_{\mathcal{H}_k} \\
&= \frac{1}{n}\sum_{i\in\mathcal{O}} \left\| k(X_i,\cdot) - k(\widetilde{X}_i,\cdot) \right\|_{\mathcal{H}_k} \\
&\leq \frac{2|\mathcal{O}|}{n} \\
&\leq 2\epsilon.
\end{aligned}$$

Consider $Q = P_{\widetilde{\theta}_n}$. Then:

$$\begin{aligned}
\mathbb{D}_k(P_{\widetilde{\theta}_n}, P^0) &\leq \mathbb{D}_k(P_{\widetilde{\theta}_n}, \widetilde{P}_n) + \mathbb{D}_k(\widetilde{P}_n, P^0) \\
&\leq \mathbb{D}_k(P_{\widehat{\theta}_n}, \widetilde{P}_n) + \mathbb{D}_k(\widetilde{P}_n, P^0) \text{ by definition of } \widetilde{\theta}_n \\
&\leq \left[2\epsilon + \mathbb{D}_k(P_{\widehat{\theta}_n}, \widehat{P}_n)\right] + \left[2\epsilon + \mathbb{D}_k(\widehat{P}_n, P^0)\right]
\end{aligned}$$

where we used the previous derivations for $Q = P_{\widehat{\theta}_n}$ and then $Q = P^0$ respectively. So:

$$\begin{aligned}
\mathbb{D}_k(P_{\widetilde{\theta}_n}, P^0) &\leq 4\varepsilon + \mathbb{D}_k(P_{\widehat{\theta}_n}, \widehat{P}_n) + \mathbb{D}_k(\widehat{P}_n, P^0) \\
&\leq 4\varepsilon + \mathbb{D}_k(P_{\theta^0}, \widehat{P}_n) + \mathbb{D}_k(\widehat{P}_n, P^0) \text{ by definition of } \widehat{\theta}_n \\
&= 4\varepsilon + 2\mathbb{D}_k(\widehat{P}_n, P^0)
\end{aligned}$$

as it is here assumed that $P^0 = P_{\theta^0}$. □

8.7.5 Proof of the results in Subsection 8.3.3

Proof of Theorem 8.3.5. With probability at least $1-\varepsilon$, for any $\theta \in \Theta$,

$$\begin{aligned}
\|p_{\widehat{\theta}_n} - p^0\|_{L^2} &\leq \|p_{\widehat{\theta}_n} - p_\theta\|_{L^2} + \|p_\theta - p^0\|_{L^2} \\
&\leq \frac{\mathbb{D}_{k_\gamma}(P_{\widehat{\theta}_n}, P_\theta)}{\mathcal{L}(\gamma)} + \|p_\theta - p^0\|_{L^2} \\
&\leq \frac{\mathbb{D}_{k_\gamma}(P_{\widehat{\theta}_n}, \widehat{P}_n) + \mathbb{D}_{k_\gamma}(P_\theta, \widehat{P}_n)}{\mathcal{L}(\gamma)} + \|p_\theta - p^0\|_{L^2} \\
&\leq \frac{2\mathbb{D}_{k_\gamma}(P_\theta, \widehat{P}_n)}{\mathcal{L}(\gamma)} + \|p_\theta - p^0\|_{L^2}
\end{aligned}$$

$$\leq \frac{2\mathbb{D}_{k_\gamma}(P_\theta, P^0) + 2\mathbb{D}_{k_\gamma}(P^0, \widehat{P}_n)}{\mathcal{L}(\gamma)} + \|p_\theta - p^0\|_{L^2}$$
$$= \left(1 + \frac{2\mathcal{U}(\gamma)}{\mathcal{L}(\gamma)}\right) \|p_\theta - p^0\|_{L^2} + \frac{2}{\mathcal{L}(\gamma)} \mathbb{D}_{k_\gamma}(\widehat{P}_n, P^0).$$

Take expectation on both sides to obtain

$$\mathbb{E}\left(\|p_{\widehat{\theta}_n} - p^0\|_{L^2}\right) \leq \left(1 + \frac{2\mathcal{U}(\gamma)}{\mathcal{L}(\gamma)}\right) \|p_\theta - p^0\|_{L^2} + \frac{2}{\mathcal{L}(\gamma)} \mathbb{E}\left(\mathbb{D}_{k_\gamma}(\widehat{P}_n, P^0)\right)$$
$$\leq \left(1 + \frac{2\mathcal{U}(\gamma)}{\mathcal{L}(\gamma)}\right) \|p_\theta - p^0\|_{L^2} + \frac{2}{\mathcal{L}(\gamma)} \sqrt{\frac{1 + 2\sum_{t=1}^n \left(1 - \frac{t}{n}\right) \varrho_t}{n}}$$

from Lemma 8.7.1. The proof of the result in probability is similar, except that we use Lemma 8.7.2 to bound $\mathbb{D}_{k_\gamma}(\widehat{P}_n, P^0)$ is probability rather than in expectation. □

Proof of Proposition 4. We remind a few properties of the Fourier transform. First,

$$\mathcal{F}[k_\gamma](t) = \gamma^d \mathcal{F}[K_1](\gamma t) = \gamma^d \mu(\gamma t).$$

Let $\star$ denote the convolution product:

$$p \star q(x) = \int p(x - t) q(t) \mathrm{d}t$$

and we remind the classical result $\mathcal{F}[p \star q] = \mathcal{F}[p]\mathcal{F}[q]$. Finally, we remind that

$$\int |p|^2(x) \mathrm{d}x = \int |\mathcal{F}[p](t)|^2 \mathrm{d}t.$$

Keeping this in mind,

$$\mathbb{D}^2_{k_\gamma}(P_\theta, P_{\theta'}) = \iint k_\gamma(y - x)[p_\theta(x) - p_{\theta'}(x)][p_\theta(y) - p_{\theta'}(y)] \mathrm{d}x \mathrm{d}y$$
$$= \int [k_\gamma \star (p_\theta - p_{\theta'})](y)[p_\theta(y) - p_{\theta'}(y)] \mathrm{d}y$$
$$= \int \mathcal{F}[k_\gamma \star (p_\theta - p_{\theta'})](t) \overline{\mathcal{F}[p_\theta - p_{\theta'}](t)} \mathrm{d}t$$
$$= \int \mathcal{F}[k_\gamma] \mathcal{F}[p_\theta - p_{\theta'}](t) \overline{\mathcal{F}[p_\theta - p_{\theta'}](t)} \mathrm{d}t$$
$$= \int \gamma^d \mu(\gamma t) \left|\mathcal{F}[p_\theta - p_{\theta'}](t)\right|^2 \mathrm{d}t.$$

To obtain the upper bound, note that:

$$\int \gamma^d \mu(\gamma t) \left|\mathcal{F}[p_\theta - p_{\theta'}](t)\right|^2 \mathrm{d}t \leq \gamma^d D \int \left|\mathcal{F}[p_\theta - p_{\theta'}](u)\right|^2 \mathrm{d}u = \gamma^d D \|p_\theta - p_{\theta'}\|^2_{L_2}.$$

The lower bound is given by

$$\int \gamma^d \mu(\gamma t) \left|\mathcal{F}[p_\theta - p_{\theta'}](t)\right|^2 \mathrm{d}t \geq b^2 \gamma^d \int_{\|\gamma t\| \leq a} \left|\mathcal{F}[p_\theta - p_{\theta'}](t)\right|^2 \mathrm{d}t$$
$$= b^2 \gamma^d \int_{\|t\| \leq \frac{a}{\gamma}} \left|\mathcal{F}[p_\theta - p_{\theta'}](t)\right|^2 \mathrm{d}t \geq b^2 \gamma^d \mathcal{A} \left(\frac{a}{\gamma}\right)^2 \|p_\theta - p_{\theta'}\|^2_{L_2}.$$

□

8.7.6 Proofs of Section 8.4

Proof of Proposition 5. We remind that $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ where $\theta \in \Theta = \mathbb{R}^d$. When X and Y are independent, respectively from P_θ and $P_{\theta'}$, we have $(X-Y) \sim \mathcal{N}(\theta-\theta', 2\sigma^2 I_d)$. Thus,

$$\frac{(X-Y)}{\sqrt{2\sigma^2}} \sim \mathcal{N}\left(\frac{(\theta-\theta')}{\sqrt{2\sigma^2}}, I_d\right)$$

and thus the square of this random variable is a noncentral chi-square random variable:

$$\frac{\|X-Y\|^2}{2\sigma^2} \sim \chi^2\left(d, \frac{\|\theta-\theta'\|^2}{2\sigma^2}\right).$$

It is known that when $U \sim \chi^2(d,m)$ we have $\mathbb{E}[\exp(tU)] = \exp(mt/(1-2t))/(1-2t)^{d/2}$. Taking $t = -(2\sigma^2)/\gamma^2$, this leads to

$$\left\langle \mu_{P_\theta}, \mu_{P_{\theta'}} \right\rangle_{\mathcal{H}} = \mathbb{E}_{X\sim P_\theta, Y\sim P_{\theta'}}\left[\exp\left(-\frac{\|X-Y\|^2}{\gamma^2}\right)\right] = \left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2}\right) \tag{8.7}$$

and thus

$$\mathbb{D}^2_{k_\gamma}(P_\theta, P_{\theta'}) = 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left[1-\exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2}\right)\right].$$

We also have:

$$\langle p_\theta, p_{\theta'}\rangle_{L^2} = \frac{\exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2}\right)}{(4\pi\sigma^2)^{d/2}} \Rightarrow \|p_\theta - p_{\theta'}\|^2_{L^2} = \frac{2\left[1-\exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2}\right)\right]}{(4\pi\sigma^2)^{d/2}}.$$

So:

$$\frac{\mathbb{D}^2_{k_\gamma}(P_\theta, P_{\theta'})}{\|p_\theta - p_{\theta'}\|^2_{L^2}} = \left(\frac{4\pi\sigma^2\gamma^2}{4\sigma^2+\gamma^2}\right)^{d/2} \frac{1-\exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2}\right)}{1-\exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2}\right)}.$$

Consider the function f defined, for $u>0$, by

$$f(u) = \frac{1-\exp(-\frac{u}{4\sigma^2+\gamma^2})}{1-\exp(-\frac{u}{4\sigma^2})}.$$

Note that $f(u) \to \frac{4\sigma^2}{4\sigma^2+\gamma^2}$ when $u \to 0$, f is nondecreasing and $f(\infty) = 1$. This leads to

$$\mathcal{L}(\gamma) = \frac{4\sigma^2}{4\sigma^2+\gamma^2}\left(\frac{4\pi\sigma^2\gamma^2}{4\sigma^2+\gamma^2}\right)^{d/2} \text{ and } \mathcal{U}(\gamma) = \left(\frac{4\pi\sigma^2\gamma^2}{4\sigma^2+\gamma^2}\right)^{d/2}.$$

Pluging this into Theorem 8.3.5 gives, with probability at least $1-\delta$,

$$\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \le \left(1+\frac{4\sigma^2+\gamma^2}{2\sigma^2}\right)\inf_{\theta\in\Theta}\|p_\theta - p^0\|_{L^2} + \frac{4\sigma^2+\gamma^2}{2\sigma^2}\left(\frac{4\sigma^2+\gamma^2}{4\pi\sigma^2\gamma^2}\right)^{d/2}\frac{2+2\sqrt{2\log(1/\delta)}}{\sqrt{n}}.$$

This proves (8.3). Note that the second term in the right-hand side is

$$\frac{1}{2\sigma^2(4\pi\sigma^2)^{\frac{d}{2}}}\frac{2+2\sqrt{2\log(1/\delta)}}{\sqrt{n}}g(\gamma)$$

where

$$g(\gamma)=\frac{(4\sigma^2+\gamma^2)^{d/2+1}}{\gamma^d}.$$

That is,

$$g'(\gamma)=\frac{2\left(\frac{d}{2}+1\right)(4\sigma^2+\gamma^2)^{d/2}\gamma^{d+1}-d\,(4\sigma^2+\gamma^2)^{d/2+1}\gamma^{d-1}}{\gamma^{2d}}$$

and thus $g'(\gamma)=0$ is equivalent to:

$$2\left(\frac{d}{2}+1\right)\gamma^2=d(4\sigma^2+\gamma^2),$$

that is $\gamma^2=2\sigma^2 d$. From now, we consider $\gamma^2=2\sigma^2 d$, this leads to

$$\frac{4\sigma^2+\gamma^2}{2\sigma^2}=\frac{4\sigma^2+2\sigma^2 d}{2\sigma^2}=2+d,$$

while

$$g(\gamma)=\left(1+\frac{2}{d}\right)^{d/2}2\sigma^2(d+2)\leq 2\mathrm{e}\sigma^2(d+2)$$

that is

$$\frac{1}{2\sigma^2(4\pi\sigma^2)^{\frac{d}{2}}}\frac{2+2\sqrt{2\log(1/\delta)}}{\sqrt{n}}g(\gamma^2)\leq\frac{\mathrm{e}(d+2)}{(4\pi\sigma^2)^{\frac{d}{2}}}\frac{2+2\sqrt{2\log(1/\delta)}}{\sqrt{n}}$$

and thus (8.3) becomes

$$\|p_{\widehat{\theta}_n}-p^0\|_{L^2}\leq(3+d)\inf_{\theta\in\Theta}\|p_\theta-p^0\|_{L^2}+\frac{\mathrm{e}(d+2)}{(4\pi\sigma^2)^{\frac{d}{2}}}\frac{2+2\sqrt{2\log(1/\delta)}}{\sqrt{n}}$$

that is (8.4).

Let us now consider the estimation of the parameter θ^0 in the context of Proposition 3. From (8.7) and Proposition 3, we obtain, with probability at least $1-\delta$,

$$2\left(\frac{d}{d+2}\right)^{d/2}\left[1-\exp\left(-\frac{\|\widetilde{\theta}_n-\theta^0\|^2}{2\sigma^2(2+d)}\right)\right]=\mathbb{D}_k^2\left(P_{\widetilde{\theta}_n},P_{\theta^0}\right)\leq 16\left(\epsilon+\frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2,$$

that is

$$\left[1-\exp\left(-\frac{\|\widetilde{\theta}_n-\theta^0\|^2}{2\sigma^2(2+d)}\right)\right]\leq 8\mathrm{e}\left(\epsilon+\frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2,$$

and thus

$$\|\widetilde{\theta}_n - \theta^0\|^2 \leq -2\sigma^2(d+2)\log\left[1 - 8\mathrm{e}\left(\epsilon + \frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2\right].$$

This is (8.5). □

Proof of Proposition 6. We have

$$\langle p_\theta, p_{\theta'}\rangle_{L^2} = \frac{2}{\pi[(\theta-\theta')^2+4]} \Rightarrow \|p_\theta - p_{\theta'}\|_{L^2}^2 = \frac{1}{\pi}\left(1 - \frac{1}{\frac{(\theta-\theta')^2}{4}+1}\right).$$

We use Theorem 8.3.5 and the upper bound $\mathcal{U}(\gamma) \leq (2\pi)^{1/4}\gamma^{1/2}$. Regarding $\mathcal{L}(\gamma)$, note that

$$\mathcal{F}[p_\theta - p_{\theta'}](t) = [\exp(-it\theta) - \exp(-it\theta')]\exp(-|t|)$$

and so

$$\begin{aligned}
\frac{\mathbb{D}_{k_\gamma}^2(P_\theta, P'_\theta)}{\|p_\theta - p'_\theta\|_{L^2}^2} &= \frac{\pi\int\gamma\mu(\gamma t)\left|[\exp(-it\theta) - \exp(-it\theta')]\exp(-|t|)\right|^2}{1 - \frac{1}{\frac{(\theta-\theta')^2}{4}+1}} \\
&\geq \frac{\pi\int\gamma\mu(\gamma t)\left|[\exp(-it\theta) - \exp(-it\theta')]\exp(-t^2-1)\right|^2}{1 - \frac{1}{\frac{(\theta-\theta')^2}{4}+1}} \\
&= \frac{2\pi\left[1 - \exp\left(-\frac{\|\theta-\theta'\|^2}{2(\gamma+2)}\right)\right]}{\exp(2)\sqrt{1+\frac{2}{\gamma}}\left(1 - \frac{1}{\frac{(\theta-\theta')^2}{4}+1}\right)} \\
&\geq \frac{\pi}{\exp(2)}\sqrt{\frac{2}{3}} \geq \frac{1}{3}
\end{aligned}$$

for $\gamma = 2$. So:

$$\mathbb{P}\left\{\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq 6(2\pi)^{1/4}\sqrt{2}\inf_{\theta\in\Theta}\|p_\theta - p^0\|_{L^2} + \frac{6+6\sqrt{2\log(1/\delta)}}{\sqrt{n}}\right\} \geq 1-\delta.$$

We actually have $6(2\pi)^{1/4}\sqrt{2} \leq 14$ and so

$$\mathbb{P}\left\{\|p_{\widehat{\theta}_n} - p^0\|_{L^2} \leq 14\inf_{\theta\in\Theta}\|p_\theta - p^0\|_{L^2} + \frac{6+6\sqrt{2\log(1/\delta)}}{\sqrt{n}}\right\} \geq 1-\delta.$$

Regarding the parameter estimation in the adversarial contamination case:

$$\begin{aligned}
(\widehat{\theta}_n - \theta_0)^2 &= 4\left[1 - \frac{1}{1-\pi\|p_{\widetilde{\theta}} - p_{\theta^0}\|^2}\right] \\
&\leq 4\left[1 - \frac{1}{1-3\pi\mathbb{D}_{K_\gamma}^2(P_{\widetilde{\theta}}, P_{\theta_0})}\right]
\end{aligned}$$

$$\leq 4\left[1-\frac{1}{1-96\pi\left(\epsilon^2+\frac{2+4\log(1/\delta)}{n}\right)}\right]$$

from Proposition 3. □

Proof of Proposition 7. For the upper bound, we still use $\mathcal{U}(\gamma) \leq (2\pi)^{1/4}\gamma^{1/2}$, and thus $\mathcal{U}(2) = 2^{3/4}\pi^{1/4} < 2.24$. Let us now focus on the lower bound. We begin, as usual, by the calculation of the L_2 norm:

$$\langle p_\theta, p_{\theta'}\rangle_{L^2} = (1-|\theta-\theta'|)_+ \Rightarrow \|p_\theta - p_{\theta'}\|_{L^2}^2 = 2\min\left(1, |\theta-\theta'|\right).$$

Then, note that

$$\begin{aligned}
\left\langle \mu_{P_\theta}, \mu_{P_{\theta'}}\right\rangle_{\mathcal{H}_{k_\gamma}} &= \int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}}\int_{\theta'-\frac{1}{2}}^{\theta'+\frac{1}{2}} \exp\left(-\frac{(x-y)^2}{\gamma^2}\right)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\theta-\theta'-\frac{1}{2}}^{\theta-\theta'+\frac{1}{2}} \exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}u\mathrm{d}v \\
&= \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{|\theta-\theta'|-\frac{1}{2}}^{|\theta-\theta'|+\frac{1}{2}} \exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}u\mathrm{d}v
\end{aligned}$$

where the last equality comes from the symmetry with respect to θ and θ'. So

$$\mathbb{D}_{k_\gamma}(P_\theta, P_{\theta'}) = \int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{d}u\left[2\int_{-\frac{1}{2}}^{\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v - 2\int_{|\theta-\theta'|-\frac{1}{2}}^{|\theta-\theta'|+\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v\right].$$

First, condider the case where $|\theta-\theta'| \leq 1$. Then

$$\begin{aligned}
\mathbb{D}_{k_\gamma}(P_\theta, P_{\theta'}) &= \int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{d}u\left[2\int_{-\frac{1}{2}}^{|\theta-\theta'|-\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v + 2\int_{\frac{1}{2}}^{|\theta-\theta'|+\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v\right] \\
&\geq \int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{d}u\left[2\int_{\frac{1}{2}}^{|\theta-\theta'|+\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v\right] \\
&\geq 2|\theta-\theta'|\exp\left(-\frac{\left(\frac{1}{2}+\frac{1}{2}+|\theta-\theta'|\right)^2}{\gamma^2}\right) \\
&= 2|\theta-\theta'|\exp\left(-\frac{4}{\gamma^2}\right).
\end{aligned}$$

For $|\theta-\theta'| > 1$, note that

$$\begin{aligned}
\mathbb{D}_{k_\gamma}(P_\theta, P_{\theta'}) &= \int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{d}u\left[2\int_{-\frac{1}{2}}^{\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v - 2\int_{|\theta-\theta'|-\frac{1}{2}}^{|\theta-\theta'|+\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v\right] \\
&\geq \int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{d}u\left[2\int_{-\frac{1}{2}}^{\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v - 2\int_{\frac{1}{2}}^{\frac{3}{2}}\exp\left(-\frac{(u-v)^2}{\gamma^2}\right)\mathrm{d}v\right].
\end{aligned}$$

As a special case, for $\gamma = 2$, we have

$$\mathbb{D}_{K_2}(P_\theta, P_{\theta'}) \geq 2\min\left\{\frac{|\theta-\theta'|}{\mathrm{e}}, \int_{-\frac{1}{2}}^{\frac{1}{2}}\mathrm{d}u\left[\int_{-\frac{1}{2}}^{\frac{1}{2}}\exp\left(-\frac{(u-v)^2}{4}\right)\mathrm{d}v - \int_{\frac{1}{2}}^{\frac{3}{2}}\exp\left(-\frac{(u-v)^2}{4}\right)\mathrm{d}v\right]\right\}$$

and a Monte-Carlo integration shows that the integral above is $\simeq 0.19 > 0.10$. We thus have

$$\mathcal{L}(2) \geq \frac{\mathbb{D}_{K_2}(P_\theta, P_{\theta'})}{\|p_\theta - p_{\theta'}\|_{L_2}} \geq \frac{\min\left(\frac{2}{\mathrm{e}}|\theta - \theta'|, \frac{2}{10}\right)}{\min(|\theta - \theta'|, 1)} = \frac{2}{10} = \frac{1}{5}.$$

Thus, Theorem 8.3.5 gives, with probability at least $1 - \delta$,

$$\begin{aligned}\|p_{\widehat{\theta}_n} - p^0\|_{L^2} &\leq (1 + 2 \times 2.24 \times 5) \inf_{\theta \in \Theta} \|p_\theta - p^0\|_{L^2} + \frac{5\left(2 + 2\sqrt{2\log(1/\delta)}\right)}{\sqrt{n}} \\ &= 23.4 \inf_{\theta \in \Theta} \|p_\theta - p^0\|_{L^2} + \frac{10 + 10\sqrt{2\log(1/\delta)}}{\sqrt{n}}.\end{aligned}$$

□

Proof of Proposition 9. We remind that $P_{0:t}$ is the distribution of (X_0, X_t). Then

$$\begin{aligned}\varrho_t &= \left|\mathbb{E}\left\langle \mu_{\delta_{X_t}} - \mu_{P^0}, \mu_{\delta_{X_0}} - \mu_{P^0}\right\rangle_{\mathcal{H}_k}\right| \\ &= \left|\int k(x,y) P_{0:t}(\mathrm{d}(x,y)) - \iint k(x,y) P^0(\mathrm{d}x) P^0(\mathrm{d}y)\right| \\ &= \left|\int \left(\int \mathbf{1}_{\{u \geq \|x-y\|\}} f(u)\mathrm{d}u\right) P_{0:t}(\mathrm{d}(x,y)) - \iint \left(\int \mathbf{1}_{\{u > \|x-y\|\}} f(u)\mathrm{d}u P^0(\mathrm{d}x) P^0(\mathrm{d}y)\right)\right| \\ &= \left|\int_0^\infty \left(\int \mathbf{1}_{\{u \geq \|x-y\|\}} P_{0:t}(\mathrm{d}(x,y)) - \iint \mathbf{1}_{\{u \geq \|x-y\|\}}\ P^0(\mathrm{d}x) P^0(\mathrm{d}y)\right) f(u)\mathrm{d}u\right|.\end{aligned}$$

For any partition $(A_i)_{i \in I}$ of $\mathbb{R}^d$ denote $I(u) = \{i \in I : (x,y) \in A_i^2 \Rightarrow \|x - y\| \leq u\}$. Then

$$\sum_{i \in I(u)} \mathbf{1}_{A_i}(x) \mathbf{1}_{A_i}(y) \leq \mathbf{1}_{\{\|x-y\| \leq u\}}$$

and moreover $\mathbf{1}_{\{\|x-y\| \leq u\}}$ is the supremum of this sum over all possible measurable partitions, that is, for any $\varepsilon > 0$, we can find a partition $(A_i)_{i \in I}$ such that

$$\mathbf{1}_{\{\|x-y\| \leq u\}} - \varepsilon \leq \sum_{i \in I(u)} \mathbf{1}_{A_i}(x) \mathbf{1}_{A_i}(y) \leq \mathbf{1}_{\{\|x-y\| \leq u\}}.$$

So,

$$\begin{aligned}\varrho_t &\leq \left|\int_0^\infty \sum_{i \in I(u)} [P_{0:t}(A_i \times A_i) - P^0(A_i)^2] f(u)\mathrm{d}u\right| + \varepsilon \\ &\leq \int_0^\infty \sum_{i \in I(u)} \left|P_{0:t}(A_i \times A_i) - P^0(A_i)^2\right| f(u)\mathrm{d}u + \varepsilon \\ &\leq \int_0^\infty \sum_{i \in I} \left|P_{0:t}(A_i \times A_i) - P^0(A_i)^2\right| f(u)\mathrm{d}u + \varepsilon \\ &\leq \int_0^\infty 2\beta_t f(u)\mathrm{d}u + \varepsilon = 2\beta_t + \varepsilon.\end{aligned}$$

□

Proof of Proposition 10. Let P_b denote the distribution of B_k. We have

$$\begin{aligned}
\varrho_t &= \left| \int k(x,y) P_{0:t}(\mathrm{d}(x,y)) - \iint k(x,y) P^0(\mathrm{d}x) P^0(\mathrm{d}y) \right| \\
&= \left| \iint k(x, G_t(x,b)) P^0(\mathrm{d}x) P_b(\mathrm{d}b) - \iiint k(x, G_t(x',b)) P^0(\mathrm{d}x) P^0(\mathrm{d}x') P_b(\mathrm{d}b) \right| \\
&\leq \iiint |k(x, G_t(x,b)) - k(x, G_t(x',b))| \, P^0(\mathrm{d}x) P^0(\mathrm{d}x') P_b(\mathrm{d}b) \\
&\leq \iiint L \, \|G_t(x,b) - G_t(x',b)\| \, P^0(\mathrm{d}x) P^0(\mathrm{d}x') P_b(\mathrm{d}b) \\
&\leq \iint L L_k \, \|x - x'\| \, P^0(\mathrm{d}x) P^0(\mathrm{d}x') \\
&\leq \int 2 L L_k \, \|x\| \, P^0(\mathrm{d}x) \\
&= 2 L L_k \mathbb{E}(\|X_0\|).
\end{aligned}$$

□

8.7.7 Proofs of Section 8.5

Proof of Proposition 11. Note that we can rewrite

$$\mathrm{Crit}(\theta) = \iint k(x,x') p_\theta(x) p_\theta(x') \mu(\mathrm{d}x) \mu(\mathrm{d}x') - \frac{2}{n} \sum_{i=1}^n \int k(x, X_i) p_\theta(x) \mu(\mathrm{d}x).$$

The assumption of the proposition ensure that we can interexchange the ∇ and $\iint$ symbols, and so

$$\begin{aligned}
\nabla_\theta \mathrm{Crit}(\theta) &= \iint k(x,x') \nabla_\theta [p_\theta(x) p_\theta(x')] \mu(\mathrm{d}x) \mu(\mathrm{d}x') - \frac{2}{n} \sum_{i=1}^n \int k(x, X_i) \nabla_\theta [p_\theta(x)] \mu(\mathrm{d}x) \\
&= 2 \iint k(x,x') p_\theta(x) p_\theta(x') \nabla_\theta [\log p_\theta(x)] \mu(\mathrm{d}x) \mu(\mathrm{d}x') \\
&\qquad - \frac{2}{n} \sum_{i=1}^n \int k(x, X_i) \nabla_\theta [\log p_\theta(x)] p_\theta(x) \mu(\mathrm{d}x) \\
&= 2 \mathbb{E}_{X,X' \sim P_\theta} \{ k(X, X') \nabla_\theta [\log p_\theta(X)] \} - \frac{2}{n} \sum_{i=1}^n \mathbb{E}_{X \sim P_\theta} \{ k(X_i, X) \nabla_\theta [\log p_\theta(X)] \} \\
&= 2 \mathbb{E}_{X,X' \sim P_\theta} \left\{ \left[k(X, X') - \frac{1}{n} \sum_{i=1}^n k(X_i, X) \right] \nabla_\theta [\log p_\theta(X)] \right\}.
\end{aligned}$$

This ends the proof. □

Proof of Proposition 12. The assumption that Θ is bounded with radius D ensures that (2.17) in Nemirovski et al. (2009) is satisfied, and the assumption on the expectation of the norm of the gradient ensures that (2.5) in Nemirovski et al. (2009) is also satisfied. Thus, (2.21) is also satisfied, ant that is exactly the statement of our (8.6). Then, we have:

$$\mathbb{E}\left[\mathbb{D}_k \left(P_{\widehat{\theta}_n^{(T)}}, P^0 \right) \right] \leq \mathbb{D}_k \left(P_{\widehat{\theta}_n^{(T)}}, \widehat{P}_n \right) + \mathbb{D}_k \left(\widehat{P}_n, P^0 \right)$$

$$= \sqrt{\mathbb{D}_k^2\left(P_{\widehat{\theta}_n^{(T)}}, \widehat{P}_n\right)} + \mathbb{D}_k\left(\widehat{P}_n, P^0\right)$$
$$\leq \sqrt{\mathbb{D}_k^2\left(P_{\widehat{\theta}_n}, \widehat{P}_n\right) + \frac{DM}{\sqrt{T}}} + \mathbb{D}_k\left(\widehat{P}_n, P^0\right)$$

thanks to (8.6). We upper bound the second term thanks to Lemma 8.7.1:

$$\mathbb{D}_k\left(\widehat{P}_n, P^0\right) \leq \sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}}.$$

For the first term, we use:

$$\sqrt{\mathbb{D}_k^2\left(P_{\widehat{\theta}_n}, \widehat{P}_n\right) + \frac{DM}{\sqrt{T}}} \leq \mathbb{D}_k\left(P_{\widehat{\theta}_n}, \widehat{P}_n\right) + \sqrt{\frac{DM}{\sqrt{T}}}$$
$$\leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \widehat{P}_n\right) + 2\sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}} + \sqrt{\frac{DM}{\sqrt{T}}}$$

thanks to Theorem 8.3.1. Putting everything together leads to

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\widehat{\theta}_n^{(T)}}, P^0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \widehat{P}_n\right) + 3\sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}} + \sqrt{\frac{DM}{\sqrt{T}}}$$

which ends the proof. □

8.8 Conclusion

Parametric estimation with MMD provides a simple way to define universally consistent, robust estimators. In many, but not in all, settings, these estimators also have optimal rates of convergence. The computation of the MMD-based estimator can generally be done through a stochastic gradient descent. We thus believe that it is a practically reasonable and nice alternative to many robust estimation procedures.

Interestingly, Proposition 5 provides a natural calibration to the kernel parameter, which is usually a problem in practice. However, in more general settings, the calibration of this parameter, and the choice of the kernel, remain important open questions.

The application of this method to more sophisticated models in statistics and in machine learning (time series models, regression) should be investigated in details and will be the object of future works.

Chapter 9

MMD-Bayes: Robust Bayesian Estimation via Maximum Mean Discrepancy

In some misspecified settings, the posterior distribution in Bayesian statistics may lead to inconsistent estimates. To fix this issue, it has been suggested to replace the likelihood by a pseudo-likelihood, that is the exponential of a loss function enjoying suitable robustness properties. In this chapter, we build a pseudo-likelihood based on the Maximum Mean Discrepancy, defined via an embedding of probability distributions into a reproducing kernel Hilbert space. We show that this MMD-Bayes posterior is consistent and robust to model misspecification. As the posterior obtained in this way might be intractable, we also prove that reasonable variational approximations of this posterior enjoy the same properties. We provide details on a stochastic gradient algorithm to compute these variational approximations. Numerical simulations indeed suggest that our estimator is more robust to misspecification than the ones based on the likelihood.

9.1 Introduction

Bayesian methods are very popular in statistics and machine learning as they provide a natural way to model uncertainty. Some subjective prior distribution π is updated using the negative log-likelihood ℓ_n via Bayes' rule to give the posterior $\pi_n(\theta) \propto \pi(\theta)\exp(-\ell_n(\theta))$. Nevertheless, the classical Bayesian methodology is not robust to model misspecification. There are many cases where the posterior is not consistent (Barron et al., 1999; Grünwald et al., 2017), and there is a need to develop methodologies yielding robust estimates. A way to fix this problem is to replace the log-likelihood ℓ_n by a relevant risk measure. This idea is at the core of the PAC-Bayes theory (Catoni, 2007) and Gibbs posteriors (Syring and Martin, 2018); its connection with Bayesian principles are discussed in Bissiri et al. (2016). Knoblauch et al. (2019) builds a general representation of Bayesian inference in the spirit of Bissiri et al. (2016) and extends the representation to the approximate inference case. In particular, the use of a robust divergence has been shown to provide an estimator that is robust to misspecification (Knoblauch et al.,

2019). For instance, Hooker and Vidyashankar (2014) investigated the case of Hellinger-based divergences, Ghosh and Basu (2016), Futami et al. (2018), and Nakagawa and Hashimoto (2019) used robust β- and γ-divergences, while Catoni (2012), Baraud and Birgé (2016) and Holland (2019b) replaced the logarithm of the log-likelihood by wisely chosen bounded functions. Refer to Jewson et al. (2018) for a complete survey on robust divergence-based Bayes inference.

In this chapter, we consider the Maximum Mean Discrepancy (MMD) as the alternative loss used in Bayes' formula, leading to a pseudo-posterior that we shall call MMD-Bayes in the following. MMD is built upon an embedding of distributions into a reproducing kernel Hilbert space (RKHS) that generalizes the original feature map to probability measures, and allows to apply tools from kernel methods in parametric estimation. Our MMD-Bayes posterior is related to the kernel-based posteriors in Fukumizu et al. (2013), Park et al. (2016) and Ridgway (2017), even though it is different. More recently, Briol et al. (2019) introduced a frequentist minimum distance estimator based on the MMD distance, that is shown to be consistent and robust to small deviations from the model. We show that our MMD-Bayes retains the same properties, i.e is consistent at the minimax optimal rate of convergence as the minimum MMD estimator, and is also robust to misspecification, including data contamination and outliers. Moreover, we show that these guarantees are still valid when considering a tractable approximation of the MMD-Bayes via variational inference, and we support our theoretical results with experiments showing that our approximation is robust to outliers for various estimation problems. All the proofs are deferred to the appendix.

9.2 Framework and definitions

Let us introduce the background and theoretical tools required to understand the rest of the paper. We consider in a measurable space $\left(\mathbb{X}, \mathcal{X}\right)$ a collection of n independent and identically distributed (i.i.d) random variables $X_1, ..., X_n \sim P_0$ where P_0 is the generating distribution. We index a statistical model $\{P_\theta / \theta \in \Theta\}$ by a parameter space Θ, without necessarily assuming that the true distribution P_0 belongs to the model.

Let us consider some integrally strictly positive definite kernel k [1] bounded by a positive constant, say 1. We then denote the associated RKHS $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ satisfying the reproducing property $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$ for any $f \in \mathcal{H}_k$ and any $x \in \mathbb{X}$. We define the notion of *kernel mean embedding*, a Hilbert space embedding that maps probability distributions into the RKHS $\mathcal{H}_k$. Given a distribution P, the kernel mean embedding $\mu_P \in \mathcal{H}_k$ is

$$\mu_P(\cdot) := \mathbb{E}_{X \sim P}[k(X, \cdot)] \in \mathcal{H}_k.$$

Then we define the MMD between two probability distributions P and Q simply as the distance in $\mathcal{H}_k$ between their kernel mean embeddings:

$$\mathbb{D}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}.$$

[1] This means that the positive definite kernel satisfies $\mathbb{E}_{X,Y \sim P}[k(X, Y)] \neq 0$ for any distribution P. This includes the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2 / \gamma^2)$. For this property, and the properties of MMD discussed in this section, we refer the reader to Muandet et al. (2017).

Under the assumptions we made on the kernel, the kernel mean embedding is injective and the maximum mean discrepancy is a metric, see Briol et al. (2019). We motivate the use of MMD as a robust metric in Appendix 9.7.4.

In this chapter, we adopt a Bayesian approach. We introduce a prior distribution π over the parameter space Θ equipped with some sigma-algebra. Then we define our pseudo-Bayesian distribution π_n^β given a prior π on Θ:

$$\pi_n^\beta(d\theta) \propto \exp\left(-\beta \cdot \mathbb{D}_k^2(P_\theta, \widehat{P}_n)\right)\pi(d\theta),$$

where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ is the empirical measure and $\beta > 0$ is a temperature parameter.

9.3 Theoretical analysis of MMD-Bayes

In this section, we show that the MMD-Bayes is consistent when the true distribution belongs to the model, and is robust to misspecification.

To obtain the concentration of posterior distributions in models that contain the generating distribution, Ghosal et al. (2000) introduced the so-called *prior mass condition* that requires the prior to put enough mass to some neighborhood (in Kullback-Leibler divergence) of the true distribution. This condition was widely studied since then for more general pseudo-posterior distributions (Bhattacharya et al., 2016; Alquier and Ridgway, 2017; Chérief-Abdellatif and Alquier, 2018). Unfortunately, this prior mass condition is (by definition) restricted to cases when the model is well-specified or at least when the true distribution is in a very close neighborhood of the model. We formulate here a robust version of the prior mass condition which is based on a neighborhood of an approximation θ^* of the true parameter instead of the true parameter itself. The following condition is suited to the MMD metric, recovers the usual prior mass condition when the model is well-specified and still makes sense in misspecified cases with potentially large deviations to the model assumptions:

Prior mass condition: *Let us denote* $\theta^* = \arg\min_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P_0\right)$ *and its neighborhood* $\mathcal{B}_n = \{\theta \in \Theta / \mathbb{D}_k\left(P_\theta, P_{\theta^*}\right) \leq n^{-1/2}\}$. *Then* (π, β) *is said to satisfy the prior mass condition* $C(\pi, \beta)$ *when* $\pi(\mathcal{B}_n) \geq e^{-\beta/n}$.

In the usual Bayesian setting, the computation of the prior mass is a major difficulty (Ghosal et al., 2000), and it can be hard to know whether the prior mass condition is satisfied or not. Nevertheless, here the condition does not only hold on the prior distribution π but also on the temperature parameter β. Hence, it is always possible to choose β large enough so that the prior mass condition is satisfied. We refer the reader to Appendix 9.7.5 for an example of computation of such a prior mass and valid values of β. The following theorem expressed as a generalization bound shows that the MMD-Bayes posterior distribution is robust to misspecification under the robust prior mass condition. Note that the rate $n^{-1/2}$ is exactly the one obtained by the frequentist MMD estimator of Briol et al. (2019) and is minimax optimal (Tolstikhin et al., 2017):

Theorem 9.3.1. *Under the prior mass condition $C(\pi, \beta)$:*

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right) \pi_n^\beta(\mathrm{d}\theta)\right] \leq 8 \inf_{\theta \in \Theta} \mathbb{D}_k^2\left(P_\theta, P_0\right) + \frac{16}{n}. \tag{9.1}$$

The second theorem investigates concentration of the MMD-Bayes posterior in the well-specified case. It shows that the prior mass condition $C(\pi, \beta)$ ensures that the MMD-Bayes concentrates to P_0 at the minimax rate $n^{-1/2}$:

Theorem 9.3.2. *Let us consider a well-specified model. Then under the prior mass condition $C(\pi, \beta)$, we have in probability for any $M_n \to +\infty$:*

$$\pi_n^\beta\left(\mathbb{D}_k(P_\theta, P_0) > M_n \cdot n^{-1/2}\right) \xrightarrow[n \to +\infty]{} 0. \tag{9.2}$$

Note that we obtain the concentration to the true distribution $P_0 = P_{\theta^*}$ at the minimax rate $n^{-1/2}$ for well-specified models.

9.4 Variational inference

Unfortunately, the MMD-Bayes is not tractable in complex models. In this section, we provide an efficient implementation of the MMD-Bayes based on VI retaining the same theoretical properties. Given a variational set of tractable distributions $\mathcal{F}$, we define the variational approximation of π_n^β as the closest approximation (in KL divergence) to the target MMD posterior:

$$\widetilde{\pi}_n^\beta = \arg\min_{\rho \in \mathcal{F}} \mathrm{KL}(\rho \| \pi_n^\beta).$$

Under similar conditions to those in Theorems 9.3.1 and 9.3.2, $\widetilde{\pi}_n^\beta$ is guaranteed to be $n^{-1/2}$-consistent as the MMD-Bayes. Most works ensuring the consistency or the concentration of variational approximations of posterior distributions use the *extended prior mass condition*, an extension of the prior mass condition that applies to variational approximations rather than on the distributions they approximate (Alquier et al., 2016; Alquier and Ridgway, 2017; Bhattacharya et al., 2018; Chérief-Abdellatif and Alquier, 2018; Chérief-Abdellatif, 2019a,b). Here, we extend our previous prior mass condition to variational approximations but also to misspecification. In addition to the prior mass condition inspired from Ghosal et al. (2000), the variational set $\mathcal{F}$ must contain probability distributions that are concentrated around the best approximation P_{θ^*}. This robust extended prior mass condition can be formulated as follows:

Assumption : *We assume that there exists a distribution $\rho_n \in \mathcal{F}$ such that:*

$$\int \mathbb{D}_k^2(P_\theta, P_{\theta^*}) \rho_n(d\theta) \leq \frac{1}{n} \text{ and } \mathrm{KL}(\rho_n \| \pi) \leq \frac{\beta}{n}. \tag{9.3}$$

Remark 9.4.1. *When the restriction of π to the MMD-ball $\mathcal{B}_n$ centered at θ^* of radius $n^{-1/2}$ belongs to $\mathcal{F}$, then Assumption (9.3) becomes the standard robust prior mass condition, i.e. $\pi(\mathcal{B}_n) \geq e^{-\beta/n}$. In particular, when $\mathcal{F}$ is the set of all probability measures – that is, in the case where there is no variational approximation – then we recover the standard condition.*

Now, we can state the following theorem for variational approximations:

Theorem 9.4.1. *Under the extended prior mass condition* (9.3),

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right) \widetilde{\pi}_n^\beta(\mathrm{d}\theta)\right] \leq 8 \inf_{\theta\in\Theta} \mathbb{D}_k^2\left(P_\theta, P_0\right) + \frac{16}{n}. \tag{9.4}$$

Moreover, if the model is well-specified, then under the prior mass condition $C(\pi, \beta)$*, we have in probability for any* $M_n \to +\infty$*:*

$$\widetilde{\pi}_n^\beta\left(\mathbb{D}_k(P_\theta, P_0) > M_n \cdot n^{-1/2}\right) \xrightarrow[n\to+\infty]{} 0. \tag{9.5}$$

9.5 Numerical experiments

In this section, we show that the variational approximation is robust in practice when estimating a Gaussian mean and a uniform distribution in the presence of outliers. We consider here a d-dimensional parametric model and a Gaussian mean-field variational set $\mathcal{F} = \{\mathcal{N}(m, \mathrm{diag}(s^2))/m \in \mathcal{M}, s \in \mathcal{S}\}$, $\mathcal{M} \subset \mathbb{R}^d, \mathcal{S} \subset \mathbb{R}^d_{>0}$, using componentwise multiplication. Inspired from the stochastic gradient descent of Dziugaite et al. (2015), Li et al. (2015) and Briol et al. (2019) based on a U-statistic approximation of the MMD criterion, we design a stochastic gradient descent that is suited to our variational objective. The algorithm is described in details in Appendix 9.7.7.

We perform short simulations to provide empirical support to our theoretical results. Indeed, we consider the problem of Gaussian mean estimation in the presence of outliers. The experiment consists in randomly sampling $n = 200$ i.i.d observations from a Gaussian distribution $\mathcal{N}(2, 1)$ but some corrupted observations are replaced by samples from a standard Cauchy distribution $\mathcal{C}(0, 1)$. The fraction of outliers used was ranging from 0 to 0.20 with a step-size of 0.025. We repeated each experiment 100 times and considered the square root of the mean square error (MSE). The plots we obtained demonstrate that our method performs comparably to the componentwise median (MED) and even better as the number of outliers increases, and clearly outperforms the maximum likelihood estimator (MLE). We also conducted the simulations for multidimensional Gaussians and for the robust estimation of the location parameter of a uniform distribution. We refer the reader to Appendix 9.7.8 for more details on these simulations.

9.6 Conclusion

In this chapter, we showed that the MMD-Bayes posterior concentrates at the minimax convergence rate and is robust to model misspecification. We also proved that reasonable variational approximations of this posterior retain the same properties, and we proposed a stochastic gradient algorithm to compute such approximations that we supported with numerical simulations. An interesting future line of research would be to investigate if the i.i.d assumption can be relaxed and if the MMD-based estimator is also robust to dependency in the data.

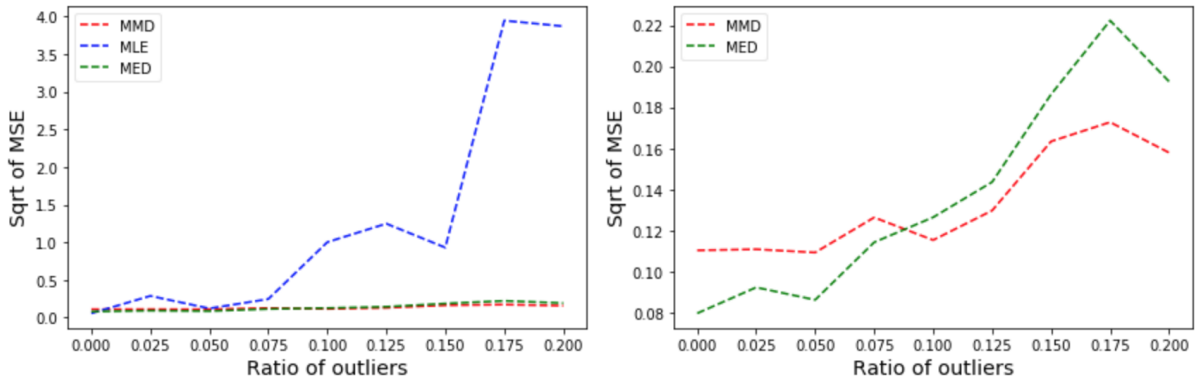


Figure 9.1: Comparison of the square root of the MSE for the MMD estimator, the MLE and the median in the robust Gaussian mean estimation problem for various values of the proportion of outliers. The MMD estimator is the mean of the variational approximation.

9.7 Proofs and additional results

9.7.1 Proof of Theorem 9.3.1.

In order to prove Theorem 9.3.1, we first need two preliminary lemmas. The first one ensures the convergence of the empirical measure $\widehat{P}_n$ to the true distribution P_0 (in MMD distance $\mathbb{D}_k$) at the minimax rate $n^{-1/2}$, and which is an expectation variant of Lemma 1 in Briol et al. (2019) that holds with high probability:

Lemma 9.7.1. *We have*

$$\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P_0\right)\right] \leq \frac{1}{n}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P_0\right)\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left[k(X_i,\cdot) - \mu_{P_0}\right]\right\|_{\mathcal{H}_k}^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\|k(X_i,\cdot) - \mu_{P_0}\|_{\mathcal{H}_k}^2 + 2\sum_{1\leq i<j\leq n}\langle k(X_i,\cdot) - \mu_{P_0}, k(X_j,\cdot) - \mu_{P_0}\rangle_{\mathcal{H}_k}\right] \\
&\leq \frac{1}{n^2}\left(n + 2\sum_{1\leq i<j\leq n} 0\right) = \frac{1}{n}.
\end{aligned}$$

□

The rate $n^{-1/2}$ is known to be minimax in this case, see Theorem 1 in Tolstikhin et al. (2017).

The second lemma is a simple triangle-like inequality that will be widely used throughout the proofs of the paper:

Lemma 9.7.2. *We have for any distributions P, P' and Q:*

$$\mathbb{D}_k^2(P, P') \leq 2\mathbb{D}_k^2(P, Q) + 2\mathbb{D}_k^2(Q, P').$$

Proof. The chain of inequalities follow directly from the triangle inequality and inequality $2ab \leq a^2 + b^2$.

$$\begin{aligned}
\mathbb{D}_k^2(P, P') &\leq \Big(\mathbb{D}_k(P, Q) + \mathbb{D}_k(Q, P')\Big)^2 \\
&= \mathbb{D}_k^2(P, Q) + \mathbb{D}_k^2(Q, P') + 2\mathbb{D}_k(P, Q)\,\mathbb{D}_k(Q, P') \\
&\leq \mathbb{D}_k^2(P, Q) + \mathbb{D}_k^2(Q, P') + \mathbb{D}_k^2(P, Q) + \mathbb{D}_k^2(Q, P') \\
&= 2\mathbb{D}_k^2(P, Q) + 2\mathbb{D}_k^2(Q, P').
\end{aligned}$$

□

Let us come back to the proof of Theorem 9.3.1. An important point is that the MMD-Bayes can also be defined using an argmin over the set $\mathcal{M}_+^1(\Theta)$ of all probability distributions absolutely continuous with respect to π and the Kullback-Leibler divergence $\mathrm{KL}(\cdot\|\cdot)$:

$$\pi_n^\beta = \operatorname*{arg\,min}_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int \mathbb{D}_k^2(P_\theta, \widehat{P}_n)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta} \right\}.$$

This is an immediate consequence of Donsker and Varadhan's variational inequality, see e.g Catoni (2007). Using the triangle inequality, Lemma 9.7.1, Lemma 9.7.2 for different settings of P, P' and Q, and Jensen's inequality:

$$\begin{aligned}
\mathbb{E}\left[\int \mathbb{D}_k^2(P_\theta, P_0)\,\pi_n^\beta(\mathrm{d}\theta)\right] &\leq 2\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\pi_n^\beta(\mathrm{d}\theta)\right] + 2\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P_0\right)\right] \\
&\leq 2\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\pi_n^\beta(\mathrm{d}\theta)\right] + \frac{2}{n} \\
&\leq 2\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\pi_n^\beta(\mathrm{d}\theta) + \frac{\mathrm{KL}(\pi_n^\beta\|\pi)}{\beta}\right] + \frac{2}{n} \\
&= 2\mathbb{E}\left[\inf_\rho \left\{\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right\}\right] + \frac{2}{n} \\
&\leq 2\inf_\rho \mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{2}{n},
\end{aligned}$$

which gives, using Lemma 9.7.1 and the triangle inequality again:

$$\begin{aligned}
\mathbb{E}\left[\int \mathbb{D}_k^2(P_\theta, P_0)\,\pi_n^\beta(\mathrm{d}\theta)\right] &\leq 2\inf_\rho \mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{2}{n} \\
&\leq 2\inf_\rho \mathbb{E}\left[\int \mathbb{D}_k^2(P_\theta, P_0)\,\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + 4\mathbb{E}\left[\mathbb{D}_k^2\left(\widehat{P}_n, P_0\right)\right] + \frac{2}{n} \\
&= 2\inf_\rho \mathbb{E}\left[2\int \mathbb{D}_k^2(P_\theta, P_0)\,\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{6}{n}
\end{aligned}$$

$$\leq 8\mathbb{D}_k^2\left(P_{\theta^*}, P_0\right) + 2\inf_{\rho} \mathbb{E}\left[4\int \mathbb{D}_k^2\left(P_\theta, P_{\theta^*}\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{6}{n}$$

We remind that $\theta^* = \arg\min_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P_0\right)$.

This bound can be formulated in the following way when ρ is chosen to be equal to π restricted to $\mathcal{B}_n$:

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right)\pi_n^\beta(\mathrm{d}\theta)\right] \leq 8\inf_{\theta\in\Theta} \mathbb{D}_k^2\left(P_\theta, P_0\right) + \frac{8}{n} + 2\frac{-\log\pi(B)}{\beta} + \frac{6}{n}.$$

Finally, as soon as the prior mass condition $C(\pi, \beta)$ is satisfied, we get:

$$\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right)\pi_n^\beta(\mathrm{d}\theta)\right] \leq 8\inf_{\theta\in\Theta} \mathbb{D}_k^2\left(P_\theta, P_0\right) + \frac{16}{n}.$$

9.7.2 Proof of Theorem 9.3.2.

In case of well-specification, Formula (9.1) simply becomes according to Jensen's inequality:

$$\mathbb{E}\left[\int \mathbb{D}_k\left(P_\theta, P_0\right)\pi_n^\beta(\mathrm{d}\theta)\right] \leq \sqrt{\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right)\pi_n^\beta(\mathrm{d}\theta)\right]} \leq \sqrt{\frac{16}{n}} = \frac{4}{\sqrt{n}}.$$

Hence, it is sufficient to show that the inequality above implies the concentration of the MMD-Bayes to the true distribution. This is a simple consequence of Markov's inequality. Indeed, for any $M_n \to +\infty$:

$$\mathbb{E}\left[\pi_n^\beta\left(\mathbb{D}_k(P_\theta, P_0) > M_n\cdot n^{-1/2}\right)\right] \leq \frac{\mathbb{E}\left[\int \mathbb{D}_k(P_\theta, P_0)\pi_n^\beta(d\theta)\right]}{M_n\cdot n^{-1/2}} \leq \frac{4n^{-1/2}}{M_n\cdot n^{-1/2}} \xrightarrow[n\to+\infty]{} 0,$$

which guarantees the convergence in mean of $\pi_n^\beta\left(\mathbb{D}_k(P_\theta, P_0) > M_n\cdot n^{-1/2}\right)$ to 0, which leads to the convergence in probability of $\pi_n^\beta\left(\mathbb{D}_k(P_\theta, P_0) > M_n\cdot n^{-1/2}\right)$ to 0, i.e. the concentration of MMD-Bayes to P_0 at rate $n^{-1/2}$.

9.7.3 Proof of theorem 9.4.1.

Formula (9.4) can be proven easily as for the proof of Theorem 9.3.1. Indeed, we use the expression of the variational approximation of the MMD-Bayes using an argmin over the set $\mathcal{F}$:

$$\widetilde{\pi}_n^\beta = \arg\min_{\rho\in\mathcal{F}}\left\{\int \mathbb{D}_k^2(P_\theta, \widehat{P}_n)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right\}.$$

This is yet an application of Donsker and Varadhan's lemma. Then, as previously:

$$\begin{aligned}\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right)\widetilde{\pi}_n^\beta(\mathrm{d}\theta)\right] &\leq \mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\widetilde{\pi}_n^\beta(\mathrm{d}\theta)\right] + \frac{2}{n} \text{ by Lemma 9.7.1} \\ &\leq 2\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\widetilde{\pi}_n^\beta(\mathrm{d}\theta) + \frac{\mathrm{KL}(\pi_n^\beta\|\pi)}{\beta}\right] + \frac{2}{n}\end{aligned}$$

$$
\begin{aligned}
&= 2\mathbb{E}\left[\inf_{\rho}\left\{\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right\}\right] + \frac{2}{n} \\
&\leq 2\inf_{\rho}\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, \widehat{P}_n\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{2}{n} \\
&\leq 2\inf_{\rho}\mathbb{E}\left[2\int \mathbb{D}_k^2\left(P_\theta, P_0\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{6}{n} \\
&\leq 8\mathbb{D}_k^2\left(P_{\theta^*}, P_0\right) + 2\inf_{\rho}\mathbb{E}\left[4\int \mathbb{D}_k^2\left(P_\theta, P_{\theta^*}\right)\rho(\mathrm{d}\theta) + \frac{\mathrm{KL}(\rho\|\pi)}{\beta}\right] + \frac{6}{n}.
\end{aligned}
$$

Hence, under the extended prior mass condition (9.3), we have directly:

$$
\mathbb{E}\left[\int \mathbb{D}_k^2\left(P_\theta, P_0\right)\widetilde{\pi}_n^\beta(\mathrm{d}\theta)\right] \leq 8\inf_{\theta\in\Theta}\mathbb{D}_k^2\left(P_\theta, P_0\right) + \frac{16}{n}.
$$

The proof of Formula (9.5) follows the lines of the proof of Theorem 9.3.2.

9.7.4 An example of robustness of the MMD distance.

In this appendix, we try to give some intuition on the choice of MMD-Bayes rather than the classical regular Bayesian distribution. To do so, we show a simple misspecified example for which the MMD distance is more suited than the classical Kullback-Leibler (KL) divergence used in the Bayes rule in the definition of the classical Bayesian posterior.

We consider the Huber's contamination model described as follows. We observe a collection of random variables $X_1, ..., X_n$. There are unobserved i.i.d random variables $Z_1, ..., Z_n \sim \mathrm{Ber}(\epsilon)$ and a distribution Q, such that the distribution of X_i given $Z_i = 0$ is a Gaussian $\mathcal{N}(\theta^0, \sigma^2)$ where the distribution of X_i given $Z_i = 0$ is Q. The observations X_i's are independent. This is equivalent to considering a true distribution $P_0 = (1-\epsilon)\mathcal{N}(\theta^0, \sigma^2) + \epsilon Q$. Here, $\epsilon \in (0, 1/2)$ is the contamination rate, σ^2 is a known variance and Q is the contamination distribution that is taken here as $\mathcal{N}(\theta_c, \sigma^2)$, where θ_c is the mean of the corrupted observations. The true parameter of interest is θ^0 and the model is composed Gaussian distributions $\{P_\theta = \mathcal{N}(\theta, \sigma^2)/\theta \in \mathbb{R}^d\}$. The goal in this appendix is to show that we exactly recover the true parameter θ^0 with the minimizer of the MMD distance to the true distribution P_0, whereas it is not the case with the KL divergence. We use a Gaussian kernel $k(x, y) = \exp(-\|x-y\|^2/\gamma^2)$.

Computation of the MMD distance to the true distribution:

We have remind that $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ where $\theta \in \Theta = \mathbb{R}^d$. For independent X and Y following respectively P_θ and $P_{\theta'}$, we get $(X - Y) \sim \mathcal{N}(\theta - \theta', \sigma^2 I_d)$. Hence,

$$
\frac{X-Y}{\sqrt{2\sigma^2}} \sim \mathcal{N}\left(\frac{\theta-\theta'}{\sqrt{2\sigma^2}}, I_d\right)
$$

and the square of this random variable is a noncentral chi-square random variable:

$$
\frac{\|X-Y\|^2}{2\sigma^2} \sim \chi^2\left(d, \frac{\|\theta-\theta'\|^2}{2\sigma^2}\right).
$$

It is known that for $U \sim \chi^2(d, m)$, we have $\mathbb{E}[\exp(tU)] = \exp(mt/(1-2t))/(1-2t)^{d/2}$, and then $t = -(2\sigma^2)/\gamma^2$ gives:

$$\begin{aligned}\langle \mu_{P_\theta}, \mu_{P_{\theta'}} \rangle_{\mathcal{H}_k} &= \mathbb{E}_{X\sim P_\theta, Y \sim P_{\theta'}} \left[\exp\left(-\frac{\|X-Y\|^2}{\gamma^2} \right) \right] \\ &= \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2} \right).\end{aligned}$$

Thus,

$$\langle \mu_{P_\theta}, \mu_{P_\theta} \rangle_{\mathcal{H}_k} = \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}},$$

$$\begin{aligned}\langle \mu_{P_\theta}, \mu_{P_0} \rangle_{\mathcal{H}_k} &= (1-\epsilon)\langle \mu_{P_\theta}, \mu_{P_{\theta^0}} \rangle_{\mathcal{H}_k} + \epsilon \langle \mu_{P_\theta}, \mu_{P_{\theta_c}} \rangle_{\mathcal{H}_k} \\ &= (1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta^0\|^2}{4\sigma^2+\gamma^2} \right) \\ &\qquad + \epsilon \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta_c\|^2}{4\sigma^2+\gamma^2} \right),\end{aligned}$$

and

$$\begin{aligned}\langle \mu_{P_0}, \mu_{P_0} \rangle_{\mathcal{H}_k} &= (1-\epsilon)^2 \langle \mu_{P_{\theta^0}}, \mu_{P_{\theta^0}} \rangle_{\mathcal{H}_k} + 2\epsilon(1-\epsilon)\langle \mu_{P_{\theta^0}}, \mu_{P_{\theta_c}} \rangle_{\mathcal{H}_k} + \epsilon^2 \langle \mu_{P_{\theta_c}}, \mu_{P_{\theta_c}} \rangle_{\mathcal{H}_k} \\ &= (1-\epsilon)^2 \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} + \epsilon^2 \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \\ &\qquad + 2\epsilon(1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta^0-\theta_c\|^2}{4\sigma^2+\gamma^2} \right) \\ &= \Big(1 - 2\epsilon(1-\epsilon)\Big) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \\ &\qquad + 2\epsilon(1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta^0-\theta_c\|^2}{4\sigma^2+\gamma^2} \right).\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{D}_k^2(P_0, P_\theta) &= \|\mu_{P_\theta} - \mu_{P_0}\|_{\mathcal{H}_k}^2 = \langle \mu_{P_\theta}, \mu_{P_\theta} \rangle_{\mathcal{H}_k} - 2\langle \mu_{P_\theta}, \mu_{P_0} \rangle_{\mathcal{H}_k} + \langle \mu_{P_0}, \mu_{P_0} \rangle_{\mathcal{H}_k} \\ &= 2\Big(1 - \epsilon(1-\epsilon)\Big) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} - 2\epsilon \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta_c\|^2}{4\sigma^2+\gamma^2} \right) \\ &\qquad + 2\epsilon(1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta^0-\theta_c\|^2}{4\sigma^2+\gamma^2} \right) \\ &\qquad - 2(1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta^0\|^2}{4\sigma^2+\gamma^2} \right) \\ &= 2(1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2} \right)^{\frac{d}{2}} \left[1 - \exp\left(-\frac{\|\theta-\theta^0\|^2}{4\sigma^2+\gamma^2} \right) \right]\end{aligned}$$

$$+ 2\epsilon \left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}} \left[1 - \exp\left(-\frac{\|\theta-\theta_c\|^2}{4\sigma^2+\gamma^2}\right)\right]$$
$$- 2\epsilon(1-\epsilon) \left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}} \left[1 - \exp\left(-\frac{\|\theta^0-\theta_c\|^2}{4\sigma^2+\gamma^2}\right)\right].$$

Hence, the minimizer of $\mathbb{D}_k\left(P_0, P_\theta\right)$ w.r.t θ, i.e the maximizer of:

$$(1-\epsilon)\exp\left(-\frac{\|\theta-\theta^0\|^2}{4\sigma^2+\gamma^2}\right) + \epsilon\exp\left(-\frac{\|\theta-\theta_c\|^2}{4\sigma^2+\gamma^2}\right).$$

is θ^0 itself as $\epsilon \leq 1/2$.

Computation of the KL divergence to the true distribution:

In this case, easy computations lead for any θ to:

$$\begin{aligned}
\mathrm{KL}(P_0\|P_\theta) &= \mathrm{KL}\Big((1-\epsilon)\mathcal{N}(\theta^0,\sigma^2) + \epsilon\mathcal{N}(\theta_c,\sigma^2)\|\mathcal{N}(\theta,\sigma^2)\Big) \\
&= C + (1-\epsilon)H(\theta^0\|\theta) + \epsilon H(\theta_c\|\theta) \\
&= C + \frac{d\log(2\pi\sigma^2)}{2} + \frac{d\sigma^2}{2} + (1-\epsilon)\frac{\|\theta-\theta^0\|^2}{2\sigma^2} + \epsilon\frac{\|\theta-\theta_c\|^2}{2\sigma^2},
\end{aligned}$$

where

$$\begin{aligned}
H(\theta'\|\theta) &= -\int \log\Big(\mathcal{N}(x|\theta,\sigma^2)\Big)\mathcal{N}(x|\theta',\sigma^2)dx \\
&= \frac{d\log(2\pi\sigma^2)}{2} + \frac{d\sigma^2}{2} + \frac{\|\theta-\theta'\|^2}{2\sigma^2}
\end{aligned}$$

is the cross-entropy of P_θ and $P_{\theta'}$, and

$$\begin{aligned}
C = (1-\epsilon)\int \log\Big((1-\epsilon)\mathcal{N}(x|\theta^0,\sigma^2) + \epsilon\mathcal{N}(x|\theta_c,\sigma^2)\Big)\mathcal{N}(x|\theta^0,\sigma^2)dx \\
+ \epsilon\int \log\Big((1-\epsilon)\mathcal{N}(x|\theta^0,\sigma^2) + \epsilon\mathcal{N}(x|\theta_c,\sigma^2)\Big)\mathcal{N}(x|\theta_c,\sigma^2)dx,
\end{aligned}$$

where $\mathcal{N}(x|m,\sigma^2)$ is the probability density function of $\mathcal{N}(m,\sigma^2)$ evaluated at x.

Hence, the minimizer of $\mathrm{KL}\left(P_0\|P_\theta\right)$ w.r.t θ, i.e the minimizer of:

$$(1-\epsilon)\|\theta-\theta^0\|^2 + \epsilon\|\theta-\theta_c\|^2.$$

is $(1-\epsilon)\theta^0 + \epsilon\theta_c$, which can be far away from θ^0 in situations when the corrupted mean θ_c is very far from the true parameter θ^0.

9.7.5 An example of computation of a robust prior mass.

In this appendix, we tackle the computation of a prior mass in the Gaussian mean estimation problem, and we show that it leads to a wide range of values of β satisfying the prior mass condition $C(\pi, \beta)$ for a standard normal prior π.

We recall that the prior mass condition $C(\pi, \beta)$ is satisfied as soon as there exists a function f such that:

$$\beta \geq -\log \pi(\mathcal{B}_n)n.$$

In practice, lower bounds of the form $\pi(\mathcal{B}_n) \geq \mathcal{L}e^{-f(\theta^*)}$ naturally appear when computing the prior mass $\pi(\mathcal{B}_n)$. Only $f(\theta^*)$ depends on the parameter θ^* corresponding to the best approximation in the model of the true distribution in the MMD sense, that is the true parameter itself when the model is well-specified. Hence, it is sufficient to choose a value of the temperature parameter $\beta \geq \Big(f(\theta^*) - \log \mathcal{L}\Big)n$ in order to obtain the prior mass condition.

We conduct the computation in a misspecified case, where we assume that a proportion $1-\epsilon$ of the observations are sampled i.i.d from a σ^2-variate Gaussian distribution of interest P_{θ^0}, but that the remaining observations are corrupted and can take any arbitrary value. We consider the model of Gaussian distributions $\{P_\theta = \mathcal{N}(\theta, \sigma^2)/\theta \in \mathbb{R}^d\}$. This adversarial contamination model is more general than Huber's contamination model presented in Appendix 9.7.4. Note that when $\epsilon = 0$, then the model is well-specified and the distribution of interest P_{θ^0} is also the true distribution P_0. We use the Gaussian kernel $k(x, y) = \exp(-\|x-y\|^2/\gamma^2)$ and the standard normal prior $\pi = \mathcal{N}(0, I_d)$.

We write the inequality defining parameters θ belonging to $\mathcal{B}_n$:

$$\mathbb{D}_k^2\left(P_{\theta^*}, P_\theta\right) \leq n^{-1}. \tag{9.6}$$

Note that when the model is well-specified, the we get $\theta^* = \theta^0$.

According to derivations performed in Appendix 9.7.4, we have for any θ:

$$\begin{aligned}\mathbb{D}_k^2\left(P_\theta, P_{\theta^*}\right) &= \langle \mu_{P_\theta}, \mu_{P_\theta}\rangle_{\mathcal{H}_k} - 2\langle \mu_{P_\theta}, \mu_{P_{\theta^*}}\rangle_{\mathcal{H}_k} + \langle \mu_{P_{\theta^*}}, \mu_{P_{\theta^*}}\rangle_{\mathcal{H}_k} \\ &= 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left[1 - \exp\left(-\frac{\|\theta-\theta^*\|^2}{4\sigma^2+\gamma^2}\right)\right].\end{aligned}$$

Hence, Inequality (9.6) is equivalent to:

$$2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left[1 - \exp\left(-\frac{\|\theta-\theta^*\|^2}{4\sigma^2+\gamma^2}\right)\right] \leq \frac{1}{n}$$

i.e to

$$1 - \frac{1}{2n}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}} \leq \exp\left(-\frac{\|\theta-\theta^*\|^2}{4\sigma^2+\gamma^2}\right)$$

We denote $s_n = \sqrt{\frac{4\sigma^2+\gamma^2}{2n}}\left(1+\frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{4}}$ and $\mathbb{B}(\theta, s_n)$ the ball of radius s_n and centered at θ. Let us compute the prior mass of $\mathcal{B}_n$:

$$\pi(\mathcal{B}_n) = \pi\left(1 - \frac{1}{2n}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}} \leq \exp\left(-\frac{\|\theta-\theta^*\|^2}{4\sigma^2+\gamma^2}\right)\right)$$

$$
\begin{aligned}
&\geq \pi\left(1 - \frac{1}{2n}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}} \leq 1 - \frac{\|\theta - \theta^*\|^2}{4\sigma^2 + \gamma^2}\right) \qquad \text{using inequality } e^{-x} \geq 1 - x \\
&= \pi\left(\|\theta - \theta^*\|^2 \leq (4\sigma^2 + \gamma^2)\frac{1}{2n}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}}\right) \\
&= \pi\Big(\theta \in \mathbb{B}(\theta^*, s_n)\Big) \\
&= \int_{\mathbb{B}(\theta^*, s_n)} (2\pi)^{-d/2} e^{-\|\theta\|^2/2} d\theta.
\end{aligned}
$$

Actually, the point that minimizes $\theta \to e^{-\|\theta\|^2/2}$ on $\mathbb{B}(\theta^*, s_n)$ is $\theta^*(1 + s_n/\|\theta^*\|)$. Thus:

$$
\begin{aligned}
\pi(\mathcal{B}_n) &\geq \int_{\mathbb{B}(\theta^*, s_n)} (2\pi)^{-d/2} \exp\left(\frac{-\|\theta\|^2}{2}\right) d\theta \\
&\geq (2\pi)^{-d/2} \exp\left(\frac{-(\|\theta^*\| + s_n)^2}{2}\right) vol\Big(\mathbb{B}(\theta^*, s_n)\Big).
\end{aligned}
$$

We recall the formula of the volume of the d-dimensional ball:

$$
vol\Big(\mathbb{B}(\theta^*, s_n)\Big) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} s_n^d.
$$

Hence:

$$
\pi(\mathcal{B}_n) \geq \frac{\left(4\sigma^2 + \gamma^2\right)^{\frac{d}{2}} \left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d^2}{4}}}{\Gamma(d/2 + 1)} \exp\left(- \frac{1}{2}\left\{\|\theta^*\| + \sqrt{\frac{4\sigma^2 + \gamma^2}{2n}}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{4}}\right\}^2\right) \frac{1}{n^{d/2}}.
$$

As could be expected for a standard normal prior, the larger the value of $\|\theta^*\|$, the smaller can be the prior mass.

We denote

$$
\mathcal{L} = \frac{\left(4\sigma^2 + \gamma^2\right)^{\frac{d}{2}} \left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d^2}{4}}}{\Gamma(d/2 + 1)} \cdot \frac{1}{n^{d/2}}
$$

and

$$
f(x) = \frac{1}{2}\left\{\|x\| + \sqrt{\frac{4\sigma^2 + \gamma^2}{2n}}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{4}}\right\}^2
$$

so that $\pi(\mathcal{B}_n) \geq \mathcal{L} e^{-f(\theta^*)}$.

Hence, for the standard normal prior π, values of β leading to consistency of the MMD-Bayes are:

$$
\begin{aligned}
&\beta \geq (f(\theta^*) - \log \mathcal{L})n \\
&= \frac{n}{2}\left\{\|\theta^*\| + \sqrt{\frac{4\sigma^2 + \gamma^2}{2n}}\left(1 + \frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{4}}\right\}^2 + \frac{dn \log n}{2} \\
&\qquad - \frac{dn}{2}\log(4\sigma^2 + \gamma^2) - \frac{d^2 n}{4}\log\left(1 + \frac{4\sigma^2}{\gamma^2}\right) + n \log \Gamma(d/2 + 1).
\end{aligned}
$$

In particular, when γ^2 is of order d, then using Stirling's approximation, we get a lower bound on the valid values of β of order (up to a logarithmic factor):

$$
n \max\left(\|\theta^*\|^2, d\right) \lesssim \beta.
$$

9.7.6 Computation of the extended prior mass.

The computation of Condition (9.3) is of major interest. We investigate here the case of a Gaussian model $P_\theta = \mathcal{N}(\theta, \sigma^2)$, a Gaussian mean-field variational approximation $\mathcal{F} = \{\mathcal{N}(m, \mathrm{diag}(s^2))/m \in \mathbb{R}^d, s \in \mathbb{R}^d_{>0}\}$, a standard Gaussian prior $\pi = \mathcal{N}(0,1)$ and a Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/\gamma^2)$.

Let us define $\rho_n = \mathcal{N}\left(\theta^*, s^2 I_d\right)$ where $s^2 = \frac{4\sigma^2+\gamma^2}{2dn}\left(1+\frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}}$. Then:

$$\begin{aligned}\mathrm{KL}(\rho_n\|\pi) &= \frac{1}{2}\sum_{j=1}^d \left\{\theta_j^{*2} + s^2 - \log(s^2) - 1\right\} \\ &= \frac{4\sigma^2+\gamma^2}{2dn}\left(1+\frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}} + \frac{d\log(2dn) + \|\theta^*\|^2 - d - d\log(4\sigma^2+\gamma^2)}{2} \\ &\qquad - \frac{d^2}{4}\log\left(1+\frac{4\sigma^2}{\gamma^2}\right),\end{aligned}$$

and

$$\begin{aligned}\int \mathbb{D}_k^2(P_{\theta^*}, P_\theta)\rho_n(d\theta) &= 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left(1 - \int \exp\left(-\frac{\|\theta-\theta^*\|^2}{4\sigma^2+\gamma^2}\right)\rho_n(d\theta)\right) \\ &= 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left(1 - \int e^{-\|\theta\|^2}\mathcal{N}\left(d\theta\middle|0, \frac{s^2}{4\sigma^2+\gamma^2}I_d\right)\right) \\ &= 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left(1 - \mathrm{Det}\left(I_d + 2\frac{s^2}{4\sigma^2+\gamma^2}I_d\right)^{-1/2}\right) \\ &= 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left(1 - \prod_{j=1}^d\left(1 + \frac{2s^2}{4\sigma^2+\gamma^2}\right)^{-1/2}\right) \\ &= 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left(1 - \left(1 + \left(1+\frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}}\frac{1}{dn}\right)^{-d/2}\right) \\ &\leq 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left(1 - \left(1 - \frac{d}{2}\left(1+\frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}}\frac{1}{dn}\right)\right) = \frac{1}{n}.\end{aligned}$$

Hence, the robust extended prior mass condition is satisfied as soon as

$$\begin{aligned}\beta \geq \frac{4\sigma^2+\gamma^2}{2d}\left(1+\frac{4\sigma^2}{\gamma^2}\right)^{\frac{d}{2}} + \frac{n(d\log(2dn) + \|\theta^*\|^2 - d - d\log(4\sigma^2+\gamma^2))}{2} \\ - \frac{d^2 n}{4}\log\left(1+\frac{4\sigma^2}{\gamma^2}\right).\end{aligned}$$

When γ^2 is of order d, this leads to a bound of order (up to a logarithmic factor):

$$n\max\left(\|\theta^*\|^2, d\right) \lesssim \beta,$$

and we recover the bound that we found for the exact MMD-Bayes.

9.7.7 Projected Stochastic Gradient Algorithm for VI.

In this section, we provide details of a stochastic gradient algorithm (PSGAVI) to compute the Gaussian mean-field approximation, with a necessary projection step if $\mathcal{M} \subsetneq \mathbb{R}^d$ and $\mathcal{S} \subsetneq \mathbb{R}^d_{>0}$. We assume that $\mathcal{M} \subset \mathbb{R}^d$ and $\mathcal{S} \subset \mathbb{R}^d_{>0}$ are closed and convex sets so that the orthogonal projection $\Pi_{\mathcal{M}}$ on $\mathcal{M}$ and $\Pi_{\mathcal{S}}$ on $\mathcal{S}$ are well-defined. We choose a standard Gaussian prior $\pi = \mathcal{N}(0, 1)$.

Another important assumption is that the model is generative, i.e that one can easily sample from distributions belonging to the model $\{P_\theta, \theta \in \Theta\}$. The main idea of the algorithm (Dziugaite et al., 2015; Li et al., 2015; Briol et al., 2019) is then to approximate the gradient of the criterion to minimize $\mathrm{KL}(\mathcal{N}(m, \mathrm{diag}(s^2))\|\pi_n^\beta)$ using an unbiased U-statistic estimate based on random samples from the generative model, and to use a projected stochastic gradient algorithm. We recall that we use the componentwise multiplication.

Criterion to minimize:

As explained in Appendix 9.7.6, the optimization program is equivalent to minimizing:

$$\underset{(m,s)\in\mathcal{M}\times\mathcal{S}}{\arg\min} \left\{ \int \mathbb{D}_k^2(P_\theta, \widehat{P}_n)\mathcal{N}(\mathrm{d}\theta|m, \mathrm{diag}(s^2)) + \frac{1}{2\beta}\sum_{j=1}^{d}\left[m_j^2 + s_j^2 - \log(s_j^2) - 1\right]\right\}.$$

We know that:

$$\mathbb{D}_k^2(P_\theta, \widehat{P}_n) = \mathbb{E}_{X,X'\sim P_\theta}[k(X, X')] - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P_\theta}[k(X_i, X)] + \frac{1}{n^2}\sum_{1\le i,j\le n} k(X_i, X_j).$$

Hence, the criterion to minimize is:

$$\begin{aligned} R_n(m, s) := &\int \mathbb{E}_{X,X'\sim P_\theta}[k(X, X')]\mathcal{N}(\mathrm{d}\theta|m, \mathrm{diag}(s^2)) \\ &- \int \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P_\theta}[k(X_i, X)]\mathcal{N}(\mathrm{d}\theta|m, \mathrm{diag}(s^2)) + \frac{1}{2\beta}\sum_{j=1}^{d}\left\{m_j^2 + s_j^2 - \log(s_j^2) - 1\right\}. \end{aligned}$$

Gradient computation:

The first-order gradient algorithm PSGAVI requires the computation of the gradient of the criterion R_n with respect to m and s. In the following, we will use componentwise operations.

The expression of R_n contains two terms that can be written as $\int f(\theta)\mathcal{N}(\mathrm{d}\theta|m, \mathrm{diag}(s^2))$, and the derivative of this expectation can be hard to evaluate. We use the so-called reparameterization trick which is very popular in the variational inference community and approximate the expectation by a stochastic gradient estimator:

$$\int \nabla_m f(m + s\theta)\mathcal{N}(\mathrm{d}\theta|0, I_d) \approx \frac{1}{M}\sum_{k=1}^{M}\nabla_m f(m + s\theta^k)$$

and

$$\int \nabla_s f(m+s\theta)\mathcal{N}(\mathrm{d}\theta|0,I_d) \approx \frac{1}{M}\sum_{k=1}^{M} \nabla_s f(m+s\theta^k)$$

where M denotes the number of samples θ^k drawn from the standard Gaussian.

Hence, the gradients of the criterion are:

$$\nabla_m R_n(m,s) \approx \frac{1}{M}\sum_{k=1}^{M} \nabla_m \mathbb{E}_{X,X'\sim P_{m+s\theta^k}}[k(X,X')] \\ - \frac{1}{M}\frac{2}{n}\sum_{k=1}^{M}\sum_{i=1}^{n} \nabla_m \mathbb{E}_{X\sim P_{m+s\theta^k}}[k(X_i,X)] + \frac{1}{\beta}\cdot m,$$

$$\nabla_s R_n(m,s) \approx \frac{1}{M}\sum_{k=1}^{M} \nabla_s \mathbb{E}_{X,X'\sim P_{m+s\theta^k}}[k(X,X')] \\ - \frac{1}{M}\frac{2}{n}\sum_{k=1}^{M}\sum_{i=1}^{n} \nabla_s \mathbb{E}_{X\sim P_{m+s\theta^k}}[k(X_i,X)] + \frac{1}{\beta}(s-s^{-1}).$$

Moreover, using the log-derivative trick for differentiable log-densities:

$$\nabla_\theta \mathbb{E}_{X,X'\sim P_\theta}[k(X,X')] = 2\mathbb{E}_{X,X'\sim P_\theta}\Big[k(X,X')\nabla_\theta[\log p_\theta(X)]\Big],$$

$$\nabla_\theta \mathbb{E}_{X\sim P_\theta}[k(X_i,X)] = \mathbb{E}_{X\sim P_\theta}\Big[k(X_i,X)\nabla_\theta[\log p_\theta(X)]\Big].$$

Hence, we obtain stochastic gradients using i.i.d samples $(Y_1,\ldots,Y_M)$ from P_θ:

$$\widehat{\nabla_m R_n}(m,s) = \frac{2}{M^2}\sum_{k=1}^{M}\sum_{j=1}^{M}\left\{\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j,Y_\ell) - \frac{1}{n}\sum_{i=1}^{n} k(X_i,Y_j)\right\}\nabla_m[\log p_{m+s\theta^k}(Y_j)] \\ + \frac{1}{\beta}\cdot m$$

and

$$\widehat{\nabla_s R_n}(m,s) = \frac{2}{M^2}\sum_{k=1}^{M}\sum_{j=1}^{M}\left\{\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j,Y_\ell) - \frac{1}{n}\sum_{i=1}^{n} k(X_i,Y_j)\right\}\nabla_s[\log p_{m+s\theta^k}(Y_j)] \\ + \frac{1}{\beta}(s-s^{-1}).$$

Note that when the log-density $\log p_\theta(x)$ is not differentiable, it is often possible to compute the stochastic gradients involving $\theta^1,...,\theta^M$ directly, without using the Monte Carlo samples $Y_1,...,Y_M$. For instance, when the model is a uniform distribution $P_\theta = \mathcal{U}([\theta-a,\theta+a])$ and when the kernel can be written as $k(x,y) = K(x-y)$ for some function K (such as Gaussian kernels), we have:

$$\mathbb{E}_{X,X'\sim P_\theta}[k(X,X')] = \int_{\theta-a}^{\theta+a}\int_{\theta-a}^{\theta+a} K(x-x')dxdx' = \int_{-a}^{+a}\int_{-a}^{+a} K(x-x')dxdx',$$

and

$$\mathbb{E}_{X\sim P_\theta}[k(X_i, X)] = \int_{\theta-a}^{\theta+a} K(x - X_i)dx = \int_{\theta-a-X_i}^{\theta+a-X_i} K(x)dx.$$

Hence,

$$\nabla_m \mathbb{E}_{X,X'\sim P_{m+s\theta^k}}[k(X, X')] = 0,$$
$$\nabla_s \mathbb{E}_{X,X'\sim P_\theta}[k(X, X')] = 0,$$

and

$$\nabla_m \mathbb{E}_{X\sim P_{m+s\theta^k}}[k(X_i, X)] = K(m + s\theta^k + a - X_i) - K(m + s\theta^k - a - X_i),$$

$$\nabla_s \mathbb{E}_{X\sim P_{m+s\theta^k}}[k(X_i, X)] = sK(m + s\theta^k + a - X_i) - sK(m + s\theta^k - a - X_i).$$

PSGAVI algorithm:

The Projected Stochastic Gradient Algorithm for Variational Inference is the following:

Algorithm 13 PSGAVI

Require: A dataset $(X_1, ..., X_n)$, a model $\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, a kernel k, a sequence of steps $(\eta_t)_{t\geq 1}$, a batch size M, a stopping time T, closed and convex sets $\mathcal{M} \subset \mathbb{R}^d$ and $\mathcal{S} \subset \mathbb{R}^d_{>0}$, an initial mean $m^{(0)} \in \mathcal{M}$, an initial covariance matrix $\mathrm{diag}(s^{(T)2})$ where $s^{(0)} \in \mathcal{S}$.

$X_0 \sim \nu_0$

for $t = 1, ..., T$ **do**

draw $(Y_1, \ldots, Y_M)$ i.i.d from $P_{m^{(t-1)}}$

$m^{(t)} = \Pi_{\mathcal{M}} \left(m^{(t-1)} - \eta_t \widehat{\nabla_m R_n}(m^{(t-1)}, s^{(t-1)}) \right)$

$s^{(t)} = \Pi_{\mathcal{S}} \left(s^{(t-1)} - \eta_t \widehat{\nabla_s R_n}(m^{(t-1)}, s^{(t-1)}) \right)$

end for

A theoretical analysis of the algorithm, in the spirit of Chérief-Abdellatif et al. (2019), goes beyond the scope of this chapter and will be the object of future works.

9.7.8 Numerical simulations.

In this section, we provide numerical experiments that support our theoretical results. We studied three different and simple problems: the robust unidimensional Gaussian mean estimation, the robust multidimensional Gaussian mean estimation, and the uniform location parameter estimation.

In each experiment, we compared the mean of the variational approximation of the MMD-Bayes to other estimators: the median estimator and the MLE in the Gaussian mean estimation problem, i.e the componentwise median and the arithmetic mean, and the method of moments and the MLE in the uniform location parameter estimation problem, i.e the arithmetic mean and the average between the largest and the lowest values. We chose a value of β of e^{nd}, a number of Monte-Carlo samples equal to n and a

step-size of $\eta_t = 1/\sqrt{t}$. We used the Gaussian kernel $k(x, y) = e^{-\|x-y\|^2/d}$ where d is the dimension and we repeated each experiment 100 times.

Gaussian mean estimation problem: for both the uni- and the multidimensional cases, we randomly sampled $n = 200$ i.i.d observations from a Gaussian distribution $\mathcal{N}(\theta, I_d)$ where I_d is the identity matrix of dimension d and θ is the vector with all components equal to 2. Some proportion $\epsilon \in [0, 0.2]$ of corrupted observations is replaced by independent samples which components are independently sampled from a standard Cauchy distribution $\mathcal{C}(0, 1)$. We compared the mean of the variational approximation with the MLE (i.e the arithmetic mean) and the componentwise median using the squared root of the MSE.

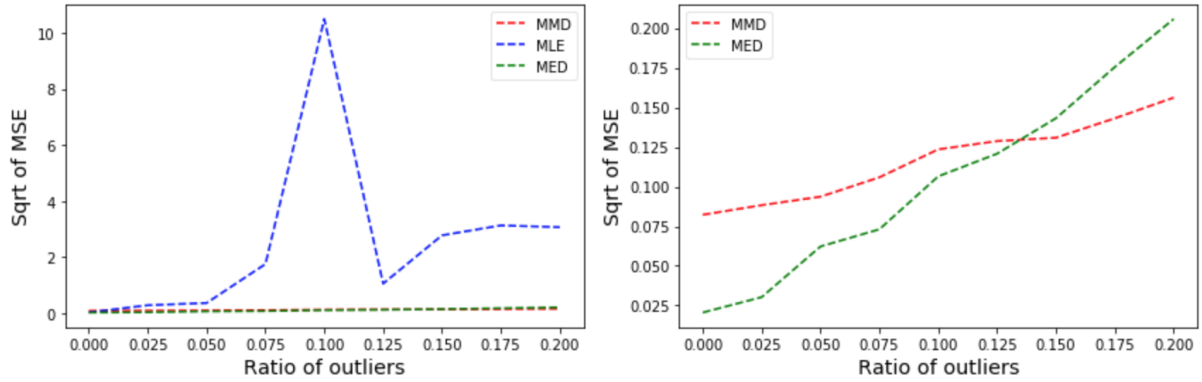


Figure 9.2: Comparison of the square root of the MSE for the MMD estimator, the MLE and the median in the robust multidimensional Gaussian mean estimation problem for various values of the proportion of outliers. Here $d = 15$.

Uniform location parameter estimation problem: we randomly sampled $n = 200$ i.i.d observations from a uniform distribution $\mathcal{U}\left([\theta - \frac{1}{2}, \theta + \frac{1}{2}]\right)$ where $\theta = 1$. Following the previous set of experiments, the proportion $\epsilon \in [0, 0.2]$ of data is replaced by outliers from a Gaussian $\mathcal{N}(20, 1)$. We compared the mean of the variational approximation with the MLE (i.e the average between the largest and the lowest values) and the method of moments estimator (i.e the arithmetic mean) using again the square root of the MSE.

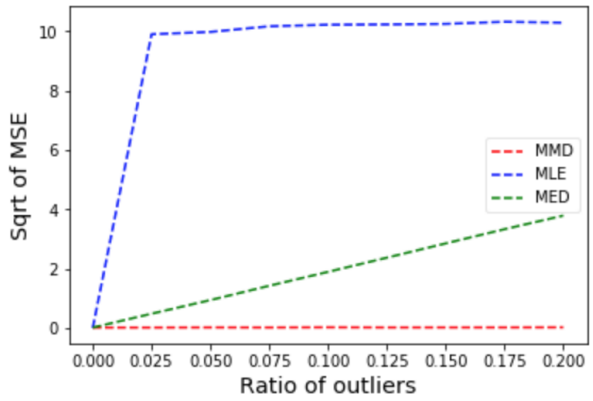


Figure 9.3: Comparison of the square root of the MSE for the MMD estimator, the MLE and the method of moments in the robust estimation of the location parameter of a uniform distribution for various values of the proportion of outliers.

Results: The error of our estimators as a function of the contamination ratio ϵ is plotted in Figures 1, 2 and 3. These plots show that our method is applicable to various problems and leads to a good estimator for all of them. Indeed, the plots in Figures 1 and 2 show that the MSE for the MMD estimator performs as well as the componentwise median and even better when the number of outliers in the dataset increases, much better than the MLE in the robust Gaussian mean estimation problem, and is not affected that much by the presence of outliers in the data. For the uniform location parameter estimation problem addressed in Figure 3, the MMD estimator is clearly the one that performs the best and is not affected by a reasonable proportion of outliers, contrary to the method of moments which square root of MSE is increasing linearly with ϵ and to the MLE that gives inconsistent estimates as soon as there is an outlier in the data.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252, Long Beach, California, USA. PMLR.

Alon, N., Matias, Y., and Szegedy, M. (2008). The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58:1.

Alquier, P. (2008a). Density estimation with quadratic loss: a confidence intervals method. *ESAIM: Probability and Statistics*, 12:438–463.

Alquier, P. (2008b). Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304.

Alquier, P. and Ridgway, J. (2017). Concentration of tempered posteriors and of their variational approximations. *to appear in the Annals of Statistics*.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414.

Ambroladze, A., Parrado-Hernández, E., and Shawe-taylor, J. S. (2007). Tighter pac-bayes bounds. In *Advances in neural information processing systems*, pages 9–16.

Amenta, N., Bern, M., Eppstein, D., and Teng, S. H. (2000). Regression depth and center points. *Discrete and Computational Geometry*, 23(3):305–323.

Andrews, D. (1984). Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4):930–934.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43.

Arlot, S., Celisse, A., and Harchaoui, Z. (2012). A kernel multiple change-point algorithm via model selection. *arXiv preprint arXiv:1202.3878*.

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646.

Ayer, S. and Sawhney, H. (1995). Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. *International Conference on Computer Vision*.

Bacharoglou, A. (2010). Approximation of probability distributions by convex mixtures of Gaussian measures. *Proceedings of the American of the American Mathematical Society*, 138(7):2619–2628.

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58.

Banerjee, A. (2006). On Bayesian bounds. In *Proceedings of ICML*, pages 81–88. ACM.

Banerjee, S., Castillo, I., and Ghosal, S. (2020). Bayesian inference in high-dimensional models.

Baraud, Y. and Birgé, L. (2016). Rho-estimators revisited: General theory and applications. *arXiv preprint arXiv:1605.05051*.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: rho-estimation. *Inventiones mathematicae*, 207(2):425–517.

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. ieee trans. on information theory 39, 930-945. *Information Theory, IEEE Transactions on*, 39:930 – 945.

Barron, A., Schervish, M. J., Wasserman, L., et al. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37(6):1554–1563.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, (53):370–418.

Behrens, G., Friel, N., and Hurn, M. (2012). Tuning tempered transitions. *Statistics and computing*, 22(1):65–78.

Bellec, G., Kappel, D., Maass, W., and Legenstein, R. (2018). Deep rewiring: Training very sparse deep networks. In *International Conference on Learning Representations*.

Bengio, Y. and Delalleau, O. (2011). On the expressive power of deep architectures. In *Proceedings of the 22Nd International Conference on Algorithmic Learning Theory*, ALT'11, pages 18–36, Berlin, Heidelberg. Springer-Verlag.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463.

Bernstein, S. (1917). Theory of probability. *Moskow*.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. (2017). Inference in generative models using the wasserstein distance. *arXiv preprint arXiv:1701.05146*.

Bhattacharya, A., Pati, D., and Yang, Y. (2016). Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125, to appear in the Annals of Statistics*.

Bhattacharya, A., Pati, D., and Yang, Y. (2018). On statistical optimality of variational Bayes. *PMLR: Proceedings of AISTAT*, 84.

Biau, G., Cadre, B., Sangnier, M., and Tanielian, U. (2018). Some theoretical properties of GANs. *arXiv preprint arXiv:1803.07819*.

Bickel, P. J. (1976). Another look at robustness: A review of reviews and some new developments. *Scand J. Statis*, 3:145–168.

Biernacki, C., Celeux, G., and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 65(2):181–237.

Birgé, L. (2006). *Model selection via testing: an alternative to (penalized) maximum likelihood estimators*. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete.

Bishop, C. (1999). Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, volume 1, pages 509–514. IEE.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM.

Blei, D. M., Ng, A., Wang, C., and Jordan, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org.

Boucheron, S., Lugosi, G., and Massart, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614.

Boucheron, S., Lugosi, G., and Massart, P. (2012). *Concentration inequalities*. Oxford University Press.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, 71:52–78.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Breiman, L., LeCam, L., and Schwartz, L. (1964). Consistent estimates and zero-one sets. *The Annals of Mathematical Statistics*, 35(1):157–161.

Briol, F.-X., Barp, A. D., B., A., and Girolami, M. (2019). Statistical inference for generative models via maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. (2013). Streaming variational Bayes. In *NIPS*, pages 1727–1735. Curran Associates, Inc.

Bubeck, S. (2011). Introduction to online optimization. Lecture notes (Princeton University).

Buchholz, A., Wenzel, F., and Mandt, S. (2018). Quasi-Monte Carlo variational inference. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 668–677, Stockholmsmässan, Stockholm Sweden. PMLR.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Sparse density estimation with ℓ_1 penalties. In Conference on Computational Learning Theory.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2010). Spades and mixture models. *The Annals of Statistics*, 38(4):2525–2558.

Cai, T., Ma, Z., and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815.

Campbell, T. and Li, X. (2019). Universal boosting variational inference. *arXiv preprint arXiv:1903.05220*.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108.

Carel, L. and Alquier, P. (2017). Simultaneous dimension reduction and clustering via the nmf-em algorithm. *arXiv preprint arXiv:1709.03346*.

Castillo, I. (2014). Bayesian nonparametrics, convergence and limiting shape of posterior distributions. *Habilitation à diriger des recherches*.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer.

Catoni, O. (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilités et statistiques*, 48(4):1148–1185.

Catoni, O. and Giulini, I. (2017). *Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector*. PAC-Bayesian trends and insights, In NIPS-2017 Workshop (Almost) 50 Shades of Bayesian Learning.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. E. (2018). *Handbook of Mixture Analysis*. CRC Press.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Challis, E. and Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *The Journal of Machine Learning Research*, 14(1):2239–2286.

Chambaz, A., Garivier, A., and Gassiat, E. (2009). A minimum description length approach to hidden markov models with poisson and gaussian emissions. application to order identification. *Journal of Statistical Planning and Inference*, 139(3):962–977.

Chan, M. T. (2004). An optimal randomized algorithm for maximum tukey depth. In Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates. *arXiv preprint arXiv:1902.01998*.

Chérief-Abdellatif, B.-E. (2019a). Consistency of ELBO maximization for model selection. In Ruiz, F., Zhang, C., Liang, D., and Bui, T., editors, *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96 of *Proceedings of Machine Learning Research*, pages 11–31. PMLR.

Chérief-Abdellatif, B.-E. (2019b). Convergence rates of variational inference in sparse deep learning. *arXiv preprint arXiv:1908.04847v2*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *arXiv preprint arXiv:1912.05737*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via Maximum Mean Discrepancy. In Zhang, C., Ruiz, F., Bui, T., Dieng, A. B., and Liang, D., editors, *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–21. PMLR.

Chérief-Abdellatif, B.-E., Alquier, P., and Khan, M. E. (2019). A generalization bound for online variational inference. In Lee, W. S. and Suzuki, T., editors, *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 662–677, Nagoya, Japan. PMLR.

Chinot, G., Lecué, G., and Lerasle, M. (2019). Robust statistical learning with lipschitz and convex loss functions. *Probability Theory and Related Fields, 0*, 178:1–44.

Collier, O. and Dalalyan, A. S. (2017). Minimax estimation of a p-dimensional linear functional in sparse Gaussian models and robust estimation of the mean. submitted 1712.05495, arXiv.

Cottet, V. and Alquier, P. (2018). 1-bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Machine Learning*, 107(3):579–603.

Cuong, N. V., Ho, L. S. T., and Dinh, V. (2013). Generalization and robustness of batched weighted average algorithm with V-geometrically ergodic Markov data. In *International Conference on Algorithmic Learning Theory*, pages 264–278. Springer.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314.

Dai, B., He, N., Dai, H., and Song, L. (2016). Provable Bayesian inference via particle mirror descent. In *AISTAT*, pages 985–994.

Dalalyan, A. S., Grappin, E., Paris, Q., et al. (2018). On the exponentially weighted aggregate with the laplace prior. *The Annals of Statistics*, 46(5):2452–2478.

Dalalyan, A. S. and Sebbar, M. (2017). Optimal kullback-leibler aggregation in mixture density estimation by maximum likelihood. *arXiv preprint arXiv:1701.05009*.

Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In Bshouty, N. and Gentile, C., editors, *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 97–111. Springer Berlin Heidelberg.

Deb, P., Gallo, W., Ayyagari, P., Fletcher, J., and Sindelar, J. (2011). The effect of job loss on overweight and drinking. *Journal of Health Economics*.

Dedecker, J., Doukhan, P., Lang, G., Rafael, L. R. J., Louhichi, S., and Prieur, C. (2007). *Weak dependence: With examples and applications*. Springer, New York.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Depersin, J. and Lecué, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2018a). Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, K., and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium*.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2018b). List-decodable robust mean estimation and learning mixtures of spherical gaussians. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing.

Diakonikolas, I., Kong, W., and Stewart, A. (2018c). Efficient algorithms and lower bounds for robust linear regression. *arXiv preprint arXiv:1806.00040*.

Do, M. N. (2003). Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, 10(4):115–118.

Domke, J. (2019). Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, pages 328–337.

Donoho, D. L. and Liu, R. C. (1988a). The automatic robustness of minimum distance functionals. *The Annals of Statistics*, pages 552–586.

Donoho, D. L. and Liu, R. C. (1988b). Pathologies of some minimum distance estimators. *The Annals of Statistics*, pages 587–608.

Doob, J. L. (1949). Application of the theory of martingales. *Le Calcul des Probabilités et ses Applications. Colloques Internationaux du CNRS*, (13):23–27.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A. and Johansen, A. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12.

Doukhan, P. (1994). *Mixing: properties and examples*, volume 85. Springer, lecture notes in statistics.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685, Long Beach, California, USA. PMLR.

Du, S. S., Balakrishnan, S., and Singh, A. (2017). Computationally efficient robust estimation of sparse functionals. *arXiv preprint arXiv:1702.07709*.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.

Forth, S., Hovland, P., Phipps, E., Utke, J., and Walther, A. (2014). *Recent Advances in Algorithmic Differentiation*. Springer Publishing Company, Incorporated.

Fukumizu, K., Song, L., and Gretton, A. (2013). Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(2013):3002–3048.

Futami, F., Sato, I., and Sugiyama, M. (2018). Variational inference based on robust divergences. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 813–822, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.

Gao, C., Liu, J., Yao, Y., and Zhu, W. (2019). Robust estimation and generative adversarial nets. *ICLR 2019*.

Gassiat, E., Rousseau, J., and Vernet, E. (2018). Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electronic Journal of Statistics*, 12(1):703–740.

Gerchinovitz, S. (2013). Sparsity regret bounds for individual sequences in online linear regression. *The Journal of Machine Learning Research*, 14(1):729–769.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). Consistency issues in bayesian nonparametric asymptotic. In *Nonparametrics and Time Series: A Tribute to Madan Lal Puri.*, pages 639–667. Marcel Dekker, Inc.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Ghosal, S., Van Der Vaart, A., et al. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263.

Ghosh, A. and Basu, A. (2016). Robust Bayes estimation using the density power divergence. *The Annals of Statistics*, pages 500–531.

Giulini, I. (2017). Robust pca and pairs of projections in a hilbert space. *Electronic Journal of Statistics*, 11(2):3903–3926.

Giulini, I. (2018). Robust dimension-free gram operator estimates. *Bernoulli*, 11(2):3864–3923.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Graves, A. (2011). Practical variational inference for neural networks. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. Advances in neural information processing systems, pages 673–681.

Grohs, P., Perekrestenko, D., Elbrächter, D., and Bölcskei, H. (2019). Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44(1):133–152.

Grünwald, P., Van Ommen, T., et al. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *Preprint arXiv:1901.05353*.

Guhaniyogi, R., Willett, R. M., and Dunson, D. B. (2013). Approximated Bayesian inference for massive streaming data. Unpublished manuscript.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774.

Hayakawa, S. and Suzuki, T. (2019). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *arXiv preprint arXiv:1905.09195*.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3–4):157–325.

Hershey, J. and Olsen, P. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4.

Hinton, G. E. and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pages 5–13, New York, NY, USA. ACM.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Holland, M. J. (2019a). Distribution-robust mean estimation via smoothed random perturbations. *arXiv preprint arXiv:1906.10300*.

Holland, M. J. (2019b). Pac-bayes under potentially heavy tails. *arXiv preprint arXiv:1905.07900*.

Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test, 23(3):556–584*.

Hopkins, S. B. (2019). Sub-gaussian mean estimation in polynomial time. *to appear in The Annals of Statistics*.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *JMLR*, 17:1–40.

Hüber, P. J. (1964). Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1):73–101.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. (2018). Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. *arXiv preprint arXiv:1809.09505*.

Imaizumi, M. and Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 869–878. PMLR.

Jaiswal, P., Honnappa, H., and Rao, V. A. (2019a). Risk-sensitive variational bayes: Formulations and bounds. *arXiv preprint arXiv:1906.01235*.

Jaiswal, P., Rao, V. A., and Honnappa, H. (2019b). Asymptotic consistency of α-rényi-approximate posteriors. *arXiv preprint arXiv:1902.01902*.

Jerfel, G. (2017). An information theoretic interpretation of variational inference based on the mdl principle and the bits-back coding scheme.

Jerrum, M., Valiant, L., and Vazirani, V. (1986). Random generation of combinatorial structures from a uniform distribution. Theoretical Computer Science, 43:186–188.

Jewson, J., Smith, J., and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. Advances in Neural Information Processing Systems (NeurIPS), 2018:262–271.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017). A linear-time kernel goodness-of-fit test. Advances in Neural Information Processing Systems, pages 262–271.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kalai, A. and Vempala, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Kawaguchi, K. (2016). Deep learning without poor local minima. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 586–594. Curran Associates, Inc.

Kawaguchi, K., Huang, J., and Kaelbling, L. P. (2019). Effect of depth and width on local minima in deep learning. *Neural Computation*, 31(6):1462–1498.

Khan, M., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. (2018). Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2611–2620, Stockholmsmässan, Stockholm Sweden. PMLR.

Khan, M. E. and Lin, W. (2017). Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR*, 54:878–887.

Khan, M. E. and Nielson, D. (2018). Fast yet simple natural-gradient descent for variational inference in complex models. Invited paper at ISITA 2018, preprint arXiv:1807.04489.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference. *arXiv preprint arXiv:1904.02063*.

Kothari, P. K., Steinhardt, J., and Steurer, D. (2018). Robust moment estimation and improved clustering via sum of squares. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing.

Kruijer, W., Rousseau, J., and Van Der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

Laforgia, A. and Natalin, P. (2013). On some inequalities for the gamma function. *Advances in Dynamical Systems and Applications*, 8(2):261–267.

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium*.

Laplace, P.-S. (1810). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur applications aux probabilités. *Memoires de l'Academie des Sciences de Paris*.

Laplace, P. S. d. (1774). Mémoire sur la probabilité des causes par les évènements. *Mémoires de Mathematique et de Physique, Presentés à l'Académie Royale des Sciences, Par Divers Savans & Lus Dans ses Assemblées*, (6):621–656.

Le Cam, L. (1970). *On the assumptions used to prove asymptotic normality of maximum likelihood estimates*. The Institute of Mathematical Statistics.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)*, pages 13–54.

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.

Lerasle, M., Szabó, Z., Mathieu, T., and Lecué, G. (2019). Monk–outlier-robust mean embedding estimation by median-of-means. In International Conference on Machine Learning.

Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions on*, 52:3396 – 3410.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In International Conference on Machine Learning, pages 1718–1727.

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Information and computation*, 108(2):212–261.

Liu, J., Huang, Y., Singh, R., Vert, J.-P., and Noble, W. (2019). Jointly embedding multiple single-cell omics measurements. *BioRxiv*, page 644310.

Louhichi, S. (1998). *Théorèmes limites pour des suites positivement ou faiblement dépendantes*. Thèse de doctorat de l'Université Paris-XI.

Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0-regularization. In *International Conference on Learning Representations*.

Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions–a survey. *arXiv preprint arXiv:1906.04280*.

Lv, J. and Liu, J. S. (2013). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):141–167.

MacKay, D. J. C. (1992a). *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology.

MacKay, D. J. C. (1992b). A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, Edited by Jean Picard.

McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

McDiarmid, C. (1989). On the method of bounded differences. in. In Siemons, J., editor, *Surveys of Combinatorics*. Mathematical Society Lecture Notes Series 141, London.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33(3):331–373.

Millar, P. W. (1981). Robust estimation via minimum distance methods. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):73–89.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. in. *Bernoulli*, 21:2308–2335.

Mishkin, A., Kunstner, F., Nielsen, D., Schmidt, M., and Khan, M. E. (2018). Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 6245–6255. Curran Associates, Inc.

Moridomi, K., Hatano, K., and Takimoto, E. (2018). Online linear optimization with the log-determinant regularizer. *IEICE Transactions on Information and Systems*, E101D(6):1511–1520.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Nakagawa, T. and Hashimoto, S. (2019). Robust bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, pages 1–18.

Nasios, N. and Bors, A. (2006). Variational learning for Gaussian mixture models. In *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, volume 36, pages 849–862.

Neal, R. M. (1995). *Bayesian learning for neural networks*. PhD thesis, University of Toronto.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Nemirovski, A. and Yudin, B. (1983). *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018). A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*.

Nguyen, C. V., Bui, T. D., Li, Y., and Turner, R. E. (2017a). Online variational Bayesian inference: Algorithms for sparse gaussian processes and theoretical bounds. ICML 2017 Time Series Workshop.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. (2017b). Variational continual learning. *Preprint arXiv:1710.10628*.

Nguyen, Q., Mukkamala, M. C., and Hein, M. (2019). On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*.

O'Hagan, A., Murphy, T. B., and Gormley, I. C. (2012). Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics & Data Analysis*, 56(12):3843–3864.

Opper, M. and Archambeau, C. (2008). The variational gaussian approximation revisited. *Neural computation*, 21:786–92.

Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. (2019). Practical deep learning with bayesian principles. In *Advances in Neural Information Processing Systems*, pages 4289–4301.

Pan, W., Lin, J., and Le, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics*, 3:117–124.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-abc: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 398–407, volume 51, pp. 51.

Parr, W. (1981). Minimum distance estimation: a bibliography. *Communications in Statistics: Theory and Methods*, 10(12):1205–1224.

Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75(371):616–624.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. (2012). Pac-bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(Dec):3507–3531.

Petersen, P. and Voigtländer, F. (2017). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*.

Peyré, G.and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rao, C. R. and Wu, Y. (2001). *On model selection*, volume Volume 38 of *Lecture Notes–Monograph Series*, pages 1–57. Institute of Mathematical Statistics, Beachwood, OH.

Ridgway, J. (2017). Probably approximate Bayesian computation: nonasymptotic convergence of abc under misspecification. *arXiv preprint arXiv:1707.05987v2*.

Ridgway, J., Alquier, P., Chopin, N., and Liang, F. (2014). PAC-Bayesian AUC classification and scoring. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 658–666. Curran Associates Inc.

Rigouste, L., Cappé, O., and Yvon, F. (2007). Inference and evaluation of the multinomial mixture model for text clustering. In *Information Processing & Management*, volume 43, pages 1260–1280.

Rio, E. (2013). On mcdiarmid's concentration inequality. *Electronic Communications in Probability*, 18.

Rio, E. (2017a). *Asymptotic theory of weakly dependent random processes*. Springer, Berlin.

Rio, E. (2017b). Inégalités de hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 330(10):905–908.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rivoirard, V. and Rousseau, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7(2):311–334.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rockova, V. and Polson, n. (2018). Posterior concentration for sparse deep learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 930–941. Curran Associates, Inc.

Rolnick, D. and Tegmark, M. (2018). The power of deeper networks for expressing natural functions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Rosenblatt, M. (1956). A central limit corollary and a strong mixing condition. In *Proc. Natl*, pages 43–47, 42. Acad. Sci. USA.

Rousseau, J. (2016). On the frequentist properties of bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536.

Salmon, J. and Dalalyan, A. (2011). Optimal aggregation of affine estimators. In Kakade, S. M. and von Luxburg, U., editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 635–660, Budapest, Hungary. PMLR.

Sato, M.-A. (2001). Online model selection based on the variational Bayes. *Neural computation*, 13(7):1649–1681.

Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with relu activation function. *arXiv*, arXiv:1708.06633.

Schwartz, L. (1965). On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Seldin, Y., Auer, P., Shawe-Taylor, J., Ortner, R., and Laviolette, F. (2011). Pac-bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems*, pages 1683–1691.

Seldin, Y. and Tishby, N. (2010). PAC–Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(Dec):3595–3646.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Shawe-Taylor, J. and Williamson, R. C. (1997a). A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, page 2–9, New York, NY, USA. Association for Computing Machinery.

Shawe-Taylor, J. and Williamson, R. C. (1997b). A PAC analysis of a Bayesian estimator. In *Tenth annual conference on Computational learning theory*, volume 6, pages 2–9.

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, 97(457):222–235.

Sheth, R. and Khardon, R. (2017). Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5151–5161. Curran Associates, Inc.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550:354–.

Singer, Y. and Warmuth, M. K. (1999). Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, MA.

Song, L. (2008). *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, University of Sydney.

Song, L. and Gretton, A. (2011). and bickson. In *Kernel belief propagation*, pages 707–715, and Low, Y., & Guestrin, C. . *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. D.

Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Stanford, J., Giardina, K., Gerhardt, G., Fukumizu, K., and Amari, S.-i. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13.

Stoneking, C. J. (2014). Bayesian inference of Gaussian mixture models with noninformative priors. *arXiv preprint arXiv:1405.4895*.

Sudderth, E. and Jordan, M. (2009). Shared segmentation of natural scenes using dependent pitman-yor processes. *In Advances in Neural Information Processing Systems*, pages 1585–1592.

Suzuki, T. (2012). PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. In *Conference on Learning Theory*, pages 8–1.

Suzuki, T. (2018). Fast generalization error bound of deep learning from a kernel perspective. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1397–1406, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Suzuki, T. (2019). Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*.

Syring, N. and Martin, R. (2018). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Tat Lee, Y., Song, Z., and Vempala, S. S. (2018). Algorithmic Theory of ODEs and Sampling from Well-conditioned Logconcave Densities. *arXiv e-prints*, page arXiv:1812.06243.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2339–2347. Curran Associates, Inc.

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2017). Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048.

Tonolini, F., Jensen, B. S., and Murray-Smith, R. (2019). Variational sparse coding. *Conference on Uncertainty in Artificial Intelligence*.

Tsuzuku, Y., Sato, I., and Sugiyama, M. (2019). Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. *Preprint arXiv:1901.04653*.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition.

Tukey, J. W. (1975). *Mathematics and the picturing of data*. Proceedings of the International Congress of Mathematicians.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838.

Vehtari, A., Tolvanen, V., Mononen, T., and Winther, O. (2014). Bayesian leave-one-out cross validation approximations for gaussian latent variable models. *Journal of Machine Learning Research*, 17.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). Understanding priors in Bayesian neural networks at the unit level. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6458–6467, Long Beach, California, USA. PMLR.

Von Mises, R. (1931). Wahrscheinlichkeitsrechnung. *Vienna: Deuticke*.

Vovk, V. G. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*.

Walker, S. and Hjort, N. L. (2001). On bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821.

Wang, C., Paisley, J., and Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760.

Wang, Y. and Blei, D. M. (2018). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, pages 1–85.

Watier, L., Richardson, S., and Green, P. (1999). Using gaussian mixtures with unknown number of components for mixed model estimation. *14th International Workshop on Statistical Modeling, Graz, Austria*.

Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, 28(1):75–88.

Wu, Y. and Yang, P. (2018). Optimal estimation of gaussian mixtures via denoised method of moments. *arXiv preprint arXiv:1807.07237*.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950.

Yarotsky, D. (2016). Error bounds for approximations with deep relu networks. *Neural Networks*, 94.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774.

Zeno, C., Golan, I., Hoffer, E., and Soudry, D. (2018). Bayesian gradient descent: Online variational Bayes learning with increased robustness to catastrophic forgetting and weight pruning. *arXiv preprint arXiv:1803.10123*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Zhang, F. and Gao, C. (2017). Convergence rates of variational posterior distributions. *Preprint arXiv:1712.02519v1, Accepted to the Annals of Statistics*.

Zhang, T. (2006). From ϵ-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210.

Zhao, S., Song, J., and Ermon, S. (2017). InfoVAE: Information maximizing variational autoencoders. *arXiv 1706.02262*. arXiv preprint.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 928–935. AAAI Press.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Contributions à l'étude théorique de l'inférence variationelle et à la robustesse

Mots clés : Statistique, Machine learning, Inférence variationnelle, Robustesse

Résumé : Cette thèse de doctorat traite de l'inférence variationnelle et de la robustesse en statistique et en machine learning. Plus précisément, elle se concentre sur les propriétés statistiques des approximations variationnelles et sur la conception d'algorithmes efficaces pour les calculer de manière séquentielle. Elle étudie par ailleurs les estimateurs basés sur le Maximum Mean Discrepancy comme règle d'apprentissage et montre leur robustesse en case de mauvaise spécification du modèle.

Ces dernières années, l'inférence variationnelle a été largement étudiée du point de vue computationnel, cependant, la littérature n'a accordé que peu d'attention à ses propriétés théoriques jusqu'à très récemment. Dans cette thèse, nous étudions la consistence des approximations variationnelles dans divers modèles statistiques et les conditions qui assurent leur consistence. En particulier, nous abordons le cas des modèles de mélange et des réseaux de neurones profonds. Nous justifions également d'un point de vue théorique l'utilisation de la stratégie de maximisation de l'ELBO, un critère numérique qui est largement utilisé dans la communauté VB pour la sélection de modèle et dont l'efficacité a déjà été confirmée en pratique.

En outre, l'inférence Bayésienne offre un cadre d'apprentissage en ligne attrayant pour analyser des données séquentielles, et offre des garanties de généralisation qui restent valables même en cas de mauvaise spécification des modèles et en présence d'adversaires. Malheureusement, l'inférence Bayésienne exacte est rarement tractable en pratique et des méthodes d'approximation sont généralement employées. Ces méthodes préservent-elles les propriétés de généralisation de l'inférence Bayésienne ? Dans cette thèse, nous montrons que c'est effectivement le cas pour certains algorithmes d'inférence variationnelle. Nous proposons de nouveaux algorithmes tempérés en ligne et nous en déduisons des bornes de généralisation. Notre résultat théorique repose sur la convexité de l'objectif variationnel, mais nous soutenons que notre résultat devrait être plus général et présentons des preuves empiriques à l'appui. Notre travail donne des justifications théoriques en faveur des algorithmes en ligne qui s'appuient sur des méthodes Bayésiennes approchées.

Une autre question d'intérêt majeur en statistique et qui est abordée dans cette thèse est la conception d'une procédure d'estimation universelle. Cette question est d'un intérêt majeur, notamment parce qu'elle conduit à des estimateurs robustes, un thème d'actualité en statistique et en machine learning. Nous abordons le problème de l'estimation universelle en utilisant un estimateur de minimisation de distance basé sur le Maximum Mean Discrepancy. Nous montrons que l'estimateur est robuste à la fois à la dépendance et à la présence de valeurs aberrantes dans le jeu de données. Nous mettons également en évidence les liens qui peuvent exister avec les estimateurs de minimisation de distance utilisant la distance L_2. Enfin, nous présentons une étude théorique d'un algorithme de descente de gradient stochastique utilisé pour calculer l'estimateur, et nous étayons nos conclusions par des simulations numériques. Nous proposons également une version Bayésienne de notre estimateur, que nous étudions à la fois d'un point de vue théorique et d'un point de vue computationnel.

Title : Contributions to the theoretical study of variational inference and robustness

Keywords : Statistics, Machine learning, Variational inference, Robustness

Abstract : This PhD thesis deals with variational inference and robustness. More precisely, it focuses on the statistical properties of variational approximations and the design of efficient algorithms for computing them in an online fashion, and investigates Maximum Mean Discrepancy based estimators as learning rules that are robust to model misspecification.

In recent years, variational inference has been extensively studied from the computational viewpoint, but only little attention has been put in the literature towards theoretical properties of variational approximations until very recently. In this thesis, we investigate the consistency of variational approximations in various statistical models and the conditions that ensure the consistency of variational approximations. In particular, we tackle the special case of mixture models and deep neural networks. We also justify in theory the use of the ELBO maximization strategy, a model selection criterion that is widely used in the Variational Bayes community and is known to work well in practice.

Moreover, Bayesian inference provides an attractive online-learning framework to analyze sequential data, and offers generalization guarantees which hold even under model mismatch and with adversaries. Unfortunately, exact Bayesian inference is rarely feasible in practice and approximation methods are usually employed, but do such methods preserve the generalization properties of Bayesian inference? In this thesis, we show that this is indeed the case for some variational inference algorithms. We propose new online, tempered variational algorithms and derive their generalization bounds. Our theoretical result relies on the convexity of the variational objective, but we argue that our result should hold more generally and present empirical evidence in support of this. Our work presents theoretical justifications in favor of online algorithms that rely on approximate Bayesian methods.

Another point that is addressed in this thesis is the design of a universal estimation procedure. This question is of major interest, in particular because it leads to robust estimators, a very hot topic in statistics and machine learning. We tackle the problem of universal estimation using a minimum distance estimator based on the Maximum Mean Discrepancy. We show that the estimator is robust to both dependence and to the presence of outliers in the dataset. We also highlight the connections that may exist with minimum distance estimators using L_2-distance. Finally, we provide a theoretical study of the stochastic gradient descent algorithm used to compute the estimator, and we support our findings with numerical simulations. We also propose a Bayesian version of our estimator, that we study from both a theoretical and a computational points of view.

Institut Polytechnique de Paris
91120 Palaiseau, France