Adaptative Monte-Carlo methods for complex models

Kamélia Daudel

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT024

Thèse de doctorat

Adaptive Monte Carlo Methods for Complex Models

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

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

Thèse présentée et soutenue à Palaiseau, le 14 octobre 2021, par

KAMÉLIA DAUDEL

Composition du Jury :

Gersende Fort Directrice de Recherche, CNRS	Présidente
Ismaël Castillo Professeur, Sorbonne Université	Rapporteur
Arnaud Doucet Professeur, University of Oxford	Rapporteur
Anne Sabourin Maîtresse de conférences, Télécom Paris	Examinatrice
François Roueff Professeur, Télécom Paris	Directeur de thèse
Randal Douc Professeur, Télécom SudParis	Co-directeur de thèse
François Portier Maître de conférences, Télécom Paris	Invité

Adaptive Monte Carlo Methods for Complex Models

KAMÉLIA DAUDEL

October 14, 2021

A ma grand-mère, Messaouda Fellahi

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Randal Douc et François Roueff ainsi qu'à mon encadrant François Portier. Randal, merci pour ton soutien indéfectible et ta patience infinie tout au long de ma thèse. J'ai beaucoup appris à tes côtés, tant mathématiquement qu'humainement, et je me sens très privilégiée d'avoir pu travailler avec toi. A ton contact, j'ai pu croire en mes idées et je ne saurai jamais assez te remercier de m'avoir sans relâche poussée à me dépasser. François R., je te suis très reconnaissante de la bienveillance dont tu as su faire preuve et de la confiance que tu m'as accordée. Ta porte était toujours ouverte en cas de besoin et tes encouragements m'ont été très précieux pour avancer dans cette thèse. François P., merci pour ton enthousiasme et tes connaissances en Importance Sampling, qui ont su motiver mon intérêt pour ce vaste domaine.

Un grand merci à Ismaël Castillo et Arnaud Doucet d'avoir accepté d'être rapporteurs pour ma thèse. Je remercie également chaleureusement Gersende Fort et Anne Sabourin de faire partie de mon jury.

Merci à l'ensemble du département IDS de m'avoir accueillie pendant ces trois années, et notamment merci à Gaël Richard pour avoir toléré mes boutades avec la mansuétude qui lui est propre, à Laurence, Delphine et Janique pour leur gentillesse (et ce malgré les impairs sur les ordres de mission...), à Ons et sa bonne humeur rafraîchissante, à Olivier qui m'a parlé le premier de Mirror Descent et à Robert qui a partagé avec moi son entrain pour les techniques d'optimisation.

Je tiens à remercier les amis qui m'ont accompagnée pendant cette thèse : mon mexicain favori José, mon cher Alberto expert en CNNs et en tortilla, my sweet and caring Ryan (when are we building that castle?), 我高富帅的朋友 Xuxu et Weiwei, Achille le talentueux dessinateur maintenant docteur, Mohammed qui est habilement passé d'IDS à INFRES, Hamid que j'ai (re)découvert pendant la pandémie, le solide trio Valentine, Tich Bao et Nicolas que je connais depuis près de 10 ans, mon binôme de Thessaloniki David, Dan which should be almost fluent in French by now...

Pour finir, merci à ma famille, et tout particulièrement à ma mère, de veiller sur moi affectueusement. Leur optimisme, leur humanité et leur générosité sans faille sont une source constante d'inspiration.

Contents

Abstract

This thesis lies in the field of Statistical Inference and more precisely in Bayesian Inference, where the goal is to model a phenomenon given some observed data while taking into account prior knowledge on the model parameters.

The availability of large datasets sparked the interest in using complex models for Bayesian Inference tasks that are able to capture potentially complicated structures inside the data. Such a context requires the development and study of adaptive algorithms that can efficiently process large volumes of data when the dimension of the model parameters is high.

Two main classes of methods attempt to fulfil this role: sampling-based Monte Carlo methods and optimisation-based Variational Inference methods. By relying on the optimisation literature and more recently on Monte Carlo methods, the latter have made it possible to construct fast algorithms that overcome some of the computational hurdles encountered in Bayesian Inference.

Yet, the theoretical results and empirical performances of Variational Inference methods are often impacted by two factors: one, an inappropriate choice of the objective function appearing in the optimisation problem and two, a search space that is too restrictive to match the target at the end of the optimisation procedure.

This thesis explores how we can remedy the two issues mentioned above in order to build improved adaptive algorithms for complex models at the intersection of Monte Carlo and Variational Inference methods.

In our work, we suggest selecting the α-divergence as a more general class of objective functions and we propose several ways to enlarge the search space beyond the traditional framework used in Variational Inference.

The specificity of our approach in this thesis is then that it derives numerically advantageous adaptive algorithms with strong theoretical foundations, in the sense that they provably ensure a systematic decrease in the α-divergence at each step. In addition, we unravel important connections between the sampling-based and the optimisation-based methodologies.

The thesis is then organised as follows:

- Chapter 1 (Introduction)

We present the central notions this thesis builds on and we sum up our main results.

- Chapter 2 (Based on Daudel, Douc, and Portier, 2021)

We introduce the (α, Γ)-descent, a novel iterative algorithm operating on measures that performs α-divergence minimisation. This gradient-based procedure extends the commonly-used variational approximation by adding a prior on the variational parameters in the form of a measure. It is shown to lead at each step to a systematic decrease in the α-divergence for a rich family of functions Γ and convergence results are also derived. It recovers the Entropic Mirror Descent algorithm as a special case and provides an alternative algorithm called the Power Descent. By resorting to Monte Carlo approximations, both algorithms can notably be used to optimise the mixture weights of any given mixture model without any information on the underlying distribution of the variational parameters. We demonstrate empirically the benefit of using the Power Descent and going beyond the Entropic Mirror Descent framework, which fails as the dimension grows.

- Chapter 3 (Based on Daudel and Douc, 2021)

We establish the full proof of the convergence of the Power Descent towards the optimal mixture weights when $\alpha < 1$. Observing that this algorithm is defined for all $\alpha \in \mathbb{R} \setminus \{1\}$ and since the α-divergence recovers the widely-used forward Kullback-Leibler when α goes to 1, we then extend the Power Descent to the case $\alpha = 1$ and show that we obtain an Entropic Mirror Descent. This leads us to further investigate the link between Power Descent and Entropic Mirror Descent: first-order approximations allow us to go beyond the (α, Γ)-descent framework and to introduce the Renyi Descent, a new algorithm for which we prove an $O(1/N)$ convergence rate. Lastly, we compare numerically the behavior of the unbiased Power Descent and of the biased Renyi Descent and we discuss the potential advantages of one algorithm over the other.

- Chapter 4 (Based on Daudel, Douc, and Roueff, 2021)

We propose a complete methodology to carry out α-divergence minimisation by ensuring a systematic decrease in the α-divergence at each step. In its most general form, our framework allows us to simultaneously optimise the weights and components parameters of a given mixture model. Our approach permits us to build on various methods previously proposed for α-divergence minimisation such as Gradient or Power Descent schemes and to enhance them. Furthermore, we shed a new light on an integrated Expectation-Maximization algorithm. By applying our work to the particular case of Gaussian Mixture Models optimisation via Monte Carlo approximations, we finally provide empirical evidence that our methodology yields improved results, all the while illustrating the numerical benefits of having introduced some flexibility through the parameter α of the α-divergence.

- Chapter 5 (Conclusion)

We provide concluding remarks and outline some future directions of research.

1 Introduction

The aim of this chapter is to introduce the main concepts arising in this thesis. We first recall the basics of Bayesian Inference and underline its core challenges when applied to complex models. Then, we explain how Monte Carlo and Variational Inference methods tackle these difficulties in order to carry out Bayesian Inference tasks, before summarising the contributions we make in the remaining chapters of this thesis.

1.1 Bayesian Inference

Statistical Inference is the process of modelling a phenomenon given some data. As a subclass of Statistical Inference, Bayesian Inference methods seek to fit a parameterised probability model to a set of observed data, with the particularity that prior knowledge on the model parameters is incorporated in the methods.

The framework of Bayesian Inference can then be defined as follows. Let $(\mathsf{Y}, \mathcal{Y}, \nu)$ be a measured space, where ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$. Assume that we have access to some observed variables $\mathscr{D}$ generated from a dominated probabilistic model with density $p(\mathscr{D}|y)$ parameterised by a hidden random variable $y \in \mathsf{Y}$ that is drawn from a certain prior with density p_0 with respect to ν. At the heart of Bayesian Inference is the posterior density of the latent variable y given the data $\mathscr{D}$:

$$p(y|\mathscr{D}) = \frac{p(y, \mathscr{D})}{p(\mathscr{D})} = \frac{p_0(y)p(\mathscr{D}|y)}{p(\mathscr{D})} \,,$$

where $p(\mathscr{D}) = \int_{\mathsf{Y}} p_0(y)p(\mathscr{D}|y)\nu(\mathrm{d}y)$ is called the *marginal likelihood* or *model evidence*. The posterior density is used to quantify the uncertainty of the parameter y after observing the data $\mathscr{D}$ through quantities of interest such as the marginal likelihood

$p(\mathscr{D})$ or the *posterior mean*

$$\int_{\mathsf{Y}} y\, p(y|\mathscr{D})\nu(\mathrm{d}y)\ .$$

Broadly speaking, given a function g defined on Y, the success of Bayesian Inference methods will rely on our ability to calculate integrals of the form

$$\int_{\mathsf{Y}} g(y)p(y|\mathscr{D})\nu(\mathrm{d}y)\ . \tag{1.1}$$

The problem above is a difficult one as there exists no general analytical form for (1.1) and even when an analytical form does exist for selected choices of probabilistic models, it might be too expensive to compute in practice (e.g. the computation of the marginal likelihood for a Bayesian Mixture of Gaussians, see Blei, Kucukelbir, and McAuliffe, 2017 for details).

This is particularly true in the context of Big Data, where modelling large amount of data with potentially complicated underlying structure inside the data will induce a complex and hard-to-compute posterior density. It is thus crucial to be able to find methods rendering Bayesian Inference computationally efficient and scalable to large datasets.

Since exact Bayesian Inference is often impossible, one may resort to *approximate* Bayesian Inference methods, which mainly fall into two broad categories: (i) Monte Carlo methods (e.g. Adaptive Importance Sampling (Oh and Berger, 1992), Markov Chain Monte Carlo (Neal, 1993), Sequential Monte Carlo (Doucet, Freitas, and Gordon, 2001)), that are *sampling* methods (ii) Variational Inference methods (e.g. Variational Bayes (Jordan et al., 1999), Expectation Propagation (Minka, 2001)), that rely on *optimisation* techniques.

In particular, Variational Inference methods are known for their numerical success when applied to large-scale learning tasks with complex probabilistic models (Hoffman et al., 2013; Kingma and Welling, 2014; Ranganath, Gerrish, and Blei, 2014). However, contrary to their Monte Carlo counterparts, Variational Inference methods use optimisation techniques over a constrained set of densities; this means that there is a possible mismatch between the posterior density and the approximation that is returned at the end of the optimisation procedure, which results in a lack of theoretical guarantees (Yao et al., 2018; Campbell and Li, 2019).

As a consequence, the literature is becoming increasingly interested in constructing scalable Variational Inference algorithms that are theoretically well-justified (e.g. Alquier, Ridgway, and Chopin, 2016; Domke, 2019; Alquier and Ridgway, 2020) while another active field of research focuses on combining Monte Carlo and Variational Inference methods (to name but a few: Burda, Grosse, and Salakhutdinov, 2016; Li and Turner, 2016; Mandt, Hoffman, and Blei, 2017; Naesseth et al., 2018; Thin et al., 2020; Naesseth, Lindsten, and Blei, 2020).

In this thesis, we will be particularly interested in investigating how Adaptive

Monte Carlo methods, and more specifically Adaptive Importance Sampling methods, can be paired up with scalable Variational Inference procedures to provide theoretically-sound algorithms for Bayesian purposes.

To this end, let us start by recalling the basics of Monte Carlo methods for Bayesian Inference up till Adaptive Importance Sampling methods.

1.2 Monte Carlo methods for Bayesian Inference

Monte Carlo methods as a whole seek to approximate integrals of the form

$$I(g) := \int_{\mathsf{Y}} g(y)p(y)\nu(\mathrm{d}y) \ ,$$

where g is an integrable function defined on Y and p is a probability density function with respect to ν on $(\mathsf{Y}, \mathcal{Y})$. Denoting by $\mathbb{P}$ the probability measure on $(\mathsf{Y}, \mathcal{Y})$ with Radon-Nikodym derivative with respect to ν given by $\mathrm{d}\mathbb{P}/\mathrm{d}\nu = p$, this problem can be reframed as the calculation of the expectation of g with respect to the probability distribution $\mathbb{P}$:

$$I(g) = \mathbb{E}_p[g(Y)] \ ,$$

where Y is a random variable defined on the probability space $(\mathsf{Y}, \mathcal{Y}, \mathbb{P})$.

The first idea of Monte Carlo methods is to replace the explicit calculation of the expectation of the random variable $g(Y)$ by an approximation involving the empirical mean of M independent realisations.

1.2.1 Vanilla Monte Carlo

Let $Y_1, Y_2, ...$ be an infinite sequence of independent and identically distributed random variables with common probability distribution $\mathbb{P}$. Setting for all $M \in \mathbb{N}^\star$

$$\hat{I}_M(g) = \frac{1}{M} \sum_{m=1}^{M} g(Y_m) \tag{1.2}$$

we obtain that the estimator $\hat{I}_M(g)$ of $I(g)$ is unbiased (i.e. $\mathbb{E}_p[\hat{I}_M(g)] = I(g)$). Notably, assuming that $I(g) = \mathbb{E}_p[|g(Y_1)|] < \infty$, the law of large numbers yields

$$\lim_{M\to\infty} \hat{I}_M(g) = I(g) \ , \quad \text{almost-surely,}$$

and if we further assume that $\mathbb{E}_p[|g(Y_1)|^2] < \infty$, we obtain by the central limit theorem that $\sqrt{M}(\hat{I}_M(g) - I(g))$ converges in distribution to $\mathcal{N}(0, \mathbb{V}\mathrm{ar}_p[g(Y_1)])$ as M goes to infinity.

To apply Monte Carlo methods for Bayesian Inference tasks, we would like to set $p(y) = p(y|\mathscr{D})$ for all $y \in \mathsf{Y}$: provided that we know how to sample from the posterior distribution, the estimator (1.2) would serve as an approximation of (1.1).

However, for many important Bayesian Inference models we do not know how to sample from the posterior distribution, nor do we know the value of the normalising constant $p(\mathscr{D})$. One example of such a model is Bayesian Logistic Regression for binary classification, as described below.

Example 1 (Bayesian Logistic Regression)**.** *We use the same setting as in Gershman, Hoffman, and Blei, 2012. We observe the data $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$ which is made of I binary class labels, $c_i \in \{-1, 1\}$, and of L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$. The hidden variables $y = \{\boldsymbol{\omega}, \beta\}$ consist of L regression coefficients $\omega_\ell \in \mathbb{R}$ and a precision parameter $\beta \in \mathbb{R}^+$. We assume the following model*

$$
\begin{aligned}
&p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \ , \\
&p_0(\omega_\ell|\beta) = \mathcal{N}(\omega_\ell; 0, \beta^{-1}) \ , \quad 1 \leqslant \ell \leqslant L \ , \\
&p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{\omega}) = \frac{1}{1 + e^{-\boldsymbol{\omega}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I \ ,
\end{aligned}
$$

where a and b are hyperparameters (shape and inverse scale, respectively) that we assume to be fixed. For all $y \in \mathsf{Y}$, we thus have $p(y, \mathscr{D}) \propto p_0(y) \prod_{i=1}^I p(c_i|\boldsymbol{x}_i, y)$ with $p_0(y) = \prod_{\ell=1}^L p_0(\omega_\ell|\beta) p_0(\beta)$. As the sigmoid does not admit a known conjugate prior, we do not know how to sample from the posterior distribution and $p(\mathscr{D})$ is intractable in this model. Consequently, the posterior predictive distribution, which given an unseen data x_{new} predicts the label c_{new}

$$
p(c_{\mathrm{new}}|\boldsymbol{x}_{\mathrm{new}}, \mathscr{D}) = \int_{\mathsf{Y}} p(c_{\mathrm{new}}|\boldsymbol{x}_{\mathrm{new}}, y) p(y|\mathscr{D}) \nu(\mathrm{d}y)
$$

is also an intractable integral.

Fortunately for us and as we shall see next, Importance Sampling methods come in handy to bypass these issues.

1.2.2 Importance Sampling

The key idea of Importance Sampling is to introduce a certain probability density q with respect to ν on $(\mathsf{Y}, \mathcal{Y})$ and to assume that (i) we know how to sample from q and (ii) the support of q contains the support of $g \times p$, that is for a given $y \in \mathsf{Y}$, $g(y)p(y) \neq 0$ implies $q(y) > 0$ (a sufficient condition being that the support of q contains the support of p). In this case, setting $w(y) = p(y)/q(y)$ for all $y \in \mathsf{Y}$, the following holds

$$
I(g) = \int_{\mathsf{Y}} g(y) p(y) \nu(\mathrm{d}y) = \int_{\mathsf{Y}} g(y) \frac{p(y)}{q(y)} q(y) \nu(\mathrm{d}y) = \mathbb{E}_q \left[w(Y) g(Y) \right] \ .
$$

Letting this time $Y_1, Y_2, \ldots$ be an infinite sequence of independent and identically distributed random variables sampled according to q and based on what we have seen so far, the novel (unbiased) estimator of $I(g)$ that comes to mind is given for all

$M \in \mathbb{N}^\star$ by

$$\hat{I}_M^{IS}(g) = \frac{1}{M} \sum_{m=1}^{M} w(Y_m) g(Y_m) \ . \tag{1.3}$$

If we put this into perspective with the Bayesian framework, we see that we have made some progress, as we do not need to be able to sample from the posterior distribution anymore to estimate integrals of the form (1.1).

One obstacle still remains, since as we underlined before the posterior density can often only be evaluated up to a proportional constant. This brings us to the Self-normalised Importance Sampling (SNIS) estimator defined below for all $M \in \mathbb{N}^\star$

$$\hat{I}_M^{SNIS}(g) = \frac{\frac{1}{M}\sum_{m=1}^{M} w(Y_m) g(Y_m)}{\frac{1}{M}\sum_{m=1}^{M} w(Y_m)} \ .$$

Contrary to the estimators we have introduced previously, $\hat{I}_M^{SNIS}(g)$ is biased. Nevertheless, assuming that the condition of support are met and $\mathbb{E}_q[|w(Y_1)g(Y_1)|] < \infty$, the law of large number allows us to obtain the almost sure convergence towards $I(g)$ for both $\hat{I}_M^{IS}(g)$ and $\hat{I}_M^{SNIS}(g)$.

Further assuming that $\mathbb{E}_q[|w(Y_1)g(Y_1)|^2] < \infty$ (and for the SNIS estimator that $\mathbb{E}_q[w(Y_1)^2(1+g(Y_1)^2)] < \infty$) also yields

$$\begin{aligned}
&\sqrt{M}(\hat{I}_M^{IS}(g) - I(g)) \to_{\mathcal{L}} \mathcal{N}(0, \mathbb{V}\mathrm{ar}_q[w(Y_1)g(Y_1)]) \\
&\sqrt{M}(\hat{I}_M^{SNIS}(g) - I(g)) \to_{\mathcal{L}} \mathcal{N}(0, \mathbb{V}\mathrm{ar}_q[w(Y_1)(g(Y_1) - I(g))]) \ ,
\end{aligned}$$

where $\to_{\mathcal{L}}$ denotes the convergence in distribution. Notably, $\mathbb{V}\mathrm{ar}_q[w(Y_1)g(Y_1)]$ and $\mathbb{V}\mathrm{ar}_q[w(Y_1)(g(Y_1) - I(g))]$ are minimal when $q \propto |g| \times p$ and $q \propto |g - I(g)|p$ respectively.

This illustrates the fact that the performance of Importance Sampling methods for Bayesian Inference purposes is tied to the choice of the sampler q (see Robert and Casella, 2005 and Figure 1.1). Interestingly, when integrals of the form (1.1) are to be computed for many functions g, it becomes less efficient to have q depend on g as per written in the above results, which supports the idea that one should target p directly (Delyon and Portier, 2021).

As we may not know in one go what a good sampler for a complex probabilistic model looks like, adaptive procedures may be constructed in order to refine the proposal progressively and this brings us to the concept of Adaptive Importance Sampling.

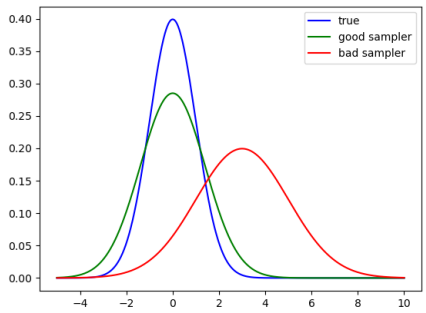


FIGURE 1.1: Here, the label "true" stands for the targeted distribution. The two remaining above samplers are then likely to provide different results in estimating $I(g)$.

1.2.3 Adaptive Importance Sampling

Starting from an initial sampler q_1, the aim of Adaptive Importance Sampling is to build an iterative sequence of sampler $(q_n)_{n\geqslant 1}$ that leads to more accurate estimators as n increases.

While originally limited to a two-step procedure (Kloek and Van Dijk, 1978; Geweke, 1989), Adaptive Importance Sampling methods have since evolved to multi-stage schemes (Oh and Berger, 1992) so that we can expect a typical Adaptive Importance Sampling algorithm to be described as in Algorithm 1.

Algorithm 1: *Adaptive Importance Sampling*

Input: N: total number of iterations, $(M_n)_{1\leqslant n\leqslant N}$: allocation policy, q_1: initial sampler.
Output: Return the pairs $(Y_{m,n}, w_n(Y_{m,n}))_{1\leqslant m\leqslant M_n, 1\leqslant n\leqslant N}$.
for $n = 1 \dots N$ **do**

1. Draw independently M_n samples $(Y_{m,n})_{1\leqslant m\leqslant M_n}$ from q_n.
2. Compute the Importance Sampling weights $(w_n(Y_{m,n}))_{1\leqslant m\leqslant M_n}$, where for all $y \in \mathsf{Y}$, $w_n(y) = p(y)/q_n(y)$.
3. Update the sampler q_n.

end

Letting $p(y) = p(y, \mathscr{D})$ for all $y \in \mathsf{Y}$, the outputted pairs in Algorithm 1 can then be used to estimate integrals of the form (1.1), e.g. by considering at time n the estimate

$$\hat{I}^{SNAIS}_{M_n,n}(g) = \frac{\frac{1}{M_n}\sum_{m=1}^{M_n} w_n(Y_{m,n})g(Y_{m,n})}{\frac{1}{M_n}\sum_{m=1}^{M_n} w_n(Y_{m,n})} .$$

Notable advances in Adaptive Importance Sampling include methods tailoring the sequence of samplers $(q_n)_{n\geqslant 1}$ according to a certain criterion (e.g. Douc et al., 2007a; Douc et al., 2007b; Cappé et al., 2008 and Portier and Delyon, 2018) as well as refinements beyond the traditional importance sampling weights (Martino et al., 2017).

We refer to Bugallo et al., 2017 for a detailed review of Adaptive Importance Sampling methods and we now move on to presenting Variational Inference methods.

1.3 Variational Inference methods for Bayesian Inference

Variational Inference methods (Jordan et al., 1999) seek to approximate the posterior density by a simpler variational density q belonging to some density family $\mathcal{Q}$ and that can be used to facilitate the computation of integrals of the form (1.1).

These approaches consider this objective purely as an optimisation problem involving a certain measure of dissimilarity D between the posterior distribution $\mathbb{P}_{|\mathscr{D}}$ and the variational distribution $\mathbb{Q}$

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \ ,$$

where $\mathbb{P}_{|\mathscr{D}}$ and $\mathbb{Q}$ are assumed to be probability measures on $(\mathsf{Y}, \mathcal{Y})$ that are absolutely continuous with respect to ν (which we denote $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$) and with associated Radon-Nikodym derivatives with respect to ν given by $q = \mathrm{d}\mathbb{Q}/\mathrm{d}\nu$ and $p(\cdot|\mathscr{D}) = \mathrm{d}\mathbb{P}_{|\mathscr{D}}/\mathrm{d}\nu$.

The core of Variational Inference methods then consists in choosing D properly and in designing approximating families $\mathcal{Q}$ which enable efficient optimisation and which are able to capture complicated structure inside the posterior density.

In this section, we will first recall the most traditional choices for D and $\mathcal{Q}$ in the Variational Inference literature. We will then detail advances in this field that are relevant for the thesis.

1.3.1 Traditional Variational Inference

A traditional choice in Variational Inference that has extensively been used in the literature corresponds to letting D be the Kullback-Leibler divergence (Kullback and Leibler, 1951), whose definition is recalled now.

Definition 1 (Kullback-Leibler divergence)**.** *Let $\mathbb{Q}$ and $\mathbb{P}$ be two probability measures on $(\mathsf{Y}, \mathcal{Y})$ that are absolutely continuous with respect to ν i.e. $\mathbb{Q} \preceq \nu$, $\mathbb{P} \preceq \nu$. Let us denote by $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$ and $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$ the Radon-Nikodym derivatives of $\mathbb{Q}$ and $\mathbb{P}$ with respect to ν. The Kullback-Leibler (KL) divergence is defined by:*

$$D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y)$$

which is always well-defined in $[0, +\infty]$.

In traditional Variational Inference, one can then seek to either minimise the forward Kullback-Leibler

$$\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1.4}$$

or to minimise the reverse Kullback-Leibler

$$\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{P}_{|\mathscr{D}}||\mathbb{Q}) \ . \tag{1.5}$$

Special interest in the Variational Inference community has notably been dedicated to attempting to solve (1.4), due to the so-called Evidence Lower BOund (ELBO) property: for all $q \in \mathcal{Q}$ it holds that

$$\begin{aligned} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) &= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D}) \\ &= -\mathrm{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D}) \ , \end{aligned}$$

where the ELBO function is defined for all $q \in \mathcal{Q}$ by

$$\mathrm{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \ . \tag{1.6}$$

The result above means that the ELBO can act as a surrogate objective function which does not involve the bothersome normalising constant $p(\mathscr{D})$ anymore. Thus, it is equivalent to consider instead of (1.4) the optimisation problem

$$\sup_{q\in\mathcal{Q}} \mathrm{ELBO}(q;\mathscr{D}) \ .$$

The name ELBO in itself then comes from the fact that Jensen's inequality applied to the strictly concave function $u \mapsto \log(u)$ implies $D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \geqslant 0$ so that we can write $\mathrm{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality if and only if $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$; this, in turn, provides a *lower bound* on the log of the marginal likelihood (i.e. the model evidence).

From there, a traditional choice for the variational family $\mathcal{Q}$ is to work within the Mean-field approximating family, which we next describe. This will allow us to explain briefly how the forward Kullback-Leibler divergence and the Mean-field approximation have been paired up together for Variational Inference purposes.

The Mean-field approximation consists in assuming that the latent variable y is made of L independent components $(y_1, \ldots, y_L) \in \mathsf{Y}_1 \times \ldots \times \mathsf{Y}_L$ such that $\mathcal{Q}$ is of the form

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\}$$

and each latent variable y_ℓ is governed by its own variational density q_ℓ with $\nu(\mathrm{d}y) =$

$\bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$. Plugging this fully factorised variational density into the ELBO (1.6) and fixing all the variational factors but the one with coordinate ℓ, the following optimal update can then be derived for this factor:

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right) \ , \text{ for } \nu_\ell\text{-almost all } y_\ell \in \mathsf{Y}_\ell \ , \tag{1.7}$$

where we have denoted by $\mathbb{E}_{-\ell}$ the expectation with respect to q omitting the factor q_ℓ. Indeed, observe that under the Mean-field assumption,

$$\begin{aligned}
\mathrm{ELBO}(q; \mathscr{D}) &= \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log q_\ell(y_\ell) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell} \\
&= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}
\end{aligned}$$

where $c_{-\ell}$ is a constant that does not depend on q_ℓ (and for convenience we have made a slight abuse of notation in $\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]$ by using the same notation for the variables $(y_k)_{1 \leqslant k \leqslant L, k \neq \ell}$ and the random variables under $\prod_{k=1, k\neq\ell}^{L} q_k$).

Then, as a consequence of Jensen's inequality the left-hand side is maximised when $q_\ell(y_\ell)$ is proportional to $\exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ for ν_ℓ-almost all $y_\ell \in \mathsf{Y}_\ell$, and we recover the aforementioned optimality condition (1.7) for the factor q_ℓ.

Based on this result, a natural idea of an iterative algorithm consists in performing the update (1.7) successively for $\ell = 1 \ldots L$ and in repeating this cycle until convergence towards a (local) optimum is reached: this procedure is called the Coordinate Ascent Variational Inference (CAVI) algorithm (Bishop, 2006) and it is summarised in Algorithm 2.

Algorithm 2: *Coordinate Ascent Variational Inference (CAVI)*

Input: $(q_\ell)_{1 \leqslant \ell \leqslant L}$: initial variational factors.
Output: Return the optimised Mean-field variational density q satisfying:
for all $y \in \mathsf{Y}$, $q(y) = \prod_{\ell=1}^{L} q_\ell(y_\ell)$.
while *the ELBO has not converged* **do**
  for $\ell = 1 \ldots L$ **do**
    set $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$, for ν_ℓ-almost all $y_\ell \in \mathsf{Y}_\ell$
  end
  Compute the ELBO.
end

To see how CAVI updates are derived in practice and observe what type of Mean-field variational density q is obtained after optimisation, we next provide in Example 2 below a toy example taken from Hernandez-Lobato et al., 2016 in which the posterior density is known.

Example 2 (Bayesian Linear Regression). *We observe the data $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$ that is made of I 1-D class labels $(c_i)_{1\leqslant i\leqslant I}$ and of I 2-D covariates $(\boldsymbol{x}_i)_{1\leqslant i\leqslant I}$, where for each datapoint $(c_i, \boldsymbol{x}_i) \in \mathbb{R}\times\mathbb{R}^2$. The hidden variables consist of two regression coefficients $y = \{y_1, y_2\} \in \mathbb{R}^2$. We assume the following model*

$$\begin{aligned} p_0(y) &= \mathcal{N}(y; \mu_0, \Lambda_0^{-1}) \ , \\ p(c_i|\boldsymbol{x}_i, y) &= \mathcal{N}(c_i; y^T\boldsymbol{x}_i, \sigma^2) \ , \quad 1 \leqslant i \leqslant I \ , \end{aligned}$$

where μ_0, Λ_0 and σ are hyperparameters that are assumed to be fixed.

Then, the posterior density is known and we have

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

with $\Lambda = \Lambda_0 + \sigma^{-2}\sum_{i=1}^I \boldsymbol{x}_i\boldsymbol{x}_i^T$ and $\Lambda\mu = \Lambda_0\mu_0 + \sigma^{-2}\sum_{i=1}^I c_i\boldsymbol{x}_i$. Under the Mean-field assumption $q(y) = q_1(y_1)q_2(y_2)$ so that for all $\ell = \{1, 2\}$, we want to find

$$q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right) \ .$$

Introducing the notation $\mu = (\mu_\ell)_{1\leqslant\ell\leqslant 2}$ and $\Lambda = (\Lambda_{\ell,k})_{1\leqslant\ell,k\leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$, we have

$$\begin{aligned} q_1(y_1) &\propto \exp\left(\mathbb{E}_{q_2}\left[-\frac{1}{2}\left\{(y_1-\mu_1)^2\Lambda_{1,1} + 2(y_1-\mu_1)(y_2-\mu_2)\Lambda_{1,2}\right\}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right) \ , \end{aligned}$$

so that we can deduce $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$ and by symmetry that $q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{q_1}[y_1] - \mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$. Consequently, denoting $m_1 = \mathbb{E}_{q_1}[y_1]$ and $m_2 = \mathbb{E}_{q_2}[y_2]$, the CAVI algorithm amounts to performing the iterations

$$\begin{aligned} m_1 &\leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2 - \mu_2)\Lambda_{1,2} \\ m_2 &\leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1 - \mu_1)\Lambda_{1,2} \ . \end{aligned}$$

Since the only stable fixed point is given by $m_1 = \mu_1$ and $m_2 = \mu_2$, we finally obtain that $q_1(y_1) = \mathcal{N}(y_1; \mu_1, \Lambda_{1,1}^{-1})$ and $q_2(y_2) = \mathcal{N}(y_2; \mu_2, \Lambda_{2,2}^{-1})$. Visually, the posterior distribution and the optimised variational distribution can be observed on Figure 1.2, where we have taken $\mu = [0, 0]$, $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = -2$.

More generally, the CAVI algorithm may result in tractable updates when applied to conjugate exponential family models, some of such instances being Gaussian Mixture Models (Bishop, 2006) and latent Dirichlet allocation (Blei, Ng, and Jordan, 2003).

To be precise, given the dataset $\mathscr{D} = (x_\ell)_{1\leqslant\ell\leqslant L}$, conjugate exponential family models introduce the latent variables $y = \{\beta, \omega_1, \ldots, \omega_L\}$ where β is seen as a global latent variable and for all $l = 1 \ldots L$, ω_ℓ is a local latent variable associated to the

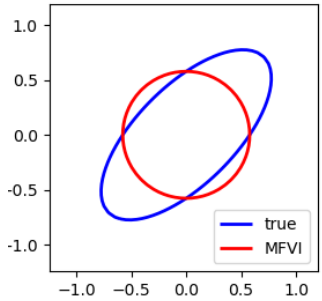


FIGURE 1.2: Mean-field approximation for the Bayesian Linear Regression from Example 2 (adapted from Hernandez-Lobato et al., 2016). The labels "true" and "MVFI" respectively stand for the posterior distribution and the Mean-field approximation obtained by forward Kullback-Leibler minimisation (with one-sigma contours).

datapoint x_ℓ so that

$$p(y, \mathscr{D}) = p(\beta) \prod_{\ell=1}^{L} p(\omega_\ell, x_\ell|\beta) \ .$$

They next consider the following Mean-field variational approximation q

$$q(y) = q(\beta|\psi) \prod_{\ell=1}^{L} q(\omega_\ell|\phi_\ell) \ ,$$

where $\{\psi, \phi_1, ..., \phi_L\}$ correspond to the variational parameters to be optimised via the CAVI algorithm. These models can then be proven to yield tractable updates for the variational parameters by making appropriate choices for $p(\beta), p(\omega_\ell, x_\ell|\beta), q(\beta|\psi)$ and $q(\omega_\ell|\phi_\ell)$ (see Blei, Kucukelbir, and McAuliffe, 2017 for details regarding Gaussian Mixture Models and latent Dirichlet allocation).

We have seen how tractable variational parameters updates based on the CAVI algorithm can be derived when D is the forward Kullback-Leibler, Q belongs to the Mean-field family and we work with well-chosen conjugate exponential family models. In the context of Big Data, one last hurdle must be overcome to obtain a fully-usable algorithm: the CAVI algorithm becomes inefficient for large datasets as it must optimise the local variational parameters $\{\phi_1, ..., \phi_L\}$ for each datapoint before re-estimating the global variational parameter ψ.

To remedy this situation, scalable methods relying on stochastic optimisation techniques (Bottou, 2010; Robbins and Monro, 1951) were developed to enable large-scale learning. These methods fall under the name of Stochastic Variational Inference (Hoffman et al., 2013) and were applied to some complex probabilistic models including latent Dirichlet allocation.

The numerical success of this approach on datasets comprising millions of datapoints has led to renewed interest in Variational Inference methods (and we refer the

reader to Blei, Kucukelbir, and McAuliffe, 2017 and Zhang et al., 2019 for comprehensive reviews around modern Variational Inference methods). In the rest of the section, we limit ourselves to revisiting the main advances in Variational Inference that are relevant in the subsequent chapters of the thesis.

1.3.2 Monte Carlo meets Variational Inference

As we have stressed previously, Variational Inference is particularly amenable to coordinate-ascent optimization when we work with the forward Kullback-Leibler divergence and under the Mean-field assumption.

However, one of the main limitations of this approach is that not only the Mean-field family restricts the choice of models but also that tractable updates are model-specific and require by-hand derivation (see Blei, Kucukelbir, and McAuliffe, 2017 and Figure 1.2).

For these reasons, Black-Box Variational Inference techniques (Ranganath, Gerrish, and Blei, 2014) have been deployed as a generic class of Variational Inference algorithms for forward Kullback-Leibler minimisation that renders Variational Inference methods applicable to a wide range of models. Letting D be the forward Kullback-Leibler and assuming that we are working with a general parametric family of the form

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \tag{1.8}$$

(where T is for example $\mathbb{R}^d$) the main idea of Black-Box Variational Inference is to use the gradient of the ELBO paired up with Monte Carlo approximations in order to carry out the optimisation procedure. Indeed, under common differentiability assumption and following Paisley, Blei, and Jordan, 2012, the gradient of the ELBO (1.6) is given by:

$$\begin{aligned}
\nabla \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D}) &= \nabla \left(\int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) \nu(\mathrm{d}y) \right) \\
&= \int_{\mathsf{Y}} \nabla k(\theta, y) \times \left[\log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) - 1 \right] \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} k(\theta, y) \nabla \left[\log k(\theta, y) \right] \log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) \nu(\mathrm{d}y) - \int_{\mathsf{Y}} \nabla k(\theta, y) \nu(\mathrm{d}y)
\end{aligned}$$

where we have used that for all $y \in \mathsf{Y}$, $\nabla k(\theta, y) = k(\theta, y) \nabla \left[\log k(\theta, y) \right]$, an operation known as the REINFORCE trick in the literature (Williams, 1992). By further noticing that $\int_{\mathsf{Y}} \nabla k(\theta, y) \nu(\mathrm{d}y) = \nabla \left(\int_{\mathsf{Y}} k(\theta, y) \nu(\mathrm{d}y) \right) = 0$, we deduce

$$\begin{aligned}
\nabla \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D}) &= \int_{\mathsf{Y}} k(\theta, y) \nabla \left[\log k(\theta, y) \right] \log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) \nu(\mathrm{d}y) \\
&= \mathbb{E}_{k(\theta,\cdot)} \left[\nabla \log \left[k(\theta, Y) \right] \times \log \left(\frac{p(Y,\mathscr{D})}{k(\theta,Y)} \right) \right]
\end{aligned}$$

so that the gradient of the ELBO can be expressed as an expectation with respect to the variational approximation $k(\theta, \cdot)$. This is where Monte Carlo techniques intervene: given M independent and identically distributed random variables $Y_1, \dots, Y_M$ sampled according to $k(\theta, \cdot)$, an unbiased estimate of the expectation above is

$$\frac{1}{M} \sum_{m=1}^{M} \nabla \log \left[k(\theta, Y_m)\right] \log \left(\frac{p(Y_m, \mathscr{D})}{k(\theta, Y)} \right) .$$

The Black-Box Variational Inference algorithm in itself then consists in introducing a sequence of learning rates $(\gamma_n)_{n \geqslant 1}$ and performing Stochastic Gradient Descent steps to construct a sequence $(\theta_n)_{n \geqslant 1}$ according to Algorithm 3 below (notice the "+" sign in the gradient step, as we seek to maximise the ELBO and thus minimise $-$ELBO).

Algorithm 3: *Black-Box Variational Inference*

Input: N: total number of iterations, $(M_n)_{n \geqslant 1}$: allocation policy, $(\gamma_n)_{n \geqslant 1}$: learning rate policy, θ_1: initial parameter value.
Output: Return the optimised parameter θ_{N+1}.
for $n = 1 \dots N$ **do**

1. Draw independently M_n samples $(Y_{m,n})_{1 \leqslant m \leqslant M_n}$ from $k(\theta_n, \cdot)$.
2. Set

$$\theta_{n+1} = \theta_n + \gamma_n \frac{1}{M_n} \sum_{m=1}^{M_n} \nabla \log \left[k(\theta, Y_{m,n})\right] |_{\theta = \theta_n} \log \left(\frac{p(Y_{m,n}, \mathscr{D})}{k(\theta_n, Y_{m,n})} \right) .$$

end

The particularity of this scheme is that Stochastic Gradient Descent steps are being performed using an *unbiased* estimate of the gradient of the ELBO. This means that, under appropriate assumptions on the learning rate policy and on the objective function (Ranganath, Gerrish, and Blei, 2014; Domke, 2019; Domke, 2020), $(k(\theta_n, \cdot))_{n \geqslant 1}$ converges towards an optimum of the ELBO, which effectively minimises (at least locally) the forward Kullback-Leibler divergence.

As Black-Box Variational Inference methods might suffer from high variances of the estimated gradients, much of the success of these schemes came from variance reduction techniques (e.g. Rao-Blackwellization, control variates (Ranganath, Gerrish, and Blei, 2014), reparametrisation (Kingma and Welling, 2014) and Quasi-Monte Carlo methods (Buchholz, Wenzel, and Mandt, 2018)).

So far, the choice of D has been limited to considering the forward and reverse Kullback-Leibler. However, another main appeal of Black-Box Variational Inference methods is that they can be used to optimise alternative objective functions beyond the particular case of the Kullback-Leibler divergence. In particular, efficient procedures have been designed when D belongs to the α-divergence family.

1.3.3 Variational Inference within the α-divergence family

Variational approximation distributions obtained by forward or reverse Kullback-Leibler minimisation are known to encounter practical issues (Minka, 2001; Hoffman et al., 2013; Blei, Kucukelbir, and McAuliffe, 2017), e.g. underestimating / overestimating the posterior variance for the forward/reverse Kullback-Leibler (posterior variance underestimation is even sometimes reinforced when additionally working under the mean-field assumption for the forward Kullback-Leibler, see Figure 1.2).

Therefore, another branch of Variational Inference methods focused on designing algorithms based on alternative families of divergences. Notably, some early works building on the α-divergence (Zhu and Rohwer, 1995a; Zhu and Rohwer, 1995b) can be found in Minka, 2004 and Minka, 2005. Before getting into the details of how the α-divergence family can be used for Variational Inference methods, let us first review basic concepts and ideas around this family.

The α-divergence family is a well-known family of divergence measures in the Information Geometry literature (e.g. Cichocki and Amari, 2010) which generalises the Kullback-Leibler divergence and whose definition for two probability measures $\mathbb{Q}$ and $\mathbb{P}$ is given below.

Definition 2. *Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Let $\mathbb{Q}$ and $\mathbb{P}$ be two probability measures on $(\mathsf{Y}, \mathcal{Y})$ that are absolutely continuous with respect to ν i.e. $\mathbb{Q} \preceq \nu$, $\mathbb{P} \preceq \nu$. Let us denote by $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$ and $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$ the Radon-Nikodym derivatives of $\mathbb{Q}$ and $\mathbb{P}$ with respect to ν. The α-divergence between $\mathbb{Q}$ and $\mathbb{P}$ is defined by :*

$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)} \right)^\alpha - 1 \right] p(y)\nu(\mathrm{d}y) \ ,$$

which is always well-defined in $[0, +\infty]$.

Under common differentiability assumptions, it holds that the α-divergence admits the forward and reverse Kullback-Leibler as limiting cases: for all $y \in \mathsf{Y}$,

$$\begin{aligned} \lim_{\alpha \to 0} \frac{1}{\alpha} \left[\frac{1}{\alpha - 1} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha - 1} \right] &= \nabla \left[\frac{1}{\alpha - 1} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha - 1} \right] \Bigg|_{\alpha=0} \\ &= -\log \left(\frac{q(y)}{p(y)} \right) \end{aligned}$$

and similarly

$$\begin{aligned} \lim_{\alpha \to 1} \frac{1}{\alpha - 1} \left[\frac{1}{\alpha} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha} \right] &= \nabla \left[\frac{1}{\alpha} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha} \right] \Bigg|_{\alpha=1} \\ &= 1 - \frac{q(y)}{p(y)} + \log \left(\frac{q(y)}{p(y)} \right) \frac{q(y)}{p(y)} \end{aligned}$$

so that $\lim_{\alpha \to 0} D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{P}||\mathbb{Q})$ and $\lim_{\alpha \to 1} D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{Q}||\mathbb{P})$.

The definition of the α-divergence can thus be extended to 0 and 1 by continuity and we will use the notation $D_0(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{P}||\mathbb{Q})$ and $D_1(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{Q}||\mathbb{P})$ from now on. Notice also that special cases of the α-divergence family include the Hellinger distance and the χ^2-divergence which correspond respectively to order $\alpha = 0.5$ and $\alpha = 2$.

Letting f_α be the *convex* function on $(0, +\infty)$ defined by $f_0(u) = u - 1 - \log(u)$, $f_1(u) = 1 - u + u\log(u)$ and $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$ for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$, we then have that for all $\alpha \in \mathbb{R}$,

$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \; . \tag{1.9}$$

Written under that form, the r.h.s of (1.9) corresponds to the general definition of the α-divergence (Cichocki and Amari, 2010). This formulation also tells us that α-divergences are members of the f-divergence family (Morimoto, 1963a; Morimoto, 1963b) through the convexity of f_α.

The fundamental properties of the α-divergence are given in the next proposition (and we refer to Minka, 2005; Cichocki and Amari, 2010; Cichocki, Cruces, and Amari, 2011; Erven and Harremoes, 2014 and Sason, 2018 for more details around the α-divergence family).

Proposition 3. *The α-divergence (extended by continuity to the cases $\alpha = 0$ and $\alpha = 1$) is always non-negative and it is equal to zero if and only if $\mathbb{Q} = \mathbb{P}$. Furthermore, it is jointly convex in $\mathbb{Q}$ and $\mathbb{P}$ and the definition of the α-divergence is invariant with respect to the transformation $\tilde{f}_{\alpha,c}(u) = f_\alpha(u) + c(u-1)$ for any arbitrary constant c, that is f_α can be equivalently replaced by $\tilde{f}_{\alpha,c}$ in (1.9).*

A more general optimisation problem than forward and reverse Kullback-Leibler minimisation as written in (1.4) and (1.5) then consists in considering

$$\inf_{q \in \mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \; . \tag{1.10}$$

Interestingly, it has been observed in Minka, 2005 that the characteristics of the resulting optimised variational density will vary depending on the value of the hyper-parameter α.

More precisely, there are two main regimes: either $\alpha \leqslant 0$ and the α-divergence is *mass-covering*, meaning that it will favor variational densities that cover all the modes or $\alpha \geqslant 1$ and the α-divergence is *mode-seeking*, that q will tend to be attracted to the mode with the largest probability mass (the case $\alpha \in (0, 1)$ corresponding to a mix of the two worlds). This comes from the fact that $D_\alpha(\mathbb{Q}||\mathbb{P})$ will blow up if the support of q is bigger than the support of p when $\alpha \geqslant 1$ and conversely, $D_\alpha(\mathbb{Q}||\mathbb{P})$ will blow up if the support of p is bigger than the support of q when $\alpha \leqslant 0$.

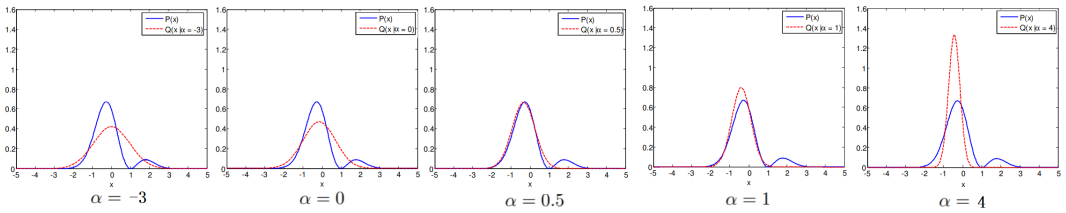


FIGURE 1.3: The Gaussian q which minimizes the α-divergence to the multimodal distribution p, for varying values of α. (adapted from Cevher's lecture notes available at https://www.ece.rice.edu/~vc3/elec633/AlphaDivergence.pdf)

An illustration of this mass-covering/mode-seeking property can be found in Figure 1.3, where given a targeted multimodal distribution, we seek to find the optimal Gaussian q in terms of the α-divergence for varying values of α.

Following up from earlier, the effect of α on the optimal Mean-field variational approximation q for the model considered in Example 2 is also depicted in Figure 1.4 (detailed derivations can be found in Appendix A.1), which further underlines the mass-covering/mode-seeking property of the α-divergence family.

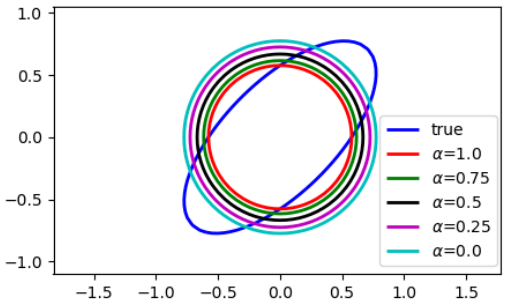


FIGURE 1.4: Optimal mean-field approximation with varying values of α for the Bayesian Linear Regression model from Example 2 (adapted from Hernandez-Lobato et al., 2016). The label "true" stands for the posterior distribution and the various Mean-field approximations are obtained by α-divergence minimisation (with one-sigma contours).

The mass-covering/mode-seeking property renders the optimisation problem (1.10) attractive for Variational Inference means, as it interpolates between the forward and reverse Kullback-Leibler divergence behaviors, which explains the interest dedicated to this family of divergences in Minka, 2004 and Minka, 2005. Yet, these works were limited to exponential family distributions.

With the advent of Monte Carlo Variational Inference, novel methods based on the α-divergence have been designed and have been found to provide promising empirical results (Hernandez-Lobato et al., 2016; Li and Turner, 2016; Dieng et al., 2017; Kuleshov and Ermon, 2017).

These methods exploit the fact that the specific form of f_α allows us to remove

the marginal likelihood $p(\mathscr{D})$ appearing in the optimisation problem (1.10) and can be classified in two groups: *biased* methods (Hernandez-Lobato et al., 2016; Li and Turner, 2016) and *unbiased* methods (Dieng et al., 2017; Kuleshov and Ermon, 2017).

Biased methods consider a slightly modified version of (1.10) which relies on the closely-related Renyi's α-divergence (Rényi, 1961; Erven and Harremoes, 2014)

$$\begin{aligned} D_\alpha^{(\mathrm{AR})}(\mathbb{Q}||\mathbb{P}) &= \frac{1}{\alpha-1}\log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{\alpha-1}\nu(\mathrm{d}y)\right) \\ &= \frac{1}{\alpha-1}\log\left(1+\alpha(\alpha-1)D_\alpha(\mathbb{Q}||\mathbb{P})\right) \ . \end{aligned}$$

In particular, Li and Turner, 2016 formalised the concept of Variational Renyi (VR) bound, a novel objective function which generalises the ELBO and is defined for all $\alpha \in \mathbb{R}\setminus\{1\}$ and for any variational density $q \in \mathcal{Q}$ by

$$\mathcal{L}_\alpha(q;\mathscr{D}) := \frac{1}{1-\alpha}\log\left(\int_{\mathsf{Y}}\left(\frac{p(y,\mathscr{D})}{q(y)}\right)^{1-\alpha} q(y)\nu(\mathrm{d}y)\right)$$

and they thus aim at finding

$$\sup_{q\in\mathcal{Q}} \mathcal{L}_\alpha(q;\mathscr{D}) \ .$$

This VR bound is shown to provide a lower or upper bound on the log-likelihood $\log p(\mathscr{D})$ depending on the sign of α and to recover the ELBO when $\alpha \to 1$ (Li and Turner, 2016, Theorem 1).

Optimisation is then carried out for a parametric family of the form (1.8) in a Black-Box Variational Inference manner by performing Stochastic Gradient Descent steps on $-\mathcal{L}_\alpha(q;\mathscr{D})$, which brings into play a biased Monte Carlo estimator of the gradient of the VR bound due to the $\log$. On the other hand, unbiased methods consider the objective function given by: for all $q \in \mathcal{Q}$,

$$\Psi_\alpha(q;\mathscr{D}) := \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y)$$

and aim at solving the following equivalent (see Appendix A.2) optimisation problem

$$\inf_{q\in\mathcal{Q}} \Psi_\alpha(q;\mathscr{D}) \tag{1.11}$$

via unbiased Stochastic Gradient Descent.

Advances in α-divergence-based Variational Inference notably include automatically tuning the hyperparameter α (Wang, Liu, and Liu, 2018) as well as attempts at getting a better theoretical and practical understanding of which approach is best between biased and unbiased α-divergence Variational Inference (Geffner and Domke, 2020a; Geffner and Domke, 2020b).

We have reviewed the basics of Adaptive Importance Sampling methods and

seen a variety of Variational Inference methods which seek to improve on the typical Mean-field Variational Inference framework with the forward Kullback-Leibler divergence, ranging from Black-Box Variational Inference techniques to considering more general objective functions. Yet and as we shall see next, some further improvements on these methods can be made in order to better capture the complexity of the posterior density.

1.4 Goal of the thesis and chapters overview

From an Adaptive Importance Sampling perspective, one cannot help but notice that Variational Inference techniques can be reframed as an instance of Step 3 in Algorithm 1 since they build a sequence of samplers that is refined iteratively in terms of a certain objective function.

Even more interestingly, in Black-Box Variational Inference techniques (be it for forward Kullback-Leibler or more generally α-divergence minimisation), the past samples generated to construct the sequence of samplers $(k(\theta_n, \cdot))_{n\geqslant 1}$ can readily be used to approximate integrals of the form (1.1).

For this reason, one can be inclined to take a Variational Inference approach to derive improved Adaptive Monte Carlo methods. In that case, since the performances of Variational Inference methods are limited by the choice of the approximating family $\mathcal{Q}$ and of the divergence D, one may wonder whether it is possible to enrich $\mathcal{Q}$ beyond the framework of Black-Box Variational Inference for α-divergence minimisation while still maintaining efficient optimisation.

To answer this question, this thesis explores novel scalable Variational Inference algorithms for α-divergence minimisation that (i) can be used in Adaptive Importance Sampling schemes and (ii) increase the expressiveness of the approximating family $\mathcal{Q}$. More precisely, our work can be decomposed in three chapters, which are based on three separate papers:

- Chapter 2 Daudel, Douc, and Portier, 2021.
"Infinite-dimensional gradient-based descent for Alpha-divergence minimisation".
To appear in the Annals of Statistics.

- Chapter 3 Daudel and Douc, 2021.
"Mixture weights optimisation for Alpha-divergence Variational Inference".
Submitted as a conference paper at the time of writing.

- Chapter 4 Daudel, Douc, and Roueff, 2021.
"Monotonic Alpha-divergence minimisation".
Submitted as a journal paper at the time of writing.

The common thread between these three works is that we were interested in creating iterative Variational Inference algorithms that ensured a *systematic decrease* in the α-divergence at each step. We provide below an overview of each chapter, giving particular emphasis to our own contributions.

1.4.1 Chapter 2: Infinite-dimensional α-divergence minimisation

In order to enlarge the parametric variational family

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

where θ is typically tuned through Stochastic Gradient Descent optimisation on either $\Psi_\alpha(k(\theta, \cdot); \mathscr{D})$ or $-\mathcal{L}_\alpha(k(\theta, \cdot); \mathscr{D})$, our first idea is to add a prior on the variational parameter θ in the form of a measure, that is we seek to perform α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\} , \tag{1.12}$$

where M is a convenient subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on T (and in this case, we equip T with a σ-field denoted by $\mathcal{T}$).

In doing so, we extend the minimizing set to a larger space since a parameter θ can be identified with its associated Dirac measure δ_θ and our approach complements already-existing Hierarchical Variational Inference methods (Ranganath, Tran, and Blei, 2016; Yin and Zhou, 2018; Titsias and Ruiz, 2019).

Indeed, while these methods restrict themselves to the forward Kullback-Leibler as objective function and consider that μ is parameterised by another parametric model so that

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \lambda_\phi(\theta) k(\theta, y) \mathrm{d}\theta \ : \ \phi \in A \right\}$$

with $\mu(\mathrm{d}\theta) = \lambda_\phi(\theta)\mathrm{d}\theta$ and where ϕ is optimised via Stochastic Gradient Descent, our framework sets the α-divergence as a more general objective function and allows us to target the important class of mixture models by taking μ as a weighted sum of Dirac measures.

Furthermore, another advantage of the approximating *infinite-dimensional* family (1.12) is that minimising the α-divergence with respect to μ between the variational density q and the targeted posterior density yields a convex optimisation problem, while the optimisation problem obtained when using a parametric variational family (be it parameterised by θ or ϕ) often does not.

More formally, letting $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ be a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density k defined on $\mathsf{T} \times \mathsf{Y}$, letting $q \in \mathcal{Q}$ be defined as in (1.12) and denoting $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}y) k(\theta, y)$ for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $y \in \mathsf{Y}$, we are interested in designing an iterative scheme that we hope will converge towards the global optimum of the α-divergence

$$\inf_{q \in \mathcal{Q}} D_\alpha(\mathbb{Q} || \mathbb{P}_{|\mathscr{D}}) = \inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; \mathscr{D}) \ .$$

For notational convenience, we define for all measurable positive function p on $(\mathsf{Y}, \mathcal{Y})$ and all probability density q with respect to ν on $(\mathsf{Y}, \mathcal{Y})$

$$\Psi_\alpha(q; p) := \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \,, \tag{1.13}$$

where we may drop the dependency on p and use the shorthand notation $\Psi_\alpha(q; \mathscr{D})$ when $p = p(\cdot, \mathscr{D})$ to denote $\Psi_\alpha(q; p(\cdot, \mathscr{D}))$, so that the general optimisation problem we consider in Chapter 2 is

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k) \,. \tag{1.14}$$

To solve the optimisation problem (1.14), we assume that we work under the mild assumption

(1.A1) The density kernel k on $\mathsf{T} \times \mathsf{Y}$, the function p on Y and the σ-finite measure ν on $(\mathsf{Y}, \mathcal{Y})$ satisfy, for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

and we introduce the exact (α, Γ)-descent, an iterative algorithm relying on a certain function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$. This algorithm is described as follows: given an initial measure $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi_\alpha(\mu_1 k) < \infty$ and $\kappa \in \mathbb{R}$, the iterative sequence of probability measures $(\mu_n)_{n \in \mathbb{N}^\star}$ is defined by setting

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \,, \qquad n \in \mathbb{N}^\star \,,$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$, we have set

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \quad \text{and} \quad b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y) \,.$$

We are able to motivate the formulation of this algorithm by considering the particular case where given $\eta > 0$, the function Γ is of the form

$$\Gamma(v) = e^{-\eta v} \,. \tag{1.15}$$

In that case, applying the transition $\mu \mapsto \mathcal{I}_\alpha(\mu)$ corresponds to performing one step of the (infinite-dimensional) Entropic Mirror Descent algorithm with the α-divergence as objective function and with a learning rate η [and we refer to Hsieh, Liu, and Cevher, 2019, Appendix A for some theoretical background on the Infinite-Dimensional Entropic Mirror Descent].

In this light, $b_{\mu,\alpha}(\theta)$ can be understood as the gradient of $\mu \mapsto \Psi_\alpha(\mu k)$. One transition of the exact (α, Γ)-descent then consists in applying a transform function Γ to the translated gradient $b_{\mu,\alpha}(\theta) + \kappa$ and projecting back onto the space of probability measures, which is why we call our approach is *infinite-dimensional* and *gradient-based*. We now describe the main results obtained in Chapter 2 regarding the (α, Γ)-descent.

1.4.1.1 Main results

Theorem 1, the first main result of Chapter 2 (and the first monotonicity result of the thesis) states conditions on Γ and κ so that one iteration of the exact (α, Γ)-descent leads to a monotonic decrease in the α-divergence. These conditions read as follows:

(1.A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v)+1 \geqslant 0, \quad v \in \mathrm{Dom}_\alpha\ .$$

Coming back to the Entropic Mirror Descent, one may for example notice that (1.A2) is satisfied with $\alpha = 1$ and $\eta \in (0,1]$ when Γ is as in (1.15) and as a consequence, we obtain that one iteration of the Entropic Mirror Descent applied to the forward Kullback-Leibler divergence systematically decreases the forward Kullback-Leibler divergence.

Another important consequence of having derived a general condition of the form (1.A2) is that it makes it possible to go beyond the Entropic Mirror Descent framework. Indeed, by letting $\alpha \in \mathbb{R}\setminus\{1\}$, $\eta \in (0,1]$, κ be such that $(\alpha-1)\kappa \geqslant 0$ and

$$\Gamma(v) = [(\alpha-1)v+1]^{\frac{\eta}{1-\alpha}}\ ,$$

one can readily check that Γ satisfies (1.A2). The resulting algorithm for this choice of function Γ is called the *Power Descent* algorithm in the following and the two cases we have just mentioned are summarised in Table 1.1 below.

TABLE 1.1: Examples of allowed (Γ, κ) in the (α, Γ)-descent according to Theorem 1.

Divergence considered	Possible choices for (Γ, κ)	
Forward KL ($\alpha = 1$)	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0,1]$	any κ
α-*divergence* with $\alpha \in \mathbb{R}\setminus\{1\}$	$\Gamma(v) = [(\alpha-1)\,v+1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0,1]$	$(\alpha-1)\kappa \geqslant 0$

Under our assumptions, the sequence $(\Psi_\alpha(\mu_n k))_{n\geqslant 1}$ is decreasing and also happens to be bounded from below, which implies its convergence. The results that follow then investigate more precisely the convergence of the algorithm.

Firstly, by strengthening the conditions on Γ (i.e. notably assuming that the function Γ is L-smooth), we obtain in Theorem 2 an $O(1/N)$ convergence rate for the exact (α, Γ)-descent of the form: for all $N \in \mathbb{N}^\star$,

$$\Psi_\alpha(\mu_N k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right]\ , \tag{1.16}$$

where the constants $L_{\alpha,1}$, $L_{\alpha,2}$ and $L_{\alpha,3}$ depend on the function Γ and are assumed to be finite. Here $\mu^\star$ is such that $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$ where $\mathrm{M}_{1,\mu_1}(\mathsf{T})$ is the set of probability measures dominated by μ_1 and we have defined $\Delta_1 = \Psi_\alpha(\mu_1 k) - \Psi_\alpha(\mu^\star k)$ as well as $KL(\mu^\star || \mu_1) = \int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu^\star}{\mathrm{d}\mu_1}\right) \mathrm{d}\mu^\star$.

Secondly, by applying the results from Theorem 2 to the Entropic Mirror Descent and the Power Descent, Theorem 3 states, under the assumption that $\theta \mapsto b_{\mu,\alpha}(\theta)$ is bounded by a constant $|b|_{\infty,\alpha}$ independent of μ, that (1.16) holds for all $N \in \mathbb{N}^\star$ when:

- $\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$ and κ is any real number (Entropic Mirror Descent),
- $\Gamma(v) = [(\alpha-1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$, $\alpha > 1$ and $\kappa > 0$ (Power Descent).

To put these results into perspective, letting $J \in \mathbb{N}^\star$, $(\theta_1, \dots, \theta_J) \in \mathsf{T}^J$ and setting $\mu_1 = J^{-1} \sum_{j=1}^J \delta_{\theta_j}$, we consider in Example 4 the case of the (this time finite-dimensional) Entropic Mirror Descent with $\alpha = 1$ and we obtain the following convergence rate for all $\eta \in (0, 1)$

$$\Psi_\alpha(\mu_N) - \Psi_\alpha(\mu^\star) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2 \log J}|b|_{\infty,\alpha}}{(1-\eta)N} \,.$$

Thus, for a constant learning rate $\eta \in (0, 1)$, the dominant term with respect to the dimension J of the simplex is in $\log J$ so that we achieve an overall $O(\log(J)/N)$ convergence rate. This improves on standard Mirror Descent results, which under similar assumptions typically only provide an $O(\sqrt{J/N})$ and $O(\sqrt{\log(J)/N})$ rate respectively for the Projected Gradient Descent and Entropic Mirror Descent by letting the learning rate be proportional to $1/\sqrt{N}$, N being fixed (see Beck and Teboulle, 2003 or Bubeck, 2015, Theorem 4.2.).

Note also that when deriving our $O(1/N)$ rate, another improvement is that we did not require the objective function to be smooth, as opposed to accelerated versions of the Mirror Descent (e.g. Mirror Prox, see Nemirovski, 2004 or Bubeck, 2015, Theorem 4.4.) that yield an $O(1/N)$ convergence rate.

Lastly, the case $\alpha < 1$ for the Power Descent being trickier, we handle it separately in Theorem 4: under the assumption that $(\mu_n)_{n \geqslant 1}$ weakly converges towards a certain $\mu^\star$, as well as (1.A3) below

(1.A3)

(i) T is a compact metric space and $\mathcal{T}$ is the associated Borel σ-field;

(ii) for all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;

(iii) we have $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} k(\theta, y) \times \sup_{\theta' \in \mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) < \infty$.

we obtain the convergence of $(\Psi_\alpha(\mu_n k))_{n\geqslant 1}$ towards $\Psi_\alpha(\mu^\star k)$ where we establish that $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$ and this concludes our theoretical results on the exact (α, Γ)-descent.

As the exact (α, Γ)-descent involves intractable integrals, notice that a practical version of this algorithm will require approximations. We thus resort to a stochastic version of the exact (α, Γ)-descent that builds a sequence $(\hat{\mu}_n)_{n\geqslant 1}$ via an unbiased Importance Sampling estimate $\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta)$ of $b_{\hat{\mu}_n,\alpha}(\theta)$ at each time n, that is $\hat{\mu}_{n+1}(\mathrm{d}\theta) = \hat{\mathcal{I}}_{\alpha,M}(\hat{\mu}_n)(\mathrm{d}\theta) \propto \hat{\mu}_n(\mathrm{d}\theta)\Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta) + \kappa)$, M being the number of samples used in the Importance Sampling estimator. Complementary theoretical results are then proved in the form of Theorem 5, Theorem 7 and Proposition 10, that we briefly detail below, before presenting the main conclusions of numerical experiments.

- Theorem 5 and Theorem 7 focus on the Entropic Mirror Descent and derive bounds under minimal assumptions. More precisely, the former is an adaptation to our framework of the classical result for Stochastic Entropic Mirror Descent from Nemirovski et al., 2009. This result yields an $O(1/\sqrt{N})$ bound on $\mathbb{E}[\Psi_\alpha(N^{-1}\sum_{n=1}^N \hat{\mu}_n k) - \Psi_\alpha(\mu^\star k)]$ for a constant learning rate that is proportional to $1/\sqrt{N}$, the number of iterations N being fixed in advance. On the other hand, the latter provides a bound on $\mathbb{E}[\Psi_\alpha(N^{-1}\sum_{n=1}^N \hat{\mu}_n k) - \Psi_\alpha(\mu^\star k)]$ of the form $O(1/N) + O(1/\sqrt{M})$, all the while keeping the learning rate constant throughout the algorithm (e.g. $\eta \in (0, 1)$ for the forward Kullback-Leibler).

- Proposition 10 deals with the Power Descent algorithm and establishes the total variation convergence of $\hat{\mathcal{I}}_{\alpha,M}(\mu)$ towards $\mathcal{I}_\alpha(\mu)$ as M goes to infinity for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\alpha \in \mathbb{R} \setminus \{1\}$.

1.4.1.2 Empirical results

For the numerical results, we let $J \in \mathbb{N}^\star$ and we consider the case where $\hat{\mu}_1$ is a weighted sum of J dirac measures, that is: $\hat{\mu}_1 = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$ with $\theta_1, \ldots, \theta_J \in \mathsf{T}$ and $\boldsymbol{\lambda} \in \mathcal{S}_J$, where $\mathcal{S}_J$ is the simplex of $\mathbb{R}^J$ and is defined by

$$\mathcal{S}_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, \ldots, J\}\,,\ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}\,. \quad (1.17)$$

In this case, for any kernel K of our choice, the (α, Γ)-descent procedure simplifies and provides an update formula for the mixture weights of the corresponding mixture model $\hat{\mu}_1 k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y)$: an immediate induction yields that for every $n \in \mathbb{N}^\star$, $\hat{\mu}_n$ can be expressed as $\hat{\mu}_n = \sum_{j=1}^J \lambda_{j,n}\delta_{\theta_j}$ where $\boldsymbol{\lambda}_n = (\lambda_{1,n}, \ldots, \lambda_{J,n}) \in \mathcal{S}_J$ satisfies the initialisation $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}$ and the update formula: for all $n \in \mathbb{N}^\star$ and all

$j \in \{1, \dots, J\}$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_i) + \kappa)} .$$

Here, the unbiased estimate $\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)$ of $b_{\mu,\alpha}(\theta_j)$ is chosen to be

$$\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta, Y_{m,n+1})}{\hat{\mu}_n k(Y_{m,n+1})} f'_\alpha\left(\frac{\hat{\mu}_n k(Y_{m,n+1})}{p(Y_{m,n+1})}\right)$$

with $Y_{1,n+1}, \dots, Y_{M,n+1} \overset{\text{i.i.d}}{\sim} \hat{\mu}_n k$ conditionally on $\mathcal{F}_n$ and where $\mathcal{F}_1 = \emptyset$ and $\mathcal{F}_n = \sigma(Y_{1,2}, \dots, Y_{M,2}, \dots, Y_{1,n}, \dots, Y_{M,n})$ for $n \geqslant 2$.

This procedure is summarised in Algorithm 6 of Chapter 2 and we now make an important remark: one main strength of the algorithm we have designed is that it does not require any information on how the $\{\theta_1, \dots, \theta_J\}$ have been obtained in order to infer the optimal weights, as it draws information from samples that are generated from $\hat{\mu}_n k$. Then, since the procedure leaves $\{\theta_1, \dots, \theta_J\}$ unchanged throughout the optimisation of the mixture weights, a natural idea is to combine this algorithm with an *Exploration step* of our choice that modifies the parameter set (Algorithm 7).

While any choice of Exploration step could be envisioned, we settle for a simple exploration step in our numerical experiments (it is detailed in Section 2.4) and we focus on investigating how the choice of α and Γ plays a role in practice.

The key message from our numerical experiments is the following: as the dimension increases the Power Descent with $\alpha < 1$ is a more scalable alternative to the Entropic Mirror Descent, which sheds light on the importance of going beyond the traditional Entropic Mirror Descent framework from the optimisation literature.

We visually support that claim in Figure 1.5 below where we target a mixture density multiplied by a constant Z and where the Entropic Mirror Descent fails as the dimension increases compared to the Power Descent (these figures correspond to Figure 2.1 and 2.2 in Section 2.4).

In addition, we also consider a Bayesian Logistic Regression on a real-world dataset in dimension 56: Figure 1.6 shows that our Power Descent algorithm has the ability to outperform a typical computationally-equivalent Adaptive Importance Sampling algorithm (see Section 2.4 for details).

This concludes our overview of Chapter 2, in which we build the novel framework of the (α, Γ)-descent and demonstrate empirically the benefit of going beyond the Entropic Mirror Descent framework for mixture weights optimisation by using the Power Descent algorithm instead. Let us now advance to summarising the content of Chapter 3.

FIGURE 1.5: Plotted on the first line is the VR bound for the Power Descent and the Entropic Mirror Descent with $\alpha = 0.5$ (0.5-Power and 0.5-Mirror) while the second line is the Log-likelihood for the Power Descent with $\alpha = 0.5$ and the Entropic Mirror Descent with $\alpha = 1$ (0.5-Power and 1-Mirror). The dimension d varies in $\{8, 16, 32\}$ from left to right and the plotted quantities are averaged over 100 replicates.

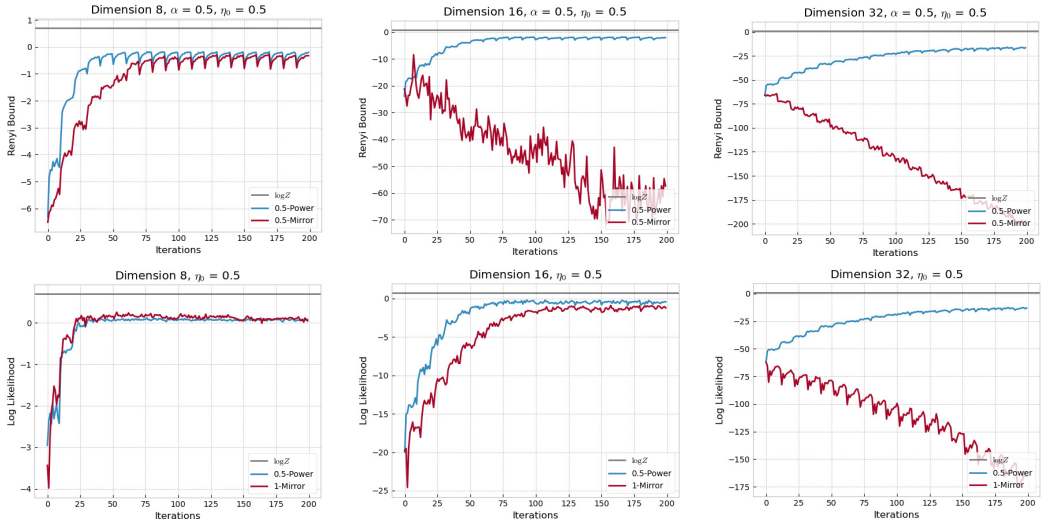
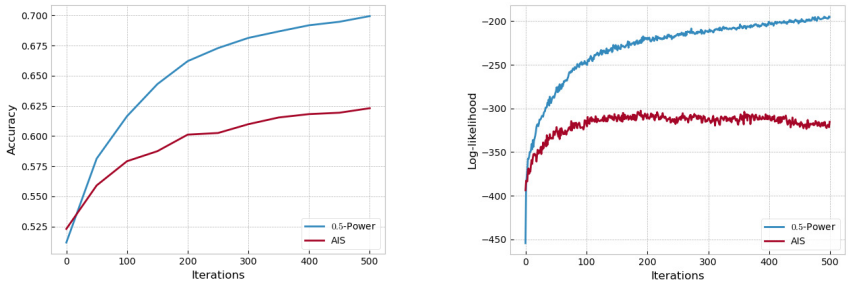


FIGURE 1.6: Plotted are the average Accuracy and Log-likelihood computed over 100 replicates for Bayesian Logistic Regression on the Covertype dataset for the Power Descent with $\alpha = 0.5$ (0.5-Power) and a computationally-equivalent Adaptive Importance Sampling algorithm (AIS).



1.4.2 Chapter 3: Mixture weights optimisation with the α-divergence

Thanks to Chapter 2, we now have access to the Power Descent, an algorithm that permits us to optimise the mixture weights of mixture models by α-divergence minimisation for all $\alpha \in \mathbb{R} \setminus \{1\}$, regardless of the underlying distribution of its mixture components parameters.

However, one may remark that the convergence of the (exact) Power Descent towards the global optimum when $\alpha < 1$ in Theorem 4 is guaranteed under the assumption that $(\mu_n)_{n \geqslant 1}$ weakly converges towards a certain $\mu^\star$, that is later proved to satisfy $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$.

This is a much stronger assumption compared to the ones made in Theorem 3 for the Entropic Mirror Descent and the Power Descent when $\alpha > 1$ and in which

convergence rates are available. Since the case $\alpha < 1$ is useful to tackle the challenges of forward Kullback-Leibler optimisation, one would be interested in alleviating this specific assumption to obtain a full proof of convergence.

Furthermore, one may also notice that the Power Descent is defined for all $\alpha \neq 1$, and thus the important case $\alpha = 1$ in (1.14) corresponding to forward Kullback-Leibler minimisation is not handled by this algorithm.

The aim of Chapter 3 is to cover both of these aspects. In particular, studying the extension of the Power Descent to the case $\alpha = 1$ will also lead us to further look into the connections between the Power Descent and the Entropic Mirror Descent.

We now describe the main results obtained in Chapter 3.

1.4.2.1 Main results

The first result of Chapter 3 establishes the full proof of the global convergence towards the optimum for the mixture weights when $\alpha < 1$.

Letting $\Theta = (\theta_1, \dots, \theta_J) \in \mathsf{T}^J$ be fixed and setting $\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$ for all $\boldsymbol{\lambda} \in \mathcal{S}_J$, Theorem 10 indeed considers the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by $\mu_1 = \mu_{\boldsymbol{\lambda}}$ and (1.4.1). This amounts to studying the sequence $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ satisfying the initialisation $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}$ and the update formula:

$$\boldsymbol{\lambda}_{n+1} = \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_n) \ , \ n \in \mathbb{N}^\star \ ,$$

where we have set $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ for every $n \in \mathbb{N}^\star$ and where for all $\boldsymbol{\lambda} \in \mathcal{S}_J$,

$$\mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}) := \left(\frac{\lambda_j \Gamma(b_{\mu_{\boldsymbol{\lambda}},\alpha}(\theta_j) + \kappa)}{\sum_{\ell=1}^J \lambda_\ell \Gamma(b_{\mu_{\boldsymbol{\lambda}},\alpha}(\theta_\ell) + \kappa)} \right)_{1\leqslant j\leqslant J}$$

with $(\alpha - 1)\kappa \geqslant 0$ and $\Gamma(v) = [(\alpha-1)v + 1]^{\frac{\eta}{1-\alpha}}$ for all $v \in \mathrm{Dom}_\alpha$.

The convergence towards the optimal mixture weights when $\alpha < 1$ is then derived in Theorem 10 under the assumption that $\{K(\theta_1, \cdot), \dots, K(\theta_J, \cdot)\}$ are linearly independent, paired up with (1.A1) and (1.A4) [where (1.A4) given below corresponds to (1.A3) in the simplified case where μ is a sum of Dirac measures].

(1.A4) (i) For all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;

(ii) we have $\displaystyle\int_{\mathsf{Y}} \max_{1\leqslant j\leqslant J} k(\theta_j, y) \times \max_{1\leqslant j'\leqslant J} \left(\frac{k(\theta_{j'}, y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) < \infty.$

If $\alpha = 0$, we assume in addition that $\displaystyle\int_{\mathsf{Y}} \max_{1\leqslant j\leqslant J} \left|\log\left(\frac{k(\theta_j, y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) < \infty.$

In terms of assumptions, notice that (1.A1) and (1.A4) are mild and that since the objective function Ψ_α depends on $\boldsymbol{\lambda}$ through $\mu_{\boldsymbol{\lambda}} K$, an identifiably condition was to be expected in Theorem 10 in order to achieve the convergence of the sequence $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$. One may then observe that this identifiably condition notably holds when K is a d-dimensional Gaussian kernel under the assumption that the $\theta_1, ..., \theta_J$ are

full-rank with $J \leqslant d$.

Next in line after the full proof of convergence for mixture weights when $\alpha < 1$ in the Power Descent is the extension of this algorithm to the case $\alpha = 1$. Proposition 19 then establishes that under typical convergence and differentiability assumptions, the Power Descent can be extended to the case $\alpha = 1$ and that we recover an Entropic Mirror Descent applied to the objective function Ψ_1.

As we already know from Theorem 3, this algorithm enjoys an $O(1/N)$ convergence rate. Yet, Proposition 19 shows that a deeper connection runs between the Power Descent and the Entropic Mirror Descent beyond the (α, Γ)-descent framework.

To better understand that connection, our idea is then to look at first-order approximations by considering the case where $\alpha \in \mathbb{R} \setminus \{1\}$ and $b_{\mu,\alpha}(\theta) \approx \mu(b_{\mu,\alpha})$ for all $\theta \in \mathsf{T}$. As a result of these calculations, letting $\eta > 0$ and $\mu \in \mathrm{M}_1(\mathsf{T})$, we find that first-order approximations for one transition for the Power Descent and for the Entropic Mirror Descent applied to $\mu \mapsto \Psi_\alpha(\mu k)$ are given by

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \mu(\mathrm{d}\theta)\left[1 - \frac{\eta}{\alpha - 1}\frac{b_{\mu,\alpha}(\theta) - \mu(b_{\mu,\alpha})}{\mu(b_{\mu,\alpha}) + \kappa + 1/(\alpha - 1)}\right] \qquad \text{(Power Descent)}$$

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \mu(\mathrm{d}\theta)\left[1 - \eta\left(b_{\mu,\alpha}(\theta) - \mu(b_{\mu,\alpha})\right)\right] \qquad \text{(Entropic Mirror Descent).}$$

Thus, these two approximations do not coincide, which brings us to introduce instead the *Renyi Descent* one-step transition

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta)\exp\left[-\eta\frac{b_{\mu,\alpha}(\theta)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right]}{\mu\left(\exp\left[-\eta\frac{b_{\mu,\alpha}}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right]\right)} \qquad \text{(Renyi Descent),}$$

since it shares the same first-order approximation as the Power Descent.

Here, the name of this one-step transition comes from the fact that it can been seen as an Entropic Mirror Descent transition for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$ applied this time to the objective function $\mu \mapsto \Psi_\alpha^{AR}(\mu k; p)$, where for all probability density q with respect to ν on $(\mathsf{Y}, \mathcal{Y})$ and all $\alpha \in \mathbb{R} \setminus \{0, 1\}$ we have set

$$\Psi_\alpha^{AR}(q; p) := \frac{1}{\alpha(\alpha - 1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right).$$

Letting $\kappa = 0$ and $p = p(\cdot, \mathscr{D})$ in $\Psi_\alpha^{AR}(q; p)$, we then recognise the VR bound $\mathcal{L}_\alpha(q; \mathscr{D})$ up to a proportional constant $-\alpha^{-1}$, hence the name Renyi Descent.

Contrary to the Power Descent, the Renyi Descent enjoys the typical $O(1/\sqrt{N})$ convergence rate results from the optimisation literature for Entropic Mirror Descent algorithms, which we further improve to an $O(1/N)$ convergence rate in Theorem 11 (and for clarity, Table 1.2 below recapitulates the theoretical contributions from Chapter 3 compared to Chapter 2).

TABLE 1.2: Comparison between the theoretical results in Chapter 3 and in Chapter 2

	Power Descent	Renyi Descent
Chapter 2	$\alpha < 1$: convergence under restrictive assumptions; $\alpha > 1$: $O(1/N)$ convergence rate	not covered
Chapter 3	$\alpha < 1$: full proof of convergence for mixture weights; extension to $\alpha = 1$	$O(1/N)$ convergence rate

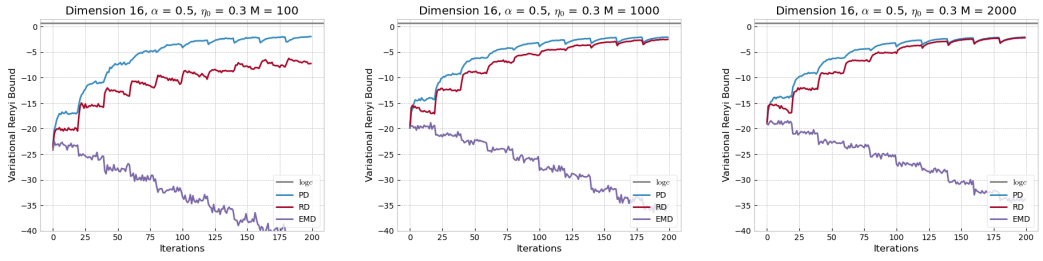
We now present some numerical results.

1.4.2.2 Empirical results

Following Chapter 2, we approximate the Power Descent and the Renyi Descent using Importance Sampling estimates, a procedure that written explicitly in Algorithm 6 and 9 and not detailed here for the sake of conciseness. We then pair them up with the same Exploration step as in Chapter 2 and we target a mixture density of Gaussian distributions multiplied by a constant c (we refer to Section 3.5 for details regarding our numerical experiments).

The plot below, which corresponds to Figure 3.1, compares the Power Descent and the Renyi Descent in dimension 16 as the number of samples M used in the Importance Sampling estimates increases. It illustrates the theoretical link between the two algorithms (and the Entropic Mirror Descent applied to Ψ_α is provided as a reference).

FIGURE 1.7: Plotted is the average VR bound for the Power Descent (PD), the Renyi Descent (RD) and the Entropic Mirror Descent applied to Ψ_α (EMD) in dimension $d = 16$ computed over 100 replicates with $\eta_0 = 0.3$ and $\alpha = 0.5$ and an increasing number of samples M.



We have thus found a novel algorithm for mixture weights optimisation that is close to the Power Descent in the sense that it shares the same first-order approximation. Theoretically-wise (and contrary to the Power descent when $\alpha < 1$), it benefits from the Entropic Mirror Descent optimisation literature so that $O(1/\sqrt{N})$ convergence rates hold, which we improve to $O(1/N)$ convergence rates.

Note that a practical use of the Power Descent and of the Renyi Descent for mixture weights optimisation involves unbiased estimates of $b_{\mu,\alpha}(\theta) + \kappa$ for the former while the latter uses biased estimates of $b_{\mu_n,\alpha}(\theta)/(\mu_n(b_{\mu_n,\alpha}) + \kappa + 1/(\alpha - 1))$. Finding

which approach is most suitable between biased and unbiased α-divergence minimisation remains an open issue in the literature (Dieng et al., 2017; Li and Turner, 2016; Geffner and Domke, 2020a; Geffner and Domke, 2020b; Dhaka et al., 2021) that is beyond the scope of this thesis.

Nevertheless, our work provides insights on potential links between unbiased and biased α-divergence methods, as both the unbiased Power Descent and the biased Renyi Descent share the same first-order approximation.

So far in Chapter 2 and Chapter 3 we have insisted on the fact that our algorithms could be paired up with any Exploration step we can think of and we have chosen to keep the Exploration step simple in our numerical experiments. One may then wonder if we can exhibit some examples of Exploration step that can successfully be combined with our framework for mixture weights optimisation. This is what we focus on in Chapter 4, which we summarise right after.

1.4.3 Chapter 4: Monotonic α-divergence minimisation

In Chapter 4, we aim at building a complete framework for mixture models optimisation that enables a systematic decrease of the α-divergence at each step. This means that on top of our updates for mixture weight optimisation, we want to derive update formulas for the mixture components parameters as well.

As we shall explain thereafter, the particularity of our work in Chapter 4 is that it will permit us to *simultaneously* optimise the weights and the components parameters of a given mixture model.

1.4.3.1 Main results

The starting point of our approach in Chapter 4 is to go back to the typical Variational Inference parametric family (1.8), that is

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \ ,$$

and to construct a sequence $(\theta_n)_{n\geqslant 1}$ so that $\Psi_\alpha(k(\theta_{n+1}, \cdot)) \leqslant \Psi_\alpha(k(\theta_n, \cdot))$ at time n.

We do so in Theorem 12, where under (1.A1), for all $\alpha \in [0, 1)$ and all initial $\theta_1 \in \mathsf{T}$ such that $\Psi_\alpha(k(\theta_1, \cdot)) < \infty$, we establish that a sufficient condition to obtain a monotonic decrease in the α-divergence at each step is for the sequence $(\theta_n)_{n\geqslant 1}$ to satisfy for all $n \geqslant 1$,

$$\int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \log\left(\frac{k(\theta_{n+1}, y)}{k(\theta_n, y)}\right) \nu(\mathrm{d}y) \leqslant 0 \ . \tag{1.18}$$

As a result, we deduce in Corollary 27 that (1.18) holds for all $n \geqslant 1$ when the sequence $(\theta_n)_{n\geqslant 1}$ is iteratively defined by

$$\theta_{n+1} = \mathrm{argmax}_{\theta\in\mathsf{T}} \int_{\mathsf{Y}} k(\theta_n, y)^{\alpha} p(y)^{1-\alpha} \log(k(\theta, y))\nu(\mathrm{d}y) \ , \quad n \geqslant 1 \ . \tag{1.19}$$

Strikingly, the above update formula is written as a maximisation problem involving the logarithm of the kernel k. This implies that it can be used to derive simple update rules for $(\theta_n)_{n\geqslant 1}$ for some well-chosen kernel k, a fact that we illustrate over several examples, namely Example 8 (Gaussian), Example 9 (Student) and Example 10 (Mean-field).

As it turned out, we can also obtain alternative schemes satisfying (1.18) beyond the intuitive update (1.19). These results require additional smoothness conditions on the sequence of functions $(g_n)_{n\geqslant 1}$ where for all $n \geqslant 1$ and all $\theta \in \mathsf{T} = \mathbb{R}^d$

$$g_n(\theta) = c_n \int_{\mathsf{Y}} \frac{k(\theta_n, y)^{\alpha} p(y)^{1-\alpha}}{\alpha - 1} \log\left(\frac{k(\theta, y)}{k(\theta_n, y)}\right) \nu(\mathrm{d}y) \tag{1.20}$$

and $(c_n)_{n\geqslant 1}$ is a positive sequence. Indeed, assuming that g_n is β_n-smooth and letting $(\gamma_n)_{n\geqslant 1}$ be valued in $(0, 1]$, Corollary 28 states that the sequence $(\theta_n)_{n\geqslant 1}$ iteratively defined by

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \nabla g_n(\theta)|_{\theta=\theta_n} \ , \quad n \geqslant 1 \ , \tag{1.21}$$

satisfies (1.18) for all $n \geqslant 1$. An important remark is then that under common differentiability assumptions, we can write: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$

$$\nabla g_n(\theta) = c_n \int_{\mathsf{Y}} \frac{k(\theta_n, y)^{\alpha} p(y)^{1-\alpha}}{\alpha - 1} \nabla(\log k(\theta, y))\nu(\mathrm{d}y) \ ,$$

so that, letting $p = p(\cdot, \mathscr{D})$, the two cases $c_n = 1$ and $c_n = (\int_{\mathsf{Y}} k(\theta_n, y)^{\alpha} p(y)^{1-\alpha} \nu(\mathrm{d}y))^{-1}$ at time n correspond to Gradient Descent steps applied to $\theta \mapsto \Psi_{\alpha}(k(\theta, \cdot); \mathscr{D})$ and $\theta \mapsto -\mathcal{L}_{\alpha}(k(\theta, \cdot); \mathscr{D})$ respectively with a learning policy proportional to $(\gamma_n \beta_n^{-1})_{n\geqslant 1}$.

We are thus able to connect our approach to typical Gradient Descent techniques for α-divergence and Renyi's α-divergence optimisation, especially since the conditions on $(g_n)_{n\geqslant 1}$ are met for a Gaussian kernel (Example 11).

At this stage, we have proposed several ways to carry out parameter optimisation for $\theta \mapsto \Psi_{\alpha}(k(\theta, \cdot))$ by decreasing the α-divergence at each step. We now move on to our main goal, which is to extend the monotonicity property to the case of mixture models: given $J \in \mathbb{N}^{\star}$, we consider the approximating family given by

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k(y) = \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} \ ,$$

where we use the notation $\Theta = (\theta_1, \dots, \theta_J) \in \mathsf{T}^J$ and $\mu_{\boldsymbol{\lambda},\Theta} = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$ for all $\boldsymbol{\lambda} \in \mathcal{S}_J$ and all $\Theta \in \mathsf{T}^J$. This approximating family simplifies to (1.8) when $J = 1$, and we are thus interested in treating the case $J > 1$.

Theorem 13, our first result for mixture models, generalises Theorem 12 and establishes sufficient conditions on both the mixture weights and the mixture components parameters leading to a monotonic decrease in the α-divergence at each step. More precisely, let us denote $\boldsymbol{\lambda}_n = (\lambda_{j,n})_{1 \leqslant j \leqslant J}$ and $\Theta_n = (\theta_{j,n})_{1 \leqslant j \leqslant J}$ for all $n \geqslant 1$. We also introduce the shorthand notation $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_{j,n}}$ and

$$\gamma_{j,\alpha}^n(y) = k(\theta_{j,n}, y) \left(\frac{\mu_n k(y)}{p(y)} \right)^{\alpha - 1} \tag{1.22}$$

for $\alpha \in [0, 1)$, all $j = 1 \dots J$, all $n \geqslant 1$ and all $y \in \mathsf{Y}$. Lastly, we define $\mathcal{S}_J^+ = \{\boldsymbol{\lambda} \in \mathcal{S}_J \, : \, \forall j \in \{1, \dots, J\}, \, \lambda_j > 0\}$.

Using these notation and under (1.A1), Theorem 13 states that for all $\alpha \in [0, 1)$ and all initial parameter set $(\boldsymbol{\lambda}_1, \Theta_1) \in \mathcal{S}_J^+ \times \mathsf{T}^J$ such that $\Psi_\alpha(\mu_1 k) < \infty$, the sequence $(\boldsymbol{\lambda}_n, \Theta_n)_{n \geqslant 1}$ defined iteratively by: $n \geqslant 1$,

$$\int_{\mathsf{Y}} \sum_{j=1}^J \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log \left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}} \right) \nu(\mathrm{d}y) \leqslant 0 \tag{1.23}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^J \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log \left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \leqslant 0 \tag{1.24}$$

ensures a systematic decrease in the α-divergence at each step, that is for all $n \geqslant 1$, we have $\Psi_\alpha(\mu_{n+1} k) \leqslant \Psi_\alpha(\mu_n k)$.

Here, a key property achieved by Theorem 13 is that (1.23) does not depend on Θ_{n+1} and similarly (1.24) does not depend on $\boldsymbol{\lambda}_{n+1}$. As a result and as we announced earlier, Theorem 13 provides a framework to optimise simultaneously the weights and components parameters of a mixture model.

The following theoretical results then focus on finding iterative schemes satisfying (1.23) and (1.24), starting with (1.23) (since the dependency in $\lambda_{j,n+1}$ appearing in (1.23) is simpler than the dependency in $\theta_{j,n+1}$ appearing in (1.24) that is expressed through the kernel k).

• Theorem 14 identifies an update formula for $(\boldsymbol{\lambda}_n)_{n \geqslant 1}$, regardless of the choice of the kernel k. Indeed, under (1.A1) and letting $\alpha \in [0, 1)$, $(\eta_n)_{n \geqslant 1}$ be valued in $(0, 1]$ and κ be such that $(\alpha - 1)\kappa \geqslant 0$, this result establishes that for all initial parameter set $(\boldsymbol{\lambda}_1, \Theta_1) \in \mathcal{S}_J^+ \times \mathsf{T}^J$, the sequence $(\boldsymbol{\lambda}_n, \Theta_n)_{n \geqslant 1}$ defined iteratively such that for all $n \geqslant 1$

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \gamma_{\ell,\alpha}^n(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}, \quad j = 1 \dots J \tag{1.25}$$

and (1.24) holds will satisfy the conditions of Theorem 13.

- Building on the update for the mixture weights from Theorem 14, Corollary 30 and Corollary 31 respectively extend the updates (1.19) and (1.21) for $J = 1$ to the more general case of mixtures models. More specifically, Corollary 30 considers the update at time n given by

$$\theta_{j,n+1} = \text{argmax}_{\theta_j \in \mathsf{T}} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \log(k(\theta_j, y)) \nu(\mathrm{d}y) \ , \quad j = 1 \dots J \tag{1.26}$$

so that the full update from $(\boldsymbol{\lambda}_n, \Theta_n)$ to $(\boldsymbol{\lambda}_{n+1}, \Theta_{n+1})$ can be written as the following optimisation problem

$$(\boldsymbol{\lambda}_{n+1}, \Theta_{n+1}) = \text{argmin}_{\boldsymbol{\lambda} \in \mathcal{S}_j^+, \Theta \in \mathsf{T}^J} \left(h_n(\boldsymbol{\lambda}) + g_n(\Theta) \right)$$

(the definition of the functions h_n and g_n can be found in (4.18) and (4.19) and we refer to Section 4.3.1 and Section 4.3.2 for details about the above formulation).

As for Corollary 31, it sets at time n

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta = \theta_{j,n}} \ , \quad j = 1 \dots J \ , \tag{1.27}$$

where for all $j = 1 \dots J$, $(\gamma_{j,n})_{n \geqslant 1}$ is valued in $(0, 1]$ and given a positive sequence $(c_{j,n})_{n \geqslant 1}$, the set of functions $(g_{j,n})_{n \geqslant 1}$ is defined by: for all $n \geqslant 1$ and all $\theta \in \mathsf{T} = \mathbb{R}^d$,

$$g_{j,n}(\theta) = c_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \tag{1.28}$$

with each function $g_{j,n}$ being assumed to be $\beta_{j,n}$-smooth. Under common differentiability assumptions one can then write: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$

$$\nabla g_{j,n}(\theta) = c_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \nabla \left(\log k(\theta, y) \right) \nu(\mathrm{d}y) \ , \quad j = 1 \dots J \ ,$$

and we recover Gradient Descent steps for α-divergence and Renyi's α-divergence minimisation by setting $c_{j,n} = \lambda_{j,n}$ and $c_{j,n} = \lambda_{j,n} (\int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y))^{-1}$ respectively when $p = p(\cdot, \mathscr{D})$. Observe as a result that $\lambda_{j,n}$ appears as a multiplicative factor by design in both definitions of $c_{j,n}$ above, which could prevent learning in practice for very small values of $\lambda_{j,n}$.

This is where the framework of Corollary 31 comes in handy, as another valid choice for $c_{j,n}$ is $c_{j,n} = (\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y))^{-1}$. In that case, $\lambda_{j,n}$ only appears through $\mu_n k$ in the mixture components parameters update. This property, also shared with the update (1.26) and that will come up again in our numerical experiments, further underlines the importance of having worked under the general conditions on $(\boldsymbol{\lambda}_n, \Theta_n)_{n \geqslant 1}$ stated in Theorem 13.

At this point, one may remark that the conditions on $(\Theta_n)_{n\geqslant 1}$ stated in (1.24) are satisfied by letting the sequence $(\Theta_n)_{n\geqslant 1}$ be constant equal to Θ throughout the algorithm and that in doing so we obtain the Power Descent algorithm (with $\eta_n = \eta/(1-\alpha)$). The updates on the mixture weights in Theorem 14 thus correspond to Power Descent one-step transitions and we have reached our announced goal, which was to pair up the work from Chapter 2 and Chapter 3 with simultaneous components parameters updates.

Furthermore, having recovered the Power Descent algorithm also tells us that it is possible to derive monotonicity result beyond the case $\alpha \in [0, 1)$ from Theorem 13, at the cost of keeping the sequence $(\Theta_n)_{n\geqslant 1}$ constant (recall indeed that our monotonicity results for the Power Descent from earlier chapters hold for all $\alpha \in \mathbb{R}$ and all $\eta \in (0, 1]$). In fact, we show in Proposition 32 that the range of possible values of η can yet again be extended when $\alpha < 0$ in the Power Descent.

Interestingly, the Power Descent is not the only special case that fits into the framework of Theorem 13. Letting $\alpha = 0$, $\kappa = 0$ and $\eta_n = 1$ for all $n \geqslant 1$ in Corollary 30, we recover the Mixture Population Monte Carlo (M-PMC) algorithm from Cappé et al., 2008, which sheds light on the link between our approach and an integrated Expectation-Maximisation algorithm from the Adaptive Importance Sampling literature (see Section 4.3.4).

Finally, practical versions of our algorithms based on our maximisation approach (1.26) and on our gradient-based approach (1.27) are derived for Gaussian Mixture Models by resorting to usual Importance Sampling estimates involving a sequence of samplers $(q_n)_{n\geqslant 1}$ (we refer to Algorithm 13 and 14 for details). While those are the algorithms we will consider in our numerical experiments, we observe that additional practical algorithms can also be derived within the Student's distribution family (e.g. Algorithm 15).

1.4.3.2 Empirical results

We revisit the example that targets a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant c. This time, the approximating family $\mathcal{Q}$ is given by the family of Gaussian Mixture Models in which the mixture weights and the mixture means are updated while the variance matrices are kept constant equal to $\sigma^2 \boldsymbol{I}_d$.

We study the impact of four parameters: (i) the learning rate policy $(\eta_n)_{n\geqslant 1}$ (ii) the constant κ (iii) the sequence of samplers $(q_n)_{n\geqslant 1}$ in the Importance Sampling estimates and (iv) the value of α. We refer to Section 4.4 for details and we now present our main conclusions.

- We first investigate (i), (ii) and (iii) and to do so, we set $\alpha = 0$ and $\sigma = 1$. Now letting the learning policy $(\eta_n)_{n\geqslant 1}$ be constant equal to η, we consider two versions of Algorithm 13:

(a) The M-PMC(η, κ) algorithm, that uses the best sampler at time n in the Importance Sampling estimates i.e. $\mu_n k$. This family of algorithms notably includes the M-PMC algorithm from Cappé et al., 2008, which also chooses this sampler in practice and is plotted as a reference algorithm under the name M-PMC$(1, 0)$.

(b) The UP-PMC(η, κ) algorithm, that uses a uniform sampler at time n.

A comparison between M-PMC(η, κ) and UP-PMC(η, κ) for different values of η and κ can then be found in Figure 1.8 below (corresponding to Figure 4.2 in Chapter 4), that we now comment. [Note as a side remark that in these plots, we are interested in the log Mean-Squared error LogMSE, however additional plots for log-likelihood (the VR bound when $\alpha = 0$) estimation can be found in Figure 4.1.]

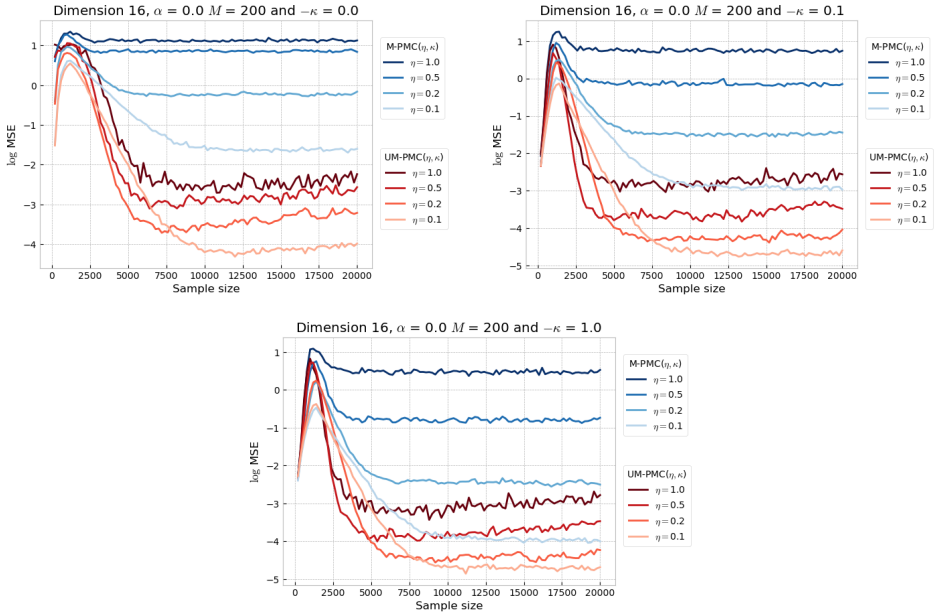


FIGURE 1.8: LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms in dimension $d = 16$ for $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$ (over 200 replicates).

The key insight from these plots is that the choice of κ, η and $(q_n)_{n \geqslant 1}$ heavily impacts the convergence of the algorithm.

Focusing first on the effect of κ and η in the M-PMC(η, κ) curves, we obtain that M-PMC$(1, 0)$ underperforms due to a learning rate η that is too high in dimension 16. As a result, it discriminates too much between the mixture weights and by setting some of them to 0 in the early stages it prevents the algorithm from visiting the mixture components parameters space well. Lowering the value of η thus results in improved results and we observe that we can even further mitigate the issue of setting mixture weights to 0 too early by enforcing a positivity of the weights via κ.

In the same vein, it appears that using a uniform sampler ensures a fairer sampling among all the components parameters throughout the algorithm and improves the estimates of the intractable integrals appearing in Algorithm 13.

Our general framework for mixture models optimisation thus strongly improves on the M-PMC algorithm via the hyperparameters κ, η and the choice of $(q_n)_{n\geqslant 1}$ and we finish up by presenting results linked to (iv).

- Regarding (iv), Figure 1.9 below (Figure 4.3 from Chapter 4) compares the performances of the UM-PMC(η, κ) algorithm depending on the value of α. Importantly, we also include, under the name RGD(η, κ), numerical experiments that use Renyi's α-divergence gradient-based updates for the mixture components parameters updates (that is there is an additional multiplicative factor $\lambda_{j,n}$ in those updates, see (4.28) and Algorithm 14). Furthermore, the M-PMC algorithm (M-PMC$(1, 0)$) and the PIMAIS algorithm (Martino et al., 2017) are provided as a reference.

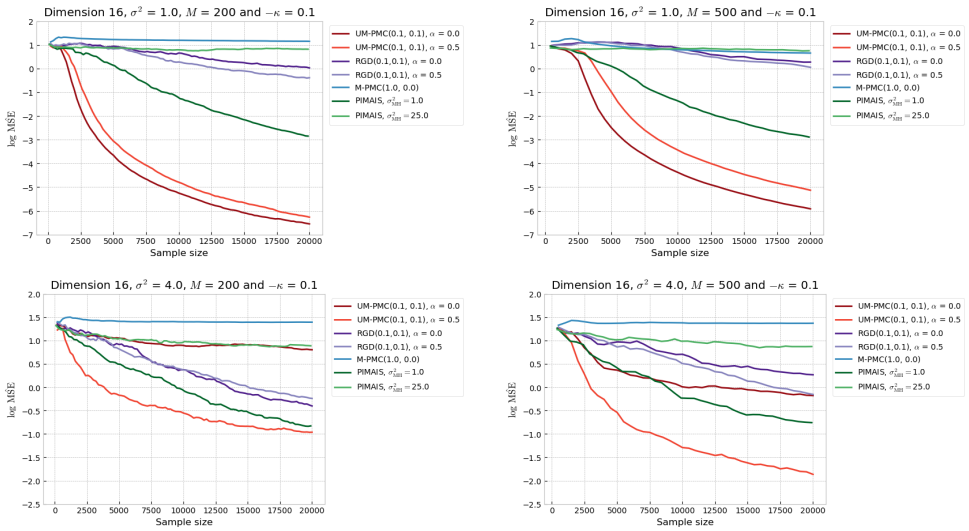


FIGURE 1.9: LogM$\hat{\text{S}}$E for UM-PMC(η, κ) and RGD(η, κ) in dimension $d = 16$ for $\alpha \in \{0., 0.5\}$, $\sigma^2 \in \{1, 4\}$, $\eta = 0.1$ and $-\kappa = 0.1$ compared with the PIMAIS algorithm and the M-PMC(1., 0.) algorithm (over 200 replicates).

We find that there are benefits of going beyond the case $\alpha = 0$ when the value of σ changes in our numerical experiments. In addition, the plots for RGD(η, κ) illustrate the fact that the multiplicative factor $\lambda_{j,n}$ appearing in components parameters updates (and that we already discussed from a theoretical point of view) does have a negative impact on the speed of convergence of the algorithm numerically.

This wraps up our overview of Chapter 4, in which we introduce a novel methodology to carry out mixture model optimisation via α-divergence minimisation. Our approach enables simultaneous updates for both the weights and components parameters and can be linked to Gradient Descent schemes. It notably recovers the Power Descent algorithm and the M-PMC algorithm as special cases. Finally, empirical evidence shows that our methodology can be used to enhance both the M-PMC algorithm and Gradient Descent schemes and we also demonstrate the importance of having some flexibility in the choice of α.

2

Infinite-dimensional α-divergence minimisation

The work presented in this chapter corresponds to the paper entitled "Infinite-dimensional gradient-based descent for Alpha-divergence minimisation" (Daudel, Douc, and Portier, 2021) that has been accepted in the Annals of Statistics.

2.1 Introduction

As stated in Chapter 1, our objective in this thesis is to figure out ways to enrich the variational approximating family $\mathcal{Q}$ beyond the framework of Black-Box Variational Inference where

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and θ is typically tuned through Stochastic Gradient Descent optimisation, all the while maintaining efficient optimisation.

Hierarchical Variational Inference methods (Ranganath, Tran, and Blei, 2016; Yin and Zhou, 2018; Titsias and Ruiz, 2019) do so by putting a prior on the variational parameter θ, that is itself parameterised by a certain ϕ to be optimised by Stochastic Gradient Descent so that

$$\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \lambda_\phi(\theta) k(\theta, y) \mathrm{d}\theta \ : \ \phi \in A\right\} .$$

In the spirit of Hierarchical Variational Inference methods, we will consider in this chapter the approximating family given by

$$\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M}\right\} ,$$

where M is a convenient subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on T.

In contrast with already-existing Hierarchical Variational Inference methods, our approach does not assume that μ is parameterised and for this reason we call it *infinite-dimensional*. We can motivate the formulation of this extended approximating family by noticing that it is large enough to include mixture models (as it corresponds to choosing μ as a weighted sum of Dirac measures).

Consequently, we aim at designing an iterative algorithm that performs infinite-dimensional α-divergence minimisation with respect to μ between the approximate variational density and the posterior density. For this purpose, let us introduce some notation that will be used in this chapter (and in the rest of the thesis too) and state more formally the optimisation problem we want to solve in this specific chapter.

Notation and problem statement Let $(\mathsf{Y}, \mathcal{Y}, \nu)$ be a measured space, where ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$ and let $(\mathsf{T}, \mathcal{T})$ be a measurable space. Furthermore, let p be a measurable positive function on $(\mathsf{Y}, \mathcal{Y})$ and for all probability density q with respect to ν on $(\mathsf{Y}, \mathcal{Y})$, we define

$$\Psi_\alpha(q; p) := \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ ,$$

where f_α is the convex function on $(0, +\infty)$ defined by $f_0(u) = u - 1 - \log(u)$, $f_1(u) = 1 - u + u\log(u)$ and $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$ for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$.

Let $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ be a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density k defined on $\mathsf{T} \times \mathsf{Y}$. Moreover, for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $y \in \mathsf{Y}$, we denote $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y)$ and we consider in this chapter the general optimisation problem

$$\operatorname{arginf}_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; p) \ , \tag{2.1}$$

where we will drop the dependency on p for notational ease and when no ambiguity occurs. As explained in Chapter 1 (and derived in Appendix A.2), the optimisation problem (2.1) with $p = p(\cdot, \mathscr{D})$ amounts to performing α-divergence minimisation with respect to μ between the variational density $q = \mu k$ and the posterior density $p(\cdot, \mathscr{D})$, so that the Bayesian case is embedded in the general framework of (2.1); we may use the shorthand notation $\Psi_\alpha(q; \mathscr{D})$ instead of $\Psi_\alpha(q; p(\cdot, \mathscr{D}))$ to designate it.

The convexity of $\mu \mapsto \Psi_\alpha(\mu k)$ is straightforward from the convexity of f_α, therefore a simple yet powerful consequence of enlarging the variational family is that the optimisation problem now involves the *convex* mapping

$$\mu \mapsto \Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ ,$$

whereas the initial optimisation problem was associated to the mapping

$$\theta \mapsto \int_{\mathsf{Y}} f_\alpha \left(\frac{k(\theta, y)}{p(y)} \right) p(y) \nu(\mathrm{d}y) \ ,$$

which is not necessarily convex. We now detail the organisation of Chapter 2.

Outline The chapter is organised as follows:

- In Section 2.2, we describe the exact (α, Γ)-descent, an iterative algorithm that performs α-divergence minimisation by updating the measure μ. We establish in Theorem 1 sufficient conditions on Γ for this algorithm to lead at each step to a systematic decrease in the α-divergence. We then investigate the convergence of the algorithm in Theorem 2, 3 and 4.

Strikingly, the Infinite-dimensional Entropic Mirror Descent (Hsieh, Liu, and Cevher, 2019, Appendix A) is included in our framework and we obtain an $O(1/N)$ convergence rate under minimal assumptions, which improves on existing results and illustrates the generality of our approach. We also introduce a novel algorithm called the Power Descent, for which we prove convergence to an optimum and obtain an $O(1/N)$ convergence rate when $\alpha > 1$.

- In Section 2.3, we define the stochastic version of the exact (α, Γ)-descent and apply it to the important case of mixture models (Jaakkola and Jordan, 1998; Gershman, Hoffman, and Blei, 2012). The resulting general-purpose algorithm is Black-Box and does not require any information on the underlying distribution of the variational parameters. This algorithm notably enjoys an $O(1/\sqrt{N})$ convergence rate in the particular case of the Entropic Mirror Descent if we know the stopping time of the algorithm (Theorem 5).

- Finally, Section 2.4 is devoted to numerical experiments. We demonstrate the benefit of using the Power Descent and thus of going beyond the Entropic Mirror Descent framework. We also compare our method to a computationally equivalent Adaptive Importance Sampling algorithm for Bayesian Logistic Regression on a large dataset.

We thus start by describing the exact (α, Γ)-descent, a novel iterative algorithm to solve (2.1).

2.2 The (α, Γ)-descent

Throughout the chapter, we will assume the following conditions on k, p and ν.

(2.A1) The density kernel k on $\mathsf{T} \times \mathsf{Y}$, the function p on Y and the σ-finite measure ν on $(\mathsf{Y}, \mathcal{Y})$ satisfy, for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ and $\int_{\mathsf{Y}} p(y) \nu(\mathrm{d}y) < \infty$.

Under (2.A1), we immediately obtain a lower bound on Ψ_α.

Lemma 4. *Suppose that (2.A1) holds. Then, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, we have*

$$\Psi_\alpha(\mu k) \geqslant \hat{f}_\alpha\left(\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)\right) > -\infty\ ,$$

where $\hat{f}_\alpha$ is defined on $(0,\infty)$ by $\hat{f}_\alpha(u) = u f_\alpha(1/u)$.

Proof. Since $\hat{f}_\alpha(u) = u f_\alpha(1/u)$, we have

$$\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} \hat{f}_\alpha\left(\frac{p(y)}{\mu k(y)}\right) \mu k(y)\nu(\mathrm{d}y)\ .$$

Recalling that f_α and hence $\hat{f}_\alpha$, is convex on $\mathbb{R}_{>0}$, Jensen's inequality applied to $\hat{f}_\alpha$ yields $\Psi_\alpha(\mu k) \geqslant \hat{f}_\alpha\left(\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)\right) > -\infty$. □

Remark 5. *Assumption (2.A1) can be extended by discarding the assumption that $p(y)$ is positive for all $y \in \mathsf{Y}$. As it complicates the expression of the constant appearing in the bound without increasing dramatically the degree of generality of the results, we chose to maintain this assumption in Chapter 2 for the sake of simplicity.*

Thus, if there exists a sequence of probability measures $\{\mu_n\ :\ n \in \mathbb{N}^\star\}$ on $(\mathsf{T}, \mathcal{T})$ such that $\Psi_\alpha(\mu_1 k) < \infty$ and $\Psi_\alpha(\mu_n k)$ is non-increasing with n, Lemma 4 guarantees that this sequence converges to a limit in $\mathbb{R}$. We now focus on constructing such a sequence $\{\mu_n\ :\ n \in \mathbb{N}^\star\}$.

For this purpose, let $\mu \in \mathrm{M}_1(\mathsf{T})$. We introduce the one-step transition of the (α,Γ)-descent which can be described as an *expectation* step and an *iteration* step:

Algorithm 4: *Exact (α,Γ)-descent one-step transition*

1. Expectation step : $b_{\mu,\alpha}(\theta) = \displaystyle\int_{\mathsf{Y}} k(\theta,y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y)$
2. Iteration step : $\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \dfrac{\mu(\mathrm{d}\theta)\cdot\Gamma(b_{\mu,\alpha}(\theta)+\kappa)}{\mu(\Gamma(b_{\mu,\alpha}+\kappa))}$

Given a certain $\kappa \in \mathbb{R}$, a certain function Γ which takes its values in $\mathbb{R}_{>0}$ and an initial measure $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi_\alpha(\mu_1 k) < \infty$, the iterative sequence of probability measures $(\mu_n)_{n\in\mathbb{N}^\star}$ is then defined by setting

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n)\ , \qquad n \in \mathbb{N}^\star\ . \tag{2.2}$$

A first remark is that under (2.A1) and for all $\alpha \in \mathbb{R}\setminus\{1\}$, $b_{\mu,\alpha}$ is well-defined. As for the case $\alpha = 1$, we will assume in the rest of the chapter that $b_{\mu,1}(\theta)$ is finite for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and $\theta \in \mathsf{T}$. The iteration $\mu \mapsto \mathcal{I}_\alpha(\mu)$ is thus well-defined if moreover

we have

$$\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty \ . \tag{2.3}$$

A second remark is that we recover the Infinite-Dimensional Entropic Mirror Descent algorithm applied to the Kullback-Leibler (and more generally to the α-divergence) objective function by choosing Γ of the form

$$\Gamma(v) = e^{-\eta v} \ .$$

We refer to Hsieh, Liu, and Cevher, 2019, Appendix A for some theoretical background on the Infinite-Dimensional Entropic Mirror Descent. In this light, $b_{\mu,\alpha}(\theta)$ can be understood as the gradient of $\mu \mapsto \Psi_\alpha(\mu k)$. Algorithm 4 then consists in applying a transform function Γ to the translated gradient $b_{\mu,\alpha}(\theta) + \kappa$ and projecting back onto the space of probability measures.

In the rest of the section, we investigate some core properties of the aforementioned sequence of probability measures $(\mu_n)_{n\in\mathbb{N}^\star}$. We start by establishing conditions on (Γ, κ) such that the (α, Γ)-descent diminishes $\Psi_\alpha(\mu_n k)$ at each iteration for all $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ satisfying $\Psi_\alpha(\mu_1 k) < \infty$.

2.2.1 Monotonicity

To establish that the (α, Γ)-descent diminishes $\Psi_\alpha(\mu_n k)$ at each iteration, we first derive a general lower-bound for the difference $\Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$. Here, (ζ, μ) is a couple of probability measures where ζ is dominated by μ which we denote by $\zeta \preceq \mu$. This first result involves the following useful quantity

$$A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha \left(\frac{g(\theta)\mu k(y)}{p(y)} \right) [1 - g(\theta)] \ , \tag{2.4}$$

where g is the density of ζ w.r.t μ, i.e. $\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta)$.

Lemma 6. *Assume (2.A1). Then, for all $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ such that $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$, we have*

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k) \ . \tag{2.5}$$

Moreover, equality holds in (2.5) if and only if $\zeta = \mu$.

Proof. To prove (2.5), we introduce the intermediate function

$$h_\alpha(\zeta, \mu) = \int_{\mathsf{Y}} \nu(\mathrm{d}y) p(y) \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta) k(\theta, y)}{\mu k(y)} f_\alpha \left(\frac{g(\theta)\mu k(y)}{p(y)} \right) \ .$$

Then, the convexity of f_α combined with Jensen's inequality implies that

$$h_\alpha(\zeta, \mu) \geqslant \int_{\mathsf{Y}} \nu(\mathrm{d}y) p(y) f_\alpha \left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) g(\theta)}{p(y)} \right) = \Psi_\alpha(\zeta k) \ . \tag{2.6}$$

Next, set $u_{\theta,y} = \frac{g(\theta)\mu k(y)}{p(y)}$ and $v_y = \frac{\mu k(y)}{p(y)}$. Since the function f_α is convex, we have that for all $\theta \in \mathsf{T}$, for all $y \in \mathsf{Y}$, $f_\alpha(v_y) \geqslant f_\alpha(u_{\theta,y}) + f'_\alpha(u_{\theta,y})(v_y - u_{\theta,y})$, that is

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \,. \tag{2.7}$$

Now integrating over T with respect to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$ and then integrating over Y with respect to $p(y)\nu(\mathrm{d}y)$ in (2.7) yields

$$\Psi_\alpha(\mu k) \geqslant h_\alpha(\zeta, \mu) + A_\alpha \,. \tag{2.8}$$

Combining this result with (2.6) gives (2.5). The case of equality is obtained using the strict convexity of f_α in (2.6) and (2.7) which shows that g is constant μ-a.e. so that $\zeta = \mu$. □

We now plan on setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Lemma 6 and obtain that one iteration of the (α, Γ)-descent yields $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$. Based on the lower-bound obtained in Lemma 6, a sufficient condition is to prove that taking $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ in (2.4) implies $A_\alpha \geqslant 0$. For this purpose, let us denote by Dom_α an interval of $\mathbb{R}$ such that for all $\theta \in \mathsf{T}$, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $b_{\mu,\alpha}(\theta) + \kappa$ and $\mu(b_{\mu,\alpha}) + \kappa \in \mathrm{Dom}_\alpha$ and let us make an assumption on (Γ, κ).

(2.A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1]\,(\log \Gamma)'(v) + 1 \geqslant 0, \quad v \in \mathrm{Dom}_\alpha \,.$$

We now state our first main theorem.

Theorem 1. *Assume (2.A1) and (2.A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that (2.3) holds and $\Psi_\alpha(\mu k) < \infty$. Then, the two following assertions hold.*

(i) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$.

(ii) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Proof. To prove (i), we set $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ in (2.4) and we will show that $A_\alpha \geqslant 0$. Then, the proof is concluded by setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Lemma 6 as

$$\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k) - A_\alpha \leqslant \Psi_\alpha(\mu k) \,. \tag{2.9}$$

We study the cases $\alpha = 1$ and $\alpha \in \mathbb{R} \setminus \{1\}$ separately.

(a) Case $\alpha = 1$. In this case $f_1'(u) = \log u$ and we have

$$\begin{aligned} A_1 &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \log\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)] \\ &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \left[\log g(\theta) + f_1'\left(\frac{\mu k(y)}{p(y)}\right)\right] [1 - g(\theta)] \\ &= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[\log g(\theta) + \int_{\mathsf{Y}} k(\theta, y) f_1'\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)\right] [1 - g(\theta)] \\ &= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[\log g(\theta) + b_{\mu,1}(\theta) + \kappa\right] [1 - g(\theta)] \ . \end{aligned}$$

where we used that $\mu[\kappa(1-g)] = 0$ in the last equality. Setting $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,1} + \kappa))$ for all $v \in \mathrm{Dom}_1$, we have $g = \tilde{\Gamma} \circ (b_{\mu,1} + \kappa)$. Let us thus consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and let V be the random variable $V(\theta) = b_{\mu,1}(\theta) + \kappa$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and we can write

$$A_1 = \mathbb{E}[(\log \tilde{\Gamma}(V) + V)(1 - \tilde{\Gamma}(V))] = \mathbb{C}\mathrm{ov}(\log \tilde{\Gamma}(V) + V, 1 - \tilde{\Gamma}(V)) \ .$$

Under (2.A2) with $\alpha = 1$, $v \mapsto \log \tilde{\Gamma}(v) + v$ and $v \mapsto 1 - \tilde{\Gamma}(v)$ are increasing on Dom_1 which implies $A_1 \geqslant 0$.

(b) Case $\alpha \in \mathbb{R} \setminus \{1\}$. In this case $f_\alpha'(u) = \frac{1}{\alpha - 1}[u^{\alpha-1} - 1]$ and we have

$$\begin{aligned} A_\alpha &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)] \\ &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} g(\theta)^{\alpha-1} [1 - g(\theta)] \\ &= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1}\right] g(\theta)^{\alpha-1} [1 - g(\theta)] \ . \end{aligned}$$

Again, setting $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$, we have $g = \tilde{\Gamma} \circ (b_{\mu,\alpha} + \kappa)$. Let us consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Then, we have $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and setting $\kappa' = \kappa - \frac{1}{\alpha - 1}$ we can write

$$A_\alpha = \mathbb{E}[(V - \kappa')\tilde{\Gamma}^{\alpha-1}(V)(1 - \tilde{\Gamma}(V))] = \mathbb{C}\mathrm{ov}((V - \kappa')\tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V)) \ .$$

Under (2.A2) with $\alpha \in \mathbb{R} \setminus \{1\}$, $v \mapsto (v - \kappa')\tilde{\Gamma}^{\alpha-1}(v)$ and $v \mapsto 1 - \tilde{\Gamma}(v)$ are increasing on Dom_α which implies $A_\alpha \geqslant 0$.

Let us now show (ii). The *if* part is obvious. As for the *only if* part, $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ combined with (2.9) yields

$$\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k) - A_\alpha \ ,$$

which is the case of equality in Lemma 6. Therefore, $\mathcal{I}_\alpha(\mu) = \mu$. □

Possible choices for (Γ, κ). At this stage, we have established conditions on (Γ, κ) such that $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$ and identified the case of equality. Notice in particular that the inequality in (2.A2) is free from the parameter κ when $\alpha = 1$, which implies that the function $\Gamma(v) = e^{-\eta v}$ satisfies (2.A2) for all $\eta \in (0, 1]$. As a consequence, the case of the Entropic Mirror Descent with the forward Kullback-Leibler divergence as objective function is included in this framework.

One can also readily check that $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ satisfies (2.A2) for all $\alpha \in \mathbb{R} \setminus \{1\}$, for all κ such that $(\alpha - 1)\kappa \geqslant 0$ and for all $\eta \in (0, 1]$. We will refer to this particular choice of Γ as the *Power Descent* thereafter. These two examples are summarised in Table 2.1 below.

TABLE 2.1: Examples of allowed (Γ, κ) in the (α, Γ)-descent according to Theorem 1.

Divergence considered	Possible choices for (Γ, κ)	
Forward KL ($\alpha = 1$)	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, 1]$	any κ
α-divergence with $\alpha \in \mathbb{R} \setminus \{1\}$	$\Gamma(v) = [(\alpha - 1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$(\alpha - 1)\kappa \geqslant 0$

Improving upon Lemma 6. In the following lemma, we derive an explicit lower-bound for $\Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu)k)$ in terms of the variance of $b_{\mu,\alpha}$. Let us thus consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and denote by $\mathbb{V}\mathrm{ar}_\mu$ the associated variance operator.

Lemma 7. *Assume (2.A1) and (2.A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that (2.3) holds and $\Psi_\alpha(\mu k) < \infty$. Then,*

$$\frac{L_{\alpha,1}}{2} \mathbb{V}\mathrm{ar}_\mu \left(b_{\mu,\alpha}\right) \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \,, \tag{2.10}$$

where

$$L_{\alpha,1} := \inf_{v \in \mathrm{Dom}_\alpha} \left\{ [(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \right\} \times \inf_{v \in \mathrm{Dom}_\alpha} -\Gamma'(v) \,.$$

Proof. On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, consider the random variable $U(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ and let V be an independent copy of U. For all $u \in \mathrm{Dom}_\alpha$, define $\tilde{\Gamma}(u) = \Gamma(u)/\mathbb{E}[\Gamma]$. Let us now prove that

$$A_\alpha \geqslant \frac{L_{\alpha,1}}{2} \mathbb{V}\mathrm{ar}_\mu(b_{\mu,\alpha}) \,.$$

We study the cases $\alpha = 1$ and $\alpha \in \mathbb{R} \setminus \{1\}$ separately.

(a) Case $\alpha = 1$. In this case,

$$A_1 = \mathbb{C}\mathrm{ov}(\log \tilde{\Gamma}(U) + U, 1 - \tilde{\Gamma}(U)) \,.$$

Using that $\mathbb{E}[1 - \tilde{\Gamma}] = 0$, we can rewrite A_1 under the form

$$\begin{aligned} A_1 &= \frac{1}{2}\mathbb{E}\left[(\log\tilde{\Gamma}(U) + U - \log\tilde{\Gamma}(V) + V)(-\tilde{\Gamma}(U) + \tilde{\Gamma}(V))\right] \\ &= \frac{1}{2}\mathbb{E}\left[\frac{\log\tilde{\Gamma}(U) + U - (\log\tilde{\Gamma}(V) + V)}{U - V}\frac{-\tilde{\Gamma}(U) + \tilde{\Gamma}(V)}{U - V}(U - V)^2\right] \\ &\geqslant \frac{L_{1,1}}{2}\mathbb{V}\mathrm{ar}_\mu(b_{\mu,1}) \ . \end{aligned}$$

(b) Case $\alpha \in \mathbb{R} \setminus \{1\}$. Set $\kappa' = \kappa - \frac{1}{\alpha - 1}$. In this case,

$$A_\alpha = \mathbb{C}\mathrm{ov}((U - \kappa')\tilde{\Gamma}^{\alpha-1}(U), 1 - \tilde{\Gamma}(U)) \ ,$$

which, using once again that $\mathbb{E}[1 - \tilde{\Gamma}] = 0$, can be rewritten as

$$\begin{aligned} A_\alpha &= \frac{1}{2}\mathbb{E}\left[((U - \kappa')\tilde{\Gamma}^{\alpha-1}(U) - (V - \kappa')\tilde{\Gamma}^{\alpha-1}(V))(-\tilde{\Gamma}(U) + \tilde{\Gamma}(V))\right] \\ &= \frac{1}{2}\mathbb{E}\left[\frac{(U - \kappa')\tilde{\Gamma}^{\alpha-1}(U) - (V - \kappa')\tilde{\Gamma}^{\alpha-1}(V)}{U - V}\frac{-\tilde{\Gamma}(U) + \tilde{\Gamma}(V)}{U - V}(U - V)^2\right] \\ &\geqslant \frac{L_{\alpha,1}}{2}\mathbb{V}\mathrm{ar}_\mu(b_{\mu,\alpha}) \ . \end{aligned}$$

Combining with (2.5) yields (2.10).

□

Lemma 7 can be interpreted in the following way: provided that $L_{\alpha,1} > 0$, (2.10) states that the case of equality is reached if and only if the variance of the gradient $b_{\mu,\alpha}$ equals zero. Such a result, which holds for any transform function Γ satisfying (2.A2), quantifies the improvement after one step of the (α, Γ)-descent.

Interestingly, monotonicity properties akin to Lemma 7 have previously been derived under stronger smoothness assumptions in the context of Projected Gradient Descent steps. For example, in the particular case where the objective function f is assumed to be β-smooth on $\mathbb{R}$, for all $u \in \mathbb{R}$ it holds (see for example Bubeck, 2015, Equation 3.5) that

$$\frac{1}{2\beta}\|\nabla f(u)\|^2 \leqslant f(u) - f\left(u - \frac{1}{\beta}\nabla f(u)\right) \ .$$

This result is then used to obtain improved convergence rates for the Projected Gradient Descent algorithm. Consequently, we are next interested in proving a rate of convergence for the exact (α, Γ)-descent by leveraging Lemma 7 .

2.2.2 Convergence

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$. We want to study the limiting behavior of the exact (α, Γ)-descent for the iterative sequence of probability measure $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (2.2). To do so,

we first introduce the two following useful quantities

$$L_{\alpha,2}^{-1} := \inf_{v\in \mathrm{Dom}_\alpha} (-\log\Gamma)'(v) \quad \text{and} \quad L_{\alpha,3}^{-1} := \inf_{v\in \mathrm{Dom}_\alpha} \Gamma(v)\ .$$

We define $\mathrm{M}_{1,\mu_1}(\mathsf{T})$ as the set of probability measures dominated by μ_1. Next, we strengthen the assumptions on Γ as follows.

(2.A3) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is L-smooth and the function $-\log\Gamma$ is concave increasing.

We are now able to derive our second main result.

Theorem 2. *Assume (2.A1), (2.A2) and (2.A3). Further assume that $L_{\alpha,1}$, $L_{\alpha,2} > 0$ and that $0 < \inf_{v\in\mathrm{Dom}_\alpha}\Gamma(v) \leqslant \sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v) < \infty$. Moreover, let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. Then, the following assertions hold.*

(i) The sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (2.2) is well-defined and we have that the sequence $(\Psi_\alpha(\mu_n k))_{n\in\mathbb{N}^\star}$ is non-increasing.

(ii) For all $N \in \mathbb{N}^\star$, we have

$$\Psi_\alpha(\mu_N k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right]\ , \tag{2.11}$$

where $\mu^\star$ is such that $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})}\Psi_\alpha(\zeta k)$ and where we have defined $\Delta_1 = \Psi_\alpha(\mu_1 k) - \Psi_\alpha(\mu^\star k)$ and $KL(\mu^\star||\mu_1) = \int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu^\star}{\mathrm{d}\mu_1}\right)\mathrm{d}\mu^\star$.

Proof of Theorem 2. We prove the assertions successively.

(i) The proof of (i) simply consists in verifying that we can apply Theorem 1. For all $\mu \in \mathrm{M}_1(\mathsf{T})$, (2.3) holds as we have

$$\mu(\Gamma(b_{\mu,\alpha}+\kappa)) \leqslant \mu\left(\sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v)\right) < \infty,$$

and since at each step $n \in \mathbb{N}^\star$, Theorem 1 combined with $\Psi_\alpha(\mu_n k) < \infty$ implies that $\Psi_\alpha(\mu_{n+1}k) \leqslant \Psi_\alpha(\mu_n k) < \infty$, we obtain by induction that $(\Psi_\alpha(\mu_n k))_{n\in\mathbb{N}^\star}$ is non-increasing.

(ii) For the sake of readability, we only treat the case $\kappa = 0$ in the proof of (ii). Note that the case $\kappa \neq 0$ unfolds similarly by replacing $b_{\mu,\alpha}$ by $b_{\mu,\alpha}+\kappa$ everywhere in the proof below. Let $n \in \mathbb{N}^\star$ and set $\Delta_n = \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu^\star k)$. We first show that

$$\Delta_n \leqslant L_{\alpha,2}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\mathrm{d}\mu^\star + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}(b_{\mu_n,\alpha})L_{\alpha,3}\right]\ . \tag{2.12}$$

The convexity of f_α implies that

$$\Delta_n \leqslant \int_{\mathsf{T}} b_{\mu_n,\alpha}(\mathrm{d}\mu_n - \mathrm{d}\mu^\star) = \int_{\mathsf{T}} (\mu_n(b_{\mu_n,\alpha}) - b_{\mu_n,\alpha})\mathrm{d}\mu^\star\ .$$

In addition, the concavity of $-\log\Gamma$ implies that for all $u, v \in \mathrm{Dom}_\alpha$,

$$-\log\Gamma(u) \leqslant -\log\Gamma(v) + (-\log\Gamma)'(v)(u-v) \;,$$

i.e

$$(-\log\Gamma)'(v)(v-u) \leqslant \log\Gamma(u) - \log\Gamma(v) \;.$$

Since by assumption $-\log\Gamma$ is increasing, $(-\log\Gamma)'(v) > 0$ and we deduce

$$v - u \leqslant \frac{\log\Gamma(u) - \log\Gamma(v)}{(-\log\Gamma)'(v)} \;. \tag{2.13}$$

We can apply (2.13) with $u = b_{\mu_n,\alpha}(\theta)$ and $v = \mu_n(b_{\mu_n,\alpha})$ which yields

$$\mu_n(b_{\mu_n,\alpha}) - b_{\mu_n,\alpha}(\theta) \leqslant \frac{\log\Gamma(b_{\mu_n,\alpha}(\theta)) - \log\Gamma(\mu_n(b_{\mu_n,\alpha}))}{(-\log\Gamma)'(\mu_n(b_{\mu_n,\alpha}))} \;.$$

Now integrating with respect to $\mathrm{d}\mu^\star$, we obtain

$$\Delta_n \leqslant \frac{1}{(-\log\Gamma)'(\mu_n(b_{\mu_n,\alpha}))} \int_{\mathsf{T}} \left[\log\Gamma(b_{\mu_n,\alpha}) - \log\Gamma(\mu_n(b_{\mu_n,\alpha}))\right] \mathrm{d}\mu^\star \;.$$

By definition of $\mu^\star$, we have that $\Delta_n \geqslant 0$ and combining with the fact that $(-\log\Gamma)'(\mu_n(b_{\mu_n,\alpha})) > 0$, we can deduce

$$\int_{\mathsf{T}} \left[\log\Gamma(b_{\mu_n,\alpha}) - \log\Gamma(\mu_n(b_{\mu_n,\alpha}))\right] \mathrm{d}\mu^\star \geqslant 0 \;.$$

Consequently, we obtain

$$\begin{aligned}
\Delta_n &\leqslant L_{\alpha,2} \int_{\mathsf{T}} \left[\log\Gamma(b_{\mu_n,\alpha}) - \log\Gamma(\mu_n(b_{\mu_n,\alpha}))\right] \mathrm{d}\mu^\star \\
&= L_{\alpha,2} \int_{\mathsf{T}} \left[\log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right) + \log\mu_n(\Gamma(b_{\mu_n,\alpha})) - \log\Gamma(\mu_n(b_{\mu_n,\alpha}))\right] \mathrm{d}\mu^\star \\
&= L_{\alpha,2} \left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right) \mathrm{d}\mu^\star + \log\mu_n(\Gamma(b_{\mu_n,\alpha})) - \log\Gamma(\mu_n(b_{\mu_n,\alpha}))\right] \;.
\end{aligned} \tag{2.14}$$

Next, we show that

$$\log\mu_n(\Gamma(b_{\mu_n,\alpha})) - \log\Gamma(\mu_n(b_{\mu_n,\alpha})) \leqslant \frac{L}{2} \mathbb{V}\mathrm{ar}_{\mu_n}(b_{\mu_n,\alpha}) L_{\alpha,3} \;.$$

By assumption Γ is L-smooth on Dom_α, thus for all $\theta \in \mathsf{T}$ and for all $n \in \mathbb{N}^\star$,

$$\begin{aligned}
\Gamma(b_{\mu_n,\alpha}(\theta)) \leqslant \Gamma(\mu_n(b_{\mu_n,\alpha})) + \Gamma'(\mu_n(b_{\mu_n,\alpha}))(b_{\mu_n,\alpha}(\theta) - \mu_n(b_{\mu_n,\alpha})) \\
+ \frac{L}{2} \left(b_{\mu_n,\alpha}(\theta) - \mu_n(b_{\mu_n,\alpha})\right)^2
\end{aligned}$$

which in turn implies

$$\mu_n(\Gamma(b_{\mu_n,\alpha})) \leqslant \Gamma(\mu_n(b_{\mu_n,\alpha})) + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}\left(b_{\mu_n,\alpha}\right) \ .$$

Finally, we obtain

$$\log \mu_n(\Gamma(b_{\mu_n,\alpha})) \leqslant \log \Gamma(\mu_n(b_{\mu_n,\alpha})) + \log\left(1 + \frac{L}{2}\frac{\mathbb{V}\mathrm{ar}_{\mu_n}(b_{\mu_n,\alpha})}{\Gamma(\mu_n(b_{\mu_n,\alpha}))}\right) \ .$$

Using that $\log(1+u) \leqslant u$ when $u \geqslant 0$ and that $1/\Gamma$ is increasing, we deduce

$$\log \mu_n(\Gamma(b_{\mu_n,\alpha})) \leqslant \log \Gamma(\mu_n(b_{\mu_n,\alpha})) + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}\left(b_{\mu_n,\alpha}\right) L_{\alpha,3} \ ,$$

which combined with (2.14) implies (2.12). To conclude, we apply Lemma 7 to $g = \frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}$ and combining with (2.12), we obtain

$$\Delta_n \leqslant L_{\alpha,2}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\mathrm{d}\mu^\star + \frac{LL_{\alpha,3}}{L_{\alpha,1}}\left(\Delta_n - \Delta_{n+1}\right)\right] \ ,$$

where by assumption $L_{\alpha,1}$, $L_{\alpha,2}$ and $L_{\alpha,3} > 0$. As the r.h.s involves two telescopic sums, we deduce

$$\frac{1}{N}\sum_{n=1}^{N}\Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{L_{\alpha,2}}{N}\Bigg[KL(\mu^\star||\mu_1) - KL(\mu^\star||\mu_{N+1}) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}(\Delta_1 - \Delta_{N+1})\Bigg] \quad (2.15)$$

and we recover (2.11) using (i), that $KL(\mu^\star||\mu_{N+1}) \geqslant 0$ and that $\Delta_{N+1} \geqslant 0$.

□

Remark 8. *Note that the convexity of the mapping $\mu \mapsto \Psi_\alpha(\mu k)$ in (2.15) implies an $O(1/N)$ convergence rate for $\bar{\mu}_N = \frac{1}{N}\sum_{n=1}^{N}\mu_n$ as well:*

$$\Psi_\alpha\left(\bar{\mu}_N k\right) - \Psi_\alpha(\mu^\star k) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right] \ .$$

We now wish to comment on the constants appearing in (2.11) and in particular the two constants $KL(\mu^\star||\mu_1)$ and Δ_1 (since the remaining constants $L_{\alpha,1}$, $L_{\alpha,2}$, $L_{\alpha,3}$ and L all involve the function Γ, which has not been chosen yet in Theorem 2).

To do so, we consider in Example 3 the finite-dimensional case where μ_1 is a weighted sum of dirac measures. As we shall explain in more details later on in Section 2.3, this case is of particular relevance to us as our procedure can then be used to optimise the mixture weights of any given mixture model.

Example 3 (Simplex Framework). *Let $J \in \mathbb{N}^\star$, let $(\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$ and let us consider $\mu_1 = J^{-1}\sum_{j=1}^{J}\delta_{\theta_j}$. Then, $\mu^\star$ is of the form $\sum_{j=1}^{J}\lambda_j^\star\delta_{\theta_j}$ where $(\lambda_1^\star, ..., \lambda_J^\star)$ belongs to*

the simplex of dimension J. Moreover, the two quantities $KL(\mu^\star||\mu_1)$ and Δ_1 can easily be bounded in terms of J. Indeed, using that $\log u \leqslant u-1$ for all $u > 0$ and that $\sum_{j=1}^J \lambda_j^{\star 2} \leqslant 1$, we obtain that

$$\begin{aligned} KL(\mu^\star||\mu_1) &= \sum_{j=1}^J \lambda_j^\star \log \lambda_j^\star + \log J \\ &\leqslant \log J \ . \end{aligned}$$

As for Δ_1, we have by convexity that

$$\Delta_1 \leqslant [\mu_1 - \mu^\star](b_{\mu_1,\alpha})$$

and, using Pinsker's inequality as well as the bound on $KL(\mu^\star||\mu_1)$ we have established just above, we can deduce

$$\begin{aligned} \Delta_1 &\leqslant [\mu_1 - \mu^\star](b_{\mu_1,\alpha} - \mathbb{E}_{\mu_1}[b_{\mu_1,\alpha}]) \\ &\leqslant \sqrt{2}\sqrt{KL(\mu^\star||\mu_1)} \max_{1\leqslant j,j'\leqslant J} |b_{\mu_1,\alpha}(\theta_j) - b_{\mu_1,\alpha}(\theta_{j'})| \\ &\leqslant \sqrt{2\log J} \max_{1\leqslant j,j'\leqslant J} |b_{\mu_1,\alpha}(\theta_j) - b_{\mu_1,\alpha}(\theta_{j'})| \ . \end{aligned}$$

In the next Theorem, we state several practical examples of couples (Γ, κ) which satisfy the assumptions from Theorem 2.

Theorem 3. *Assume (2.A1). Define $|b|_{\infty,\alpha} := \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)|$ and assume that $|b|_{\infty,\alpha} < \infty$. Let (Γ, κ) belong to any of the following cases.*

(i) Forward Kullback-Leibler divergence ($\alpha = 1$): $\Gamma(v) = e^{-\eta v}$, $\eta \in (0, 1)$ and κ is any real number (Entropic Mirror Descent);

(ii) Reverse Kullback-Leibler ($\alpha = 0$) and α-Divergence with $\alpha \in \mathbb{R} \setminus \{0, 1\}$:

(a) $\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$ and κ is any real number (Entropic Mirror Descent);

(b) $\Gamma(v) = [(\alpha-1)v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$, $\alpha > 1$ and $\kappa > 0$ (Power Descent);

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. Then, the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (2.2) is well-defined and the sequence $(\Psi_\alpha(\mu_n k))_{n\in\mathbb{N}^\star}$ is non-increasing with a convergence rate characterized by (2.11).

The proof of Theorem 3 can be found in Section 2.A.1. In terms of assumptions, we only require the gradients of the function $\mu \mapsto \Psi_\alpha(\mu k)$ to be bounded in l_∞-norm, which is a standard assumption, and the objective function to be finite at the starting measure μ_1, i.e. $\Psi_\alpha(\mu_1 k) < \infty$, which again is a mild assumption that can even be discarded for all $\alpha \neq 0$, as written in Remark 9 below.

Remark 9 (Assumption $\Psi_\alpha(\mu_1 k) < \infty$ in Theorem 3)**.** *The assumption $\Psi_\alpha(\mu_1 k) < \infty$ can be discarded in Theorem 3 for all $\alpha \neq 0$. Indeed, for all $\alpha \in \mathbb{R}$ and for all $u > 0$, we have that $u f'_\alpha(u) = \alpha f_\alpha(u) + u - 1$ and thus we can write*

$$\mu_1(b_{\mu_1,\alpha}) = \alpha \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu_1 k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) - \int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) + 1 \,.$$

Under (2.A1), it holds that $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$. Combined with the fact that we have assumed that $|b|_{\infty,\alpha} < \infty$ in Theorem 3, we obtain that $\Psi_\alpha(\mu_1 k) < \infty$.

Let us now illustrate the benefits of our approach with an example where the different constants appearing in (2.11) are bounded explicitly and where we compare the convergence rate we obtain with typical Mirror Descent convergence results from the optimisation literature.

Example 4 (Simplex framework and forward Kullback-Leibler)**.** *Let $J \in \mathbb{N}^\star$, let $(\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$ and let us consider $\mu_1 = J^{-1}\sum_{j=1}^J \delta_{\theta_j}$. In addition, let $\alpha = 1$ and $\Gamma(v) = e^{-\eta v}$ with $v \in \mathrm{Dom}_\alpha = [-|b|_{\infty,1} + \kappa, |b|_{\infty,1} + \kappa]$ and $\kappa \in \mathbb{R}$. Then, we have $L_{1,1} = (1-\eta)\eta e^{-\eta|b|_{\infty,\alpha} - \eta\kappa}$, $L_{1,2} = \eta^{-1}$, $L_{1,3} = e^{\eta|b|_{\infty,\alpha} + \eta\kappa}$ and $L = \eta^2 e^{\eta|b|_{\infty,\alpha} - \eta\kappa}$.*

In the particular case of the Entropic Mirror Descent, the constant κ does not appear in the update formula (2.2) due to the normalisation, so we can choose it however we want without impacting the convergence of the algorithm. Notice then that by choosing $\kappa = -3|b|_{\infty,\alpha}$ and based on Example 3, we obtain the following convergence rate for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_N k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,\alpha}}{(1-\eta)N} \,.$$

Thus, in the particular case of Example 4, the dominant term in (2.11) with respect to the dimension J of the simplex is in $\log J$ so that we achieve an overall $O(\frac{\log J}{N})$ convergence rate. Furthermore, the range of possible values for η is stated explicitly, since the result holds for all $\eta \in (0,1)$.

This is an improvement compared to standard Mirror Descent results, which under similar assumptions only provide an $O(1/\sqrt{N})$ convergence rate and assume an $O(1/\sqrt{N})$ learning rate (see Beck and Teboulle, 2003 or Bubeck, 2015, Theorem 4.2.). Indeed, Projected Gradient Descent and Entropic Mirror Descent typically achieve an $O(\sqrt{J/N})$ and $O(\sqrt{\log(J)/N})$ convergence rate respectively in the Simplex framework. This means that Theorem 3 improves with respect to both N and J compared to Projected Gradient Descent and that it improves with respect to N for the Entropic Mirror Descent with a small cost in terms of the dimension J of the simplex.

Moreover, while accelerated versions of the Mirror Descent (e.g. Mirror Prox, see Nemirovski, 2004 or Bubeck, 2015, Theorem 4.4.) also yield an $O(1/N)$ convergence

rate, they require the objective function to be sufficiently smooth, an additional assumption that we have bypassed when deriving our results.

The case of the Power Descent for $\alpha < 1$ is not included in Theorem 3. This case is trickier and must be handled separately in order to obtain the convergence of the algorithm. For this purpose, we first introduce the following additive set of assumptions

(2.A4)

(i) T is a compact metric space and $\mathcal{T}$ is the associated Borel σ-field;

(ii) for all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;

(iii) we have $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} k(\theta, y) \times \sup_{\theta' \in \mathsf{T}} \left(\frac{k(\theta', y)}{p(y)} \right)^{\alpha - 1} \nu(\mathrm{d}y) < \infty$.

If $\alpha = 0$, assume in addition that $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} \left| \log \left(\frac{k(\theta, y)}{p(y)} \right) \right| p(y) \nu(\mathrm{d}y) < \infty$.

Here, the condition (2.A4)-(iii) implies that $b_{\mu,\alpha}(\theta)$ and $\Psi_\alpha(\mu k)$ are uniformly bounded with respect to μ and θ, which is rather weak condition under (2.A4)-(i) since we consider a supremum taken over a compact set (and T will always be chosen as such in practice). We then have the following theorem, which states that the possible weak limits of $(\mu_n)_{n \in \mathbb{N}^\star}$ correspond to the global infimum of $\mu \mapsto \Psi_\alpha(\mu k)$.

Theorem 4. *Assume (2.A1) and (2.A4). Let $\alpha < 1$, $\kappa \leqslant 0$ and set $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ for all $v \in \mathrm{Dom}_\alpha$. Then, for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, any $\eta > 0$ satisfies (2.3) and $\Psi_\alpha(\zeta k) < \infty$.*

Let $\eta \in (0, 1]$. Further assume that there exist $\mu_1, \mu^\star \in \mathrm{M}_1(\mathsf{T})$ such that the (well-defined) sequence $(\mu_n)_{n \in \mathbb{N}^\star}$ defined by (2.2) weakly converges to $\mu^\star$ as $n \to \infty$. Then the following assertions hold

(i) $(\Psi_\alpha(\mu_n k))_{n \in \mathbb{N}^\star}$ *is non-increasing,*

(ii) $\mu^\star$ *is a fixed point of* $\mathcal{I}_\alpha$*,*

(iii) $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$.

The proof of Theorem 4 is deferred to Section 2.A.2. Intuitively, we expect $\mu^\star$ to be a fixed point of $\mathcal{I}_\alpha$ based on Theorem 1. The core difficulty of the proof is then to prove Assertion (iii) and to do so, we proceed by contradiction: we assume there exists $\bar{\mu} \in \mathrm{M}_{1,\mu_1}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star k) > \Psi_\alpha(\bar{\mu} k)$ and we contradict the fact that $(\mu_n)_{n \in \mathbb{N}^\star}$ converges to a fixed point.

The impact of Theorem 3 and Theorem 4 is twofold: not only our results improve on the $O(1/\sqrt{N})$ convergence rates previously established for Mirror Descent algorithms but they also allow us to go beyond the typical Entropic Mirror Descent framework by introducing the Power Descent.

Another interesting aspect is that the range of allowed values for the learning rate η is given explicitly in some cases (namely, the Power Descent and the Entropic Mirror Descent with the forward Kullback-Leibler). This is in contrast with usual Mirror Descent convergence results where the optimal learning rate depends on $|b|_{\infty,\alpha}$, the Lipschitz constant of Ψ_α, which might be unknown in practice.

The results we obtained thus far are summarized in Table 2.2 below.

TABLE 2.2: Examples of allowed (Γ, κ) in the (α, Γ)-descent according to Theorem 3 and Theorem 4.

Divergence considered	Possible choice of (Γ, κ)	
Forward KL ($\alpha = 1$)	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, 1)$	any κ
α-*divergence* with $\alpha \in \mathbb{R} \setminus \{1\}$	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{\lvert\alpha-1\rvert \lvert b\rvert_{\infty,\alpha}+1})$	any κ
	$\alpha > 1, \Gamma(v) = [(\alpha-1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa > 0$
	$\alpha < 1, \Gamma(v) = [(\alpha-1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa \leqslant 0$

As Algorithm 4 typically involves an intractable integral in the Expectation step, we now turn to a stochastic version of this algorithm.

2.3 Stochastic (α, Γ)-descent

We start by introducing the notation for the stochastic version of Algorithm 4. Let $M \in \mathbb{N}^\star$ and let $\mu \in \mathrm{M}_1(\mathsf{T})$. The Stochastic (α, Γ)-descent algorithm one-step transition is defined as follows.

Algorithm 5: *Stochastic (α, Γ)-descent one-step transition*

1. Sampling step : Draw independently $Y_1, \ldots, Y_M \sim \mu k$
2. Expectation step : $\hat{b}_{\mu,\alpha,M}(\theta) = \dfrac{1}{M} \displaystyle\sum_{m=1}^{M} \dfrac{k(\theta, Y_m)}{\mu k(Y_m)} f'_\alpha \left(\dfrac{\mu k(Y_m)}{p(Y_m)} \right)$
3. Iteration step : $\hat{\mathcal{I}}_{\alpha,M}(\mu)(\mathrm{d}\theta) = \dfrac{\mu(\mathrm{d}\theta) \cdot \Gamma(\hat{b}_{\mu,\alpha,M}(\theta) + \kappa)}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M} + \kappa))}$

Let us now denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space and by $\mathbb{E}$ the associated expectation operator. Given $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$, the stochastic version of the exact iterative scheme defined by (2.2) is then given by

$$\hat{\mu}_{n+1} = \hat{\mathcal{I}}_{\alpha,M}(\hat{\mu}_n) \,, \qquad n \in \mathbb{N}^\star \,, \tag{2.16}$$

where we have defined for all $\theta \in \mathsf{T}$ and for all $n \geqslant 1$,

$$\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta, Y_{m,n+1})}{\hat{\mu}_n k(Y_{m,n+1})} f'_\alpha\left(\frac{\hat{\mu}_n k(Y_{m,n+1})}{p(Y_{m,n+1})}\right) \tag{2.17}$$

with $Y_{1,n+1}, \dots, Y_{M,n+1} \overset{\text{i.i.d}}{\sim} \hat{\mu}_n k$ conditionally on $\mathcal{F}_n$ and where $\mathcal{F}_1 = \emptyset$ and $\mathcal{F}_n = \sigma(Y_{1,2}, \dots, Y_{M,2}, \dots, Y_{1,n}, \dots, Y_{M,n})$ for $n \geqslant 2$. Notice that we use $\hat{\mu}_n k$ as a sampler instead of $k(\theta, \cdot)$ in (2.17). As our algorithm optimises over μ, sampling with respect to $\hat{\mu}_n k$ is not only cheaper computationally, but it also gives preference to the interesting regions of the parameter space.

A first idea to study this algorithm is to adapt Theorem 2 to the stochastic case. This can be done for the Entropic Mirror Descent algorithm and in that case a bound on $\mathbb{E}[\Psi_\alpha(N^{-1}\sum_{n=1}^{N}\hat{\mu}_n k) - \Psi_\alpha(\mu^\star k)]$ of the form $O(1/N) + O(1/\sqrt{M})$ can be derived for a wide range of constant learning rates η (see Section 2.A.3 for the formal statement of the result and its proof). Maintaining an $O(1/N)$ bound however requires $M \geqslant N^2$, which yields an overall computational cost of order N^3. Another option consists in adapting Nemirovski et al., 2009 to our framework. This option involves a learning rate policy $(\eta_n)_{n\in\mathbb{N}}$ and notably yields an $O(1/\sqrt{N})$ bound for a constant policy $\eta_n = \eta_0/\sqrt{N}$, as written in Theorem 5 below.

Theorem 5. *Assume (2.A1). Let $M \in \mathbb{N}^\star$ and let $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$. Given a sequence of positive learning rates $(\eta_n)_{n\in\mathbb{N}}$, we let $(\hat{\mu}_n)_{n\in\mathbb{N}^\star}$ be defined by $\frac{\mathrm{d}\hat{\mu}_{n+1}}{\mathrm{d}\hat{\mu}_n} \propto e^{-\eta_n \hat{b}_{\hat{\mu}_n,\alpha,M}}$ and we set $w_n = \frac{\eta_n}{\sum_{n=1}^{N}\eta_n}$, $n \geqslant 1$. Further assume that*

$$B_\alpha := \left(\sup_{\mu\in\mathrm{M}_1(\mathsf{T})}\int_{\mathsf{Y}}\sup_{\theta,\theta'\in\mathsf{T}}\frac{k(\theta,y)^2}{k(\theta',y)}\left|f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\right|^2\nu(\mathrm{d}y)\right)^{1/2} < \infty\,, \tag{2.18}$$

and define $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\hat{\mu}_1}(\mathsf{T})}\Psi_\alpha(\zeta k)$. Then, for any $N \in \mathbb{N}^\star$,

$$\mathbb{E}\left[\Psi_\alpha\left(\sum_{n=1}^{N} w_n\hat{\mu}_n k\right) - \Psi_\alpha(\mu^\star k)\right] \leqslant \frac{B_\alpha^2\sum_{n=1}^{N}\eta_n^2/2}{\sum_{n=1}^{N}\eta_n} + \frac{KL(\mu^\star||\hat{\mu}_1)}{\sum_{n=1}^{N}\eta_n}\,, \tag{2.19}$$

In particular, the decreasing policy $\eta_n = \eta_0/\sqrt{n}$ yields an $O(\log(N)/\sqrt{N})$ bound in (2.19). Furthermore, the constant policy $\eta_n = \eta_0/\sqrt{N}$ yields an $O(1/\sqrt{N})$ bound in (2.19), which is minimal for $\eta_0 = B_\alpha^{-1}\sqrt{2KL(\mu^\star||\hat{\mu}_1)}$.

The proof of Theorem 5 can be found in Section 2.A.3 and we give below an example satisfying condition (2.18).

Example 5. *Consider the case $\mathsf{Y} = \mathbb{R}^d$ and $\alpha = 1$. Let $r > 0$ and let $\mathsf{T} = \mathcal{B}(0, r) \subset \mathbb{R}^d$. Furthermore, let K_h be a Gaussian transition kernel with bandwidth h and denote by k_h its associated kernel density. Finally, let p be a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant Z such that for all $y \in \mathsf{Y}$, $p(y) = Z \times$*

$[0.5\mathcal{N}(y;\theta_1^\star, \boldsymbol{I_d}) + 0.5\mathcal{N}(y;\theta_2^\star, \boldsymbol{I_d})]$, *where* $\theta_1^\star, \theta_2^\star \in \mathsf{T}$ *and* $\boldsymbol{I_d}$ *is the identity matrix. Then,* (2.18) *holds and we can apply Theorem 5 (see Section 2.A.3 for details).*

Notice that the $O(1/\sqrt{N})$ convergence rate from Theorem 5 holds under minimal assumptions on Ψ_α. However, bridging the gap with the $O(1/N)$ convergence rate in Theorem 3 typically requires much stronger smoothness and strong-convexity assumptions on Ψ_α which can be hard to satisfy in practice (see Bubeck, 2015, Theorem 6.2 for the statement of this result and Chérief-Abdellatif, Alquier, and Khan, 2019 for an example in Online Variational Inference). Bypassing any of these assumptions like we did in the ideal case in Theorem 3 in order to improve on Theorem 5 constitutes an interesting area of research which is beyond the scope of this thesis.

As for the stochastic version of Power Descent, we establish the total variation convergence of $\hat{\mathcal{I}}_{\alpha,M}(\mu)$ towards $\mathcal{I}_\alpha(\mu)$ as M goes to infinity for all $\mu \in \mathrm{M}_1(\mathsf{T})$. To do so, consider i.i.d random variables $Y_1, Y_2, \ldots$ with common density μk w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote by $\mathbb{E}$ the associated expectation operator. We then have Proposition 10 below.

Proposition 10. *Assume (2.A1). Let* $\alpha \in \mathbb{R} \setminus \{1\}$, $\eta > 0$, κ *be such that* $(\alpha-1)\kappa \geqslant 0$ *and set* $\Gamma(v) = [(\alpha-1)\,v + 1]^{\eta/(1-\alpha)}$ *for all* $v \in \mathrm{Dom}_\alpha$. *Let* $\mu \in \mathrm{M}_1(\mathsf{T})$ *be such that* $\Psi_\alpha(\mu k) < \infty$, (2.3) *holds and*

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta)\mathbb{E}\left[\left\{\frac{k(\theta, Y_1)}{\mu k(Y_1)}\left(\frac{\mu k(Y_1)}{p(Y_1)}\right)^{\alpha-1} + (\alpha-1)\kappa\right\}^{\frac{\eta}{1-\alpha}}\right] < \infty\,. \tag{2.20}$$

Then,

$$\lim_{M\to\infty} \left\|\hat{\mathcal{I}}_{\alpha,M}(\mu) - \mathcal{I}_\alpha(\mu)\right\|_{TV} = 0\,, \quad \mathbb{P}-\text{a.s.}$$

The proof is deferred to Section 2.A.7. The crux of the proof consists in applying a Dominated Convergence Theorem to non-negative real-valued $(\mathcal{T}\otimes\mathcal{F}, \mathcal{B}(\mathbb{R}_{\geqslant 0}))$-measurable functions, which requires to consider a Generalized version of the Dominated Convergence Theorem (Lemma 17) and an Integrated Law of Large Numbers (Lemma 18).

Mixture Models. We now address the case where $\hat{\mu}_1$ corresponds to a weighted sum of Dirac measures. This case is of particular interest to us since as we shall see, for any kernel K of our choice, the (α, Γ)-descent procedure simplifies and provides an update formula for the mixture weights of the corresponding mixture model $\hat{\mu}_1 K$.

Let $J \in \mathbb{N}^\star$ and let $\theta_1, \ldots, \theta_J \in \mathsf{T}$ be fixed. We start by introducing the simplex of $\mathbb{R}^J$

$$\mathcal{S}_J = \left\{\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J) \in \mathbb{R}^J \;:\; \forall j \in \{1, \ldots, J\}\,,\ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1\right\}\,,$$

and for all $\boldsymbol{\lambda} \in \mathcal{S}_J$, we define $\mu_{\boldsymbol{\lambda}} \in \mathrm{M}_1(\mathsf{T})$ by $\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$. Then, $\mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y)$ corresponds to a mixture model and if we let $(\hat{\mu}_n)_{n \in \mathbb{N}^\star}$ be defined by $\hat{\mu}_1 = \mu_{\boldsymbol{\lambda}}$ and

$$\hat{\mu}_{n+1} = \hat{\mathcal{I}}_{\alpha,M}(\hat{\mu}_n) \;, \qquad n \in \mathbb{N}^\star \;,$$

an immediate induction yields that for every $n \in \mathbb{N}^\star$, $\hat{\mu}_n$ can be expressed as $\hat{\mu}_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ where $\boldsymbol{\lambda}_n = (\lambda_{1,n}, \ldots, \lambda_{J,n}) \in \mathcal{S}_J$ satisfies the initialisation $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}$ and the update formula: for all $n \in \mathbb{N}^\star$ and all $j \in \{1, \ldots, J\}$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_i) + \kappa)} \;,$$

with $\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)$ given by (2.17) for all $j = 1 \ldots J$ and $Y_{1,n+1}, \ldots, Y_{M,n+1}$ drawn independently from $\hat{\mu}_n k$ conditionally on $\mathcal{F}_n$. This leads to Algorithm 6 below.

Algorithm 6: *Mixture Stochastic (α, Γ)-descent*

Input: p: measurable positive function, K: Markov transition kernel, M: number of samples, $\Theta_J = \{\theta_1, \ldots, \theta_J\} \subset \mathsf{T}$: parameter set.

Output: Optimised weights $\boldsymbol{\lambda}$.

Set $\boldsymbol{\lambda} = [\lambda_{1,1}, \ldots, \lambda_{J,1}]$.

while *not converged* **do**

Sampling step : Draw independently M samples $Y_1, \ldots, Y_M$ from $\mu_{\boldsymbol{\lambda}} k$.

Expectation step : Compute $\boldsymbol{B}_{\boldsymbol{\lambda}} = (b_j)_{1 \leqslant j \leqslant J}$ where for all $j = 1 \ldots J$

$$b_j = \frac{1}{M} \sum_{m=1}^M \frac{k(\theta_j, Y_m)}{\mu_{\boldsymbol{\lambda}} k(Y_m)} f'_\alpha \left(\frac{\mu_{\boldsymbol{\lambda}} k(Y_m)}{p(Y_m)} \right)$$

and deduce $\boldsymbol{W}_{\boldsymbol{\lambda}} = (\lambda_j \Gamma(b_j + \kappa))_{1 \leqslant j \leqslant J}$ and $w_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \Gamma(b_j + \kappa)$.

Iteration step : Set

$$\boldsymbol{\lambda} \leftarrow \frac{1}{w_{\boldsymbol{\lambda}}} \boldsymbol{W}_{\boldsymbol{\lambda}}$$

end

In this particular framework, most of the computing effort at each step lies within the computation of the vector $(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j))_{1 \leqslant j \leqslant J}$. Interestingly, these computations can also be used to obtain an estimate of the Evidence Lower Bound (resp. the VR bound (Li and Turner, 2016)) when $p = p(\cdot, \mathscr{D})$. These two quantities, which are

recalled in Chapter 1 and are given in our particular case by

$$\mathrm{ELBO}(\hat{\mu}_n k; \mathscr{D}) = -\sum_{j=1}^{J} \lambda_{j,n} b_{\hat{\mu}_n,\alpha}(\theta_j)$$

$$\mathcal{L}_\alpha(\hat{\mu}_n k; \mathscr{D}) = \frac{1}{1-\alpha} \log\left((\alpha - 1) \sum_{j=1}^{J} \lambda_{j,n} b_{\hat{\mu}_n,\alpha}(\theta_j) + 1 \right)$$

allow us to assess the convergence of the algorithm and provide a bound on the log-likelihood (see Li and Turner, 2016, Theorem 1). Note also that if there is a need for very large J, one can approximate the summation appearing in $\hat{\mu}_n k$ using subsampling.

An important point is that Algorithm 6 does not require any information on how the $\{\theta_1, \ldots, \theta_J\}$ have been obtained in order to infer the optimal weights as it draws information from samples that are generated from $\mu_{\boldsymbol{\lambda}} k$. Since the algorithm leaves $\{\theta_1, \ldots, \theta_J\}$ unchanged throughout the optimisation of the mixture weights (we call it an *Exploitation Step*), we then combine Algorithm 6 with an *Exploration step* that modifies the parameter set, which gives Algorithm 7 below.

Algorithm 7: *Complete Exploitation-Exploration Algorithm*

Input: p: measurable positive function, α: α-divergence parameter, (Γ, κ): chosen as per Table 2.1, q_0: initial sampler, K: Markov transition kernel, $(M_t)_t$: number of samples, $(J_t)_t$: dimension of the parameter set.

Output: Optimised weights $\boldsymbol{\lambda}$ and parameter set Θ.

Draw $\theta_{1,0}, \ldots, \theta_{J_0,0}$ from q_0. Set $t = 0$.

while *not converged* **do**

Exploitation step : Set $\Theta = \{\theta_{1,t}, \ldots, \theta_{J_t,t}\}$. Perform Mixture Stochastic (α, Γ)-descent and obtain $\boldsymbol{\lambda}$.

Exploration step : Perform any exploration step of our choice and obtain $\theta_{1,t+1}, \ldots, \theta_{J_{t+1},t+1}$. Set $t = t + 1$.

end

Note that this algorithm is very general, as any Exploration Step can be envisioned. We also have several other levels of generality in our algorithm since we are free to choose the kernel K, the α-divergence being optimised and we have stated different possible choices for the couple (Γ, κ).

As a side remark, notice also that we recover the mixture weights update rules from the Population Monte Carlo algorithm applied to reverse Kullback-Leibler minimisation (Douc et al., 2007a) by considering the Power Descent with $\alpha = 0$ and $\eta = 1$. We have thus embedded this special case into a more general framework.

We now move on to numerical experiments in the next section.

2.4 Numerical experiments

In this part, we want to assess how Algorithm 7 performs on both toy and real-world examples. To do so, we first need to specify the kernel K and an algorithm for the Exploration Step.

Kernel. Let K_h be a Gaussian transition kernel with bandwidth h and denote by k_h its associated kernel density. Given $J \in \mathbb{N}^\star$ and $\theta_1, \dots, \theta_J \in \mathsf{T}$, we then work within the approximating family

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\} .$$

Exploration Step. At time $t = 1 \dots T$, we resample among $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ according to the optimised mixture weights $\boldsymbol{\lambda}$. The obtained sample $\{\theta_{1,t+1}, \dots, \theta_{J_{t+1},t+1}\}$ is then perturbed stochastically using the Gaussian transition kernel K_{h_t}, which gives us our new parameter set. The hyperparameter h_t is adjusted according to the number of particles so that $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space (the optimal rate in nonparametric estimation when the function is at least 2-times continuously differentiable and the kernel has order 2 (Stone, 1982)).

Next, we are interested in the choice of α. The hyperparameter α allows us to choose between *mass-covering* divergences which tend to cover all the modes ($\alpha \ll 0$) and *mode-seeking* divergences that are attracted to the mode with the largest probability mass ($\alpha \gg 1$), the case $\alpha \in (0, 1)$ corresponding to a mix of the two worlds (see for example Minka, 2005).

Depending on the learning task, the optimal α may differ and understanding how to select the value of α is still an area of ongoing research. However, the case $\alpha < 1$ presents the advantage that $\hat{b}_{\mu,\alpha,M}$ is always finite. Indeed, for all $\alpha \in \mathbb{R} \setminus \{1\}$, we have

$$b_{\mu,\alpha}(\theta) = \frac{1}{\alpha - 1} \int_{\mathsf{Y}} \frac{k(\theta, y)}{\mu k(y)} \left(\frac{p(y, \mathscr{D})}{\mu k(y)} \right)^{1-\alpha} \mu k(y) \nu(\mathrm{d}y) - \frac{1}{\alpha - 1} ,$$

and as the dimension grows, the conditions of support are often not met in practice, meaning that there exists $A \in \mathcal{Y}$ such that $p(A, \mathscr{D}) = 0$ and $\mu k(A) > 0$. This implies that whenever $\alpha > 1$ we might have that $\hat{b}_{\mu,\alpha,M}(\theta) = \infty$ and that the α-divergence (or equivalently the VR bound) is infinite, which is the sort of behavior we would like to avoid. Thus, we restrict ourselves to the case $\alpha \leqslant 1$ in the following numerical experiments. Note that the limiting case $\alpha = 1$, corresponding to the

commonly-used forward Kullback-Leibler objective function, also suffers from this poor behavior, but is still considered in the experiments as a reference.

We now move on to our first example where we investigate the impact of different choices of Γ. The code for all the subsequent numerical experiments is available at https://github.com/kdaudel/AlphaGammaDescent.

2.4.1 Toy Example

Following Example 5, the target p is a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant Z such that

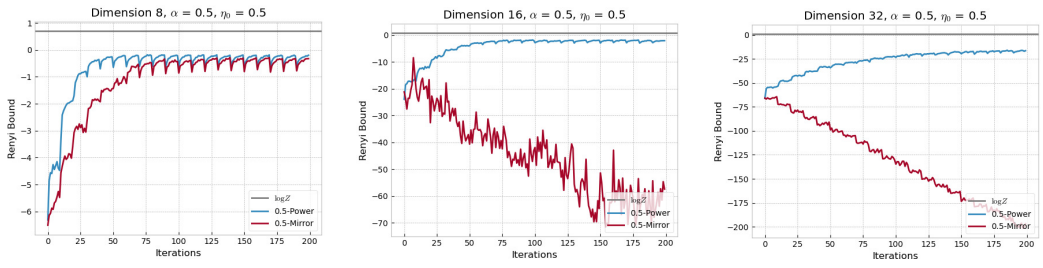
$$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})] \ ,$$

where $\boldsymbol{u_d}$ is the d-dimensional vector whose coordinates are all equal to 1, $s = 2$, $Z = 2$ and $\boldsymbol{I_d}$ is the identity matrix. $(J_t)_t$ and (M_t) are kept constant equal to $J = M = 100$, $\kappa = 0$ and the initial weights are set to be $[1/J, \dots, 1/J]$. The number of inner iterations in the (α, Γ)-descent is set to $N = 10$ and for all $n = 1 \dots N$, we use the adaptive learning rate $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.5$. We set the initial sampler to be a centered normal distribution with covariance matrix $5\boldsymbol{I_d}$. We compare three versions of the (α, Γ)-algorithm:

- 0.5-Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,
- 0.5-Power Descent : $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$,
- 1-Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$.

For each of them, we run $T = 20$ iterations of Algorithm 7 and we replicate the experiment 100 times for $d = \{8, 16, 32\}$. The results for the 0.5-Mirror and 0.5-Power Descent are displayed on Figure 2.1.

FIGURE 2.1: Plotted is the average VR bound (the Renyi-Bound axis) for the 0.5-Power and 0.5-Mirror Descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.



A first remark is that we are able to observe the monotonicity property from Theorem 2 (the VR bound varies like $\Psi_\alpha(\mu_n k)^{\alpha-1}$) for the 0.5-Power Descent, the jumps in the VR bound corresponding to an update of the parameter set. Furthermore, we see that the 0.5-Mirror Descent (which would have been the default choice based on

the existing optimisation literature) converges more slowly than the 0.5-Power Descent in dimension 8. An even more striking aspect however is that, as the dimension grows, the 0.5-Mirror Descent is unable to learn and the algorithm diverges.

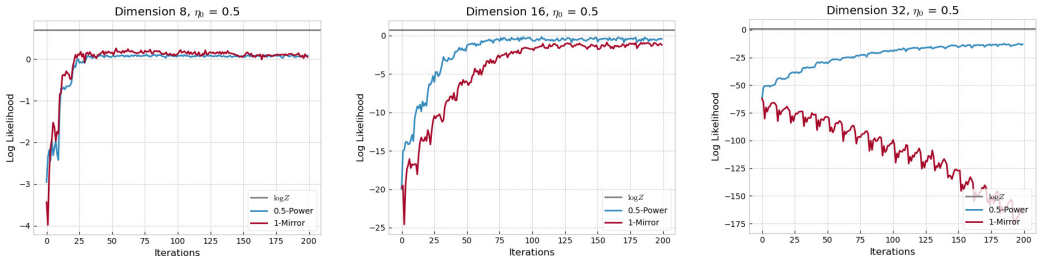
These two different behaviors for the Power and Mirror Descent can be explained by rewriting the update formulas for any $\alpha < 1$ under the form

$$\begin{aligned} \text{Mirror}: \quad & \lambda_{j,n} \propto e^{\frac{\eta}{1-\alpha}\left[(\alpha-1)b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j)+(\alpha-1)\kappa\right]} \\ \text{Power}: \quad & \lambda_{j,n} \propto e^{\frac{\eta}{1-\alpha}\log\left[(\alpha-1)b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j)+(\alpha-1)\kappa\right]} . \end{aligned}$$

In the Power case, an extra log transformation has been added, which allows to discriminate between small values of $b_{\mu_{\boldsymbol{\lambda}_n},\alpha}$. Since the values of $b_{\mu_{\boldsymbol{\lambda}_n},\alpha}$ tend to get smaller as the dimension grows, the impact of adding an extra log transformation becomes increasingly visible: the Mirror Descent becomes more and more unable to differentiate between the different particles $\{\theta_1, \dots, \theta_J\}$ and is thus unable to learn.

Finally, we compare how the 0.5-Power and 1-Mirror Descent perform at approximating the log-likelihood in dimension $d = \{8, 16, 32\}$. The results are plotted on Figure 2.2. Again, the 0.5-Power Descent comes across as faster and more stable compared to the 1-Mirror Descent as the dimension grows. Furthermore, it also does not fail in dimension 32, unlike the 1-Mirror Descent.

FIGURE 2.2: Plotted is the average Log-likelihood for 0.5-Power and 1-Mirror Descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.



Consequently, we see on this simple yet illustrative example that the Power Descent is a suitable alternative to the Mirror Descent as the dimension grows.

We are next interested in seeing how the (α, Γ)-descent performs on a real-data example. Based on the numerical results obtained so far, we rule out the Mirror Descent for $\alpha \leqslant 1$ and we focus on the Power Descent in our second example.

2.4.2 Bayesian Logistic Regression

We consider the Bayesian Logistic Regression from Example 1 with $a = 1$ and $b = 0.01$.

We test our algorithm for the *Covertype* dataset ($581,012$ data points and 54 features, available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html). Computing $p(y, \mathscr{D})$ constitutes the major computation

bottleneck here, since $p(y, \mathscr{D}) = p_0(y) \prod_i p(x_i|y)$ with a very large number of data points. We can conveniently address this problem by approximating $p(y, \mathscr{D})$ with subsampled mini-batches. We adopt this strategy here and consider mini-batches of size 100.

We set $\alpha = 0.5$, $N = 1$, $T = 500$, $\kappa = 0$, $J_0 = M_0 = 20$ and $J_{t+1} = M_{t+1} = J_t + 1$ for $t = 1 \dots T$ in Algorithm 7. The initial weights in the (α, Γ)-descent are set to $\boldsymbol{\lambda}_{init,t} = [1/J_t, \dots, 1/J_t]$ and the learning rate is set to $\eta_0 = 0.05$.

One thing that is very specific to the Exploration step that we used to run our experiments (and sampling-based Exploration steps algorithms in general) is that the particles $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ are sampled from a known distribution at each Exploration step. This means that we are able to infer information on $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ using Importance Sampling (IS) weights. We thus compare the Power (α, Γ)-descent with a state-of-the-art Adaptive Importance Sampling-based (AIS) algorithm (see for example Oh and Berger, 1992; Kloek and Van Dijk, 1978; Chopin, 2004 and Delyon and Portier, 2021).

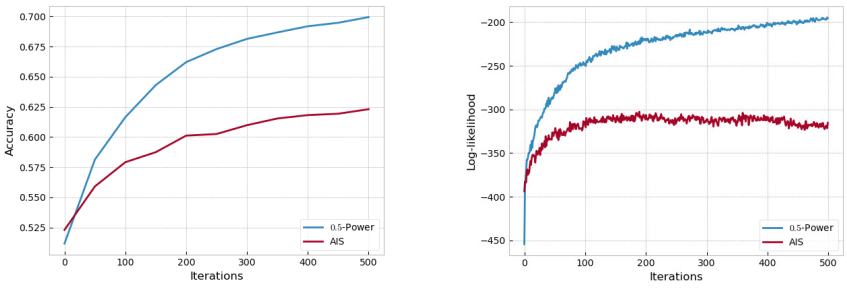
We initialise $\{\theta_{1,0}, \dots, \theta_{J_0,0}\}$ by sampling J_0 points independently from the prior $p_0(y) = p_0(\beta)p_0(w|\beta)$ and we set $q_0 = p_0$. Given q_t at time t, we draw J_t i.i.d samples $(\theta_{j,t})_{1 \leqslant j \leqslant J_t}$ from q_t and we define $q_{t+1}(y) = \sum_{j=1}^{J_t} \lambda_{j,t} k_{h_t}(y - \theta_{j,t})$ where

$$\lambda_{j,t} \propto \begin{cases} \frac{p(\theta_{j,t}, \mathscr{D})}{q_t(\theta_{j,t})} & \text{(AIS)} , \\ \Gamma(\hat{b}_{\mu_{\boldsymbol{\lambda}_{init,t}}, \alpha, M}(\theta_{j,t}) + \kappa) & \text{(Power)} . \end{cases} \tag{2.21}$$

Note that these two algorithms are computationally equivalent. Indeed, we choose $J_t = M_t$ and $N = 1$, that is we use an average of one sample from each $k(\theta_{j,t}, \cdot)$ to infer information on the relevance of the $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ with respect to one another. Comparatively, the AIS algorithm uses information directly available by computing the IS weights for $\{\theta_{1,t}, \dots, \theta_{J,t}\}$.

We replicate the experiments 100 times. The Accuracy and Log-likelihood averaged over the 100 trials for both algorithms are displayed on Figure 2.3 and we see that the 0.5-Power Descent outperforms the AIS algorithm.

FIGURE 2.3: Plotted are the average Accuracy and Log-likelihood computed over 100 replicates for Bayesian Logistic Regression on the Covertype dataset for the 0.5-Power Descent and the AIS algorithm.



2.5 Conclusion and perspectives

We introduced the (α, Γ)-descent and studied its convergence. Our framework recovers the Entropic Mirror Descent and allows us to introduce the Power Descent. Furthermore, our procedure provides a gradient-based method to optimise the mixture weights of any given mixture model, without any information on the underlying distribution of the variational parameters. We demonstrated empirically the benefit of going beyond the Entropic Mirror Descent framework by using the Power Descent algorithm instead, which is a more scalable alternative.

At this stage, we can think of several directions to extend our work on both a theoretical and a practical level.

(i) *Convergence.* One could seek to establish additional convergence results for the Power Descent. For example, since the case $\alpha < 1$ is advantageous in practice due to its mass-covering property, one may want to alleviate some of the hard-to-satisfy assumptions in Theorem 4 leading to the convergence of the Power Descent when $\alpha < 1$.

(ii) *Numerical results.* One can also be interested in understanding more precisely why the Entropic Mirror Descent appears to fail numerically compared to the Power Descent, even though these algorithms are linked to one another via the (α, Γ)-descent framework.

(iii) *Exploration Step.* As the (α, Γ)-descent does not make assumptions on the variational parameter θ, many methods can be envisioned as an Exploration step and combined with the (α, Γ)-descent besides the one we have used in Chapter 2 for illustrative purposes. One may then attempt to find Exploration steps that can efficiently be paired up with the (α, Γ)-descent.

In the following chapter, we will focus on the aspects raised in (i) and (ii), while Chapter 4 is devoted to (iii).

2.A Deferred results

2.A.1 Proof of Theorem 3

We start with a side note on Dom_α. A typical choice for Dom_α is

$$\mathrm{Dom}_\alpha = [-|b|_{\infty,\alpha} + \kappa, |b|_{\infty,\alpha} + \kappa] \,. \tag{2.22}$$

However, when $\alpha \in \mathbb{R} \setminus \{1\}$, we might consider instead

$$\mathrm{Dom}_\alpha = \begin{cases} [\frac{1}{1-\alpha} + \kappa, |b|_{\infty,\alpha} + \kappa], & \text{if } \alpha > 1 \\ [-|b|_{\infty,\alpha} + \kappa, \frac{1}{1-\alpha} + \kappa], & \text{if } \alpha < 1 \end{cases} \tag{2.23}$$

to underline the fact that for all $v \in \mathrm{Dom}_\alpha$, $(\alpha-1)v + 1 \geqslant (\alpha-1)\kappa$. Unless specified otherwise, we let Dom_α be as in (2.23) whenever $\alpha \in \mathbb{R} \setminus \{1\}$.

Proof of Theorem 3. Let us recall the different conditions that must be met in order to verify that we can apply Theorem 2 in each of the cases mentioned in Theorem 3:

1. $0 < \inf_{v \in \mathrm{Dom}_\alpha} \Gamma(v)$ and $\sup_{v \in \mathrm{Dom}_\alpha} \Gamma(v) < \infty$.

2. The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v) + 1 \geqslant 0 \,.$$

3. $L_{\alpha,1} = \inf_{v \in \mathrm{Dom}_\alpha} \{[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v) + 1\} \times \inf_{v \in \mathrm{Dom}_\alpha} -\Gamma'(v) > 0$.

4. The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is L-smooth and the function $-\log\Gamma$ is concave increasing.

5. $L_{\alpha,2} = (\inf_{v \in \mathrm{Dom}_\alpha} (-\log\Gamma)'(v))^{-1} > 0$.

(i) Forward Kullback-Leibler divergence ($\alpha = 1$): $\Gamma(v) = e^{-\eta v}$, $\eta \in (0,1)$, any real κ. Since the update formula does not depend on κ, there is no constraint on κ and we assume that $\kappa = 0$ for simplicity.

- Condition 1 is satisfied since $|b|_{\infty,1}$ is finite.
- Condition 2 is satisfied with $\Gamma'(v) = -\eta e^{-\eta}$ and $(\log\Gamma)'(v) = -\eta$.
- Condition 3 is satisfied with $L_{1,1} \geqslant (1-\eta)\eta e^{-\eta|b|_{\infty,1}}$.
- Condition 4 is satisfied.
- Condition 5 is satisfied with $L_{1,2} = \frac{1}{\eta}$.

(ii) Reverse Kullback-Leibler ($\alpha = 0$) and α-Divergence with $\alpha \in \mathbb{R}\,\{0,1\}$:

(a) $\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any real κ. The only difference with the previous case lies in the inequality (i.e. Condition 2), which can be rewritten for all $v \in \mathrm{Dom}_\alpha$ as

$$1 \geqslant \eta\left[(\alpha-1)(v-\kappa)+1\right] ,$$

Since $0 \leqslant (\alpha-1)(v-\kappa)+1 \leqslant |\alpha-1||b|_{\infty,\alpha}+1$, this inequality is then satisfied for $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$.

(b) Case $\alpha > 1$. $\Gamma(v) = ((\alpha-1)v+1)^{\frac{\eta}{1-\alpha}}$, $\eta \in (0,1]$ and κ satisfies $(\alpha-1)\kappa > 0$. Then, the condition $(\alpha-1)\kappa > 0$ ensures that Γ is well-defined on Dom_α. From there, we deduce:

- Condition 1 is satisfied since $|b|_{\infty,\alpha}$ is finite.
- Condition 2 is satisfied: $\Gamma'(v) = -\eta((\alpha-1)v+1)^{\frac{\eta}{1-\alpha}-1}$, $(\log\Gamma)'(v) = \frac{-\eta}{(\alpha-1)v+1}$ and the inequality can be rewritten for all $v \in \mathrm{Dom}_\alpha$ as

$$1 \geqslant \eta\left[1 - \frac{(\alpha-1)\kappa}{(\alpha-1)v+1}\right] ,$$

which is satisfied for $\eta \in (0,1]$.
- Condition 3 is satisfied (the condition $(\alpha-1)\kappa > 0$ is of crucial importance here).
- Condition 4 is satisfied with $(-\log\Gamma)''(v) = \frac{\eta(1-\alpha)}{((\alpha-1)v+1)^2}$ (note that we need $\alpha > 1$ here).
- Condition 5 is satisfied and here again we use that $(\alpha-1)\kappa > 0$.

□

2.A.2 Proof of Theorem 4

In this part, recall that we focus on the particular case $\alpha < 1$, $\kappa \leqslant 0$ and $\Gamma(v) = [(\alpha-1)\,v+1]^{\eta/(1-\alpha)}$ for all $v \in \mathrm{Dom}_\alpha$. In the following, we use the notation $\mu_n \Rightarrow \mu^\star$ for the weak convergence of measures in $\mathrm{M}_1(\mathsf{T})$. For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, for all $\theta \in \mathsf{T}$, define

$$g_\zeta(\theta) = (\alpha-1)(b_{\zeta,\alpha}(\theta)+\kappa)+1 .$$

We first derive four useful lemmas.

Lemma 11. *Assume (2.A1) and (2.A4). Suppose that $\mu_n \Rightarrow \mu^\star$. Then the following assertions hold.*

(i) For all $y \in \mathsf{Y}$, $\mu_n k(y)$ tends to $\mu^\star k(y)$ as $n \to \infty$.

(ii) For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, the function $\theta \mapsto g_\zeta(\theta)$ is continuous. Furthermore for all $\theta \in \mathsf{T}$, $g_{\mu_n}(\theta)$ tends to $g_{\mu^\star}(\theta)$ as $n \to \infty$.

(iii) There exist $0 < m_- < m_+ < \infty$ such that, for all $\zeta \in \mathrm{M}_1(\mathsf{T})$ and $\theta \in \mathsf{T}$, $g_\zeta(\theta) \in [m_-, m_+]$.

(iv) For all continuous, positive and bounded function h,

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)\Gamma(b_{\mu_n,\alpha}(\theta)+\kappa)h(\theta) = \int_{\mathsf{T}} \mu^\star(\mathrm{d}\theta)\Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)h(\theta)\ .$$

Proof. We prove the assertions successively.

Proof of (i). For all $y \in \mathsf{Y}$, the function $\theta \mapsto k(\theta, y)$ is continuous on a compact set, hence bounded. The weak convergence $\mu_n \Rightarrow \mu^\star$ thus implies the pointwise convergence of $\mu_n k$ to $\mu^\star k$.

Proof of (ii). For all $\theta \in \mathsf{T}$ and $\zeta \in \mathrm{M}_1(\mathsf{T})$, we write

$$g_\zeta(\theta) = \int_{\mathsf{Y}} a_\zeta(\theta, y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\ ,$$

where we set for all $(\theta, y) \in \mathsf{T}\times\mathsf{Y}$, $a_\zeta(\theta, y) = k(\theta, y)\left(\frac{\zeta k(y)}{p(y)}\right)^{\alpha-1}$. The continuity of $g_\zeta(\theta)$ follows from the Dominated Convergence Theorem, since for all $y \in \mathsf{Y}$, the function $\theta \mapsto a_\zeta(\theta, y)$ is continuous on T by (2.A4)-(ii) and for all $(\theta, y) \in \mathsf{T}\times\mathsf{Y}$, we have

$$|a_\zeta(\theta, y)| \leqslant \sup_{\theta'\in\mathsf{T}} k(\theta', y) \times \sup_{\theta''\in\mathsf{T}} \left(\frac{k(\theta'', y)}{p(y)}\right)^{\alpha-1}\ , \tag{2.24}$$

which is integrable w.r.t $\nu(\mathrm{d}y)$ by (2.A4)-(iii). The second part of (ii) is obtained similarly. Using (i) and that $u \mapsto u^{\alpha-1}$ is C^1, we get that, for all $(\theta, y) \in \mathsf{T}\times\mathsf{Y}$,

$$\lim_{n\to\infty} k(\theta, y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1} = k(\theta, y)\left(\frac{\mu^\star k(y)}{p(y)}\right)^{\alpha-1}\ ,$$

i.e. $\lim_{n\to\infty} a_{\mu_n}(\theta, y) = a_{\mu^\star}(\theta, y)$. The bound (2.24) and (2.A4)-(iii) provide a domination criterion and we get that $g_{\mu_n}(\theta)$ tends to $g_{\mu^\star}(\theta)$ as $n \to \infty$, which concludes the proof of (ii).

Proof of (iii). For all $(\theta, \zeta) \in \mathsf{T}\times\mathrm{M}_1(\mathsf{T})$, we have $g_\zeta(\theta) \in [m_-, m_+]$ where

$$\begin{aligned} m_- &:= \int_{\mathsf{Y}} \inf_{\theta'\in\mathsf{T}} k(\theta', y) \times \inf_{\theta''\in\mathsf{T}} \left(\frac{k(\theta'', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) + (\alpha-1)\kappa\ , \\ m_+ &:= \int_{\mathsf{Y}} \sup_{\theta'\in\mathsf{T}} k(\theta', y) \times \sup_{\theta''\in\mathsf{T}} \left(\frac{k(\theta'', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) + (\alpha-1)\kappa\ . \end{aligned} \tag{2.25}$$

We have that m_+ is finite by (2.A4)-(iii). Furthermore, $u \mapsto u^{\alpha-1}$ does not vanish on $(0, \infty)$. Together with (2.A1), we thus have that for any $y \in \mathsf{Y}$, the functions $\theta \mapsto k(\theta, y)$ and $\theta \mapsto (k(\theta, y)/p(y))^{\alpha-1}$ are continuous and positive on the compact set T, from which we deduce that $m_- > 0$.

Proof of (iv). Using (ii), the function $\theta \mapsto \Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)h(\theta)$ is continuous, and, since T is compact, $\mu_n \Rightarrow \mu^\star$ gives that

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)\Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)h(\theta) = \int_{\mathsf{T}} \mu^\star(\mathrm{d}\theta)\Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)h(\theta)\ . \tag{2.26}$$

Next we show that

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) \left|\Gamma(b_{\mu_n,\alpha}(\theta)+\kappa) - \Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)\right| h(\theta) = 0 \tag{2.27}$$

ie

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) \left| g_{\mu_n}(\theta)^{\frac{\eta}{1-\alpha}} - g_{\mu^\star}(\theta)^{\frac{\eta}{1-\alpha}} \right| h(\theta) = 0$$

Using (iii), since $u \mapsto u^{\frac{\eta}{1-\alpha}}$ is Lipschitz on $[m_-, m_+]$, there exists a constant C such that

$$\begin{aligned}\mu_n\left[\left| g_{\mu_n}(\theta)^{\frac{\eta}{1-\alpha}} - g_{\mu^\star}(\theta)^{\frac{\eta}{1-\alpha}} \right| h\right] &\leqslant C \sup_{\theta\in\mathsf{T}} h(\theta) \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) \left| g_{\mu_n}(\theta) - g_{\mu^\star}(\theta)\right| \\ &= C \sup_{\theta\in\mathsf{T}} h(\theta) \int_{\mathsf{Y}} |a_n(y)| \nu(\mathrm{d}y)\end{aligned}$$

where $a_n(y) := \mu_n k(y) \left\{ \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1} - \left(\frac{\mu^\star k(y)}{p(y)}\right)^{\alpha-1} \right\}$. Now, for all $y \in \mathsf{Y}$,

$$|a_n(y)| \leqslant 2 \sup_{\theta\in\mathsf{T}} k(\theta, y) \times \sup_{\theta'\in\mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} ,$$

which is integrable w.r.t ν by (2.A4)-(iii). Moreover, by (i) and by continuity of $u \mapsto u^{\alpha-1}$, we have $\lim_{n\to\infty} a_n(y) = 0$, and (2.27) follows by dominated convergence. Finally, combining (2.26), (2.27) and

$$\begin{aligned}\mu_n\left[\Gamma(b_{\mu_n,\alpha}(\theta)+\kappa)h\right] = \mu_n\left[\Gamma(b_{\mu_n,\alpha}(\theta)+\kappa)h - \Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)h\right] \\ + \mu_n\left[\Gamma(b_{\mu^\star,\alpha}(\theta)+\kappa)h\right] ,\end{aligned}$$

we obtain (iv), and the proof is concluded. □

Lemma 12. *Assume (2.A1). Let $\mu^\star, \mu \in \mathrm{M}_1(\mathsf{T})$ and assume that there exists $\bar{\mu} \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi_\alpha(\bar{\mu}k) < \Psi_\alpha(\mu^\star k)$. Then, there exists $\delta > 1$ such that*

$$\bar{\mu}(g_{\mu^\star} > \delta\mu^\star(g_{\mu^\star})) > 0 . \tag{2.28}$$

Proof. Let $\zeta, \zeta' \in \mathrm{M}_1(\mathsf{T})$. Then, by convexity of f_α we have,

$$\int_{\mathsf{T}} [\zeta - \zeta'](\mathrm{d}\theta) b_{\zeta',\alpha}(\theta) \leqslant \Psi_\alpha(\zeta k) - \Psi_\alpha(\zeta' k) ,$$

that is

$$\int_{\mathsf{T}} [\zeta - \zeta'](\mathrm{d}\theta) g_{\zeta'}(\theta) \geqslant (\alpha-1)\left(\Psi_\alpha(\zeta k) - \Psi_\alpha(\zeta' k)\right) . \tag{2.29}$$

Furthermore, for all $\delta > 1$, $(\delta-1)\mu^\star(g_{\mu^\star}) \geqslant 0$. Let us define $A_\delta = \{g_{\mu^\star} >$

$\delta\mu^\star(g_{\mu^\star})\}$ and show that $\bar{\mu}(A_\delta) > 0$ for some $\delta > 1$. To do so, we proceed by contradiction. Suppose that $\bar{\mu}(A_\delta) = 0$ for all $\delta > 1$, so that

$$\bar{\mu}[g_{\mu^\star} - \mu^\star(g_{\mu^\star})] = \bar{\mu}[(g_{\mu^\star} - \mu^\star(g_{\mu^\star}))\,\mathbf{1}_{A_\delta^c}] \leqslant (\delta - 1)\mu^\star(g_{\mu^\star}) \,.$$

Using (2.29), we get that, for all $\delta > 1$,

$$0 < (\alpha - 1)\left(\Psi_\alpha(\bar{\mu}k) - \Psi_\alpha(\mu^\star k)\right) \leqslant \bar{\mu}[(g_{\mu^\star} - \mu^\star(g_{\mu^\star}))] \leqslant (\delta - 1)\mu^\star(g_{\mu^\star}) \,.$$

Letting $\delta \downarrow 1$, we obtain a contradiction, which finishes the proof. □

Lemma 13. *Assume (2.A1). Let $\mu^\star \in \mathrm{M}_1(\mathsf{T})$ be a fixed point of $\mathcal{I}_\alpha$ and let $\eta > 0$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ and assume that there exists $\bar{\mu} \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star k) > \Psi_\alpha(\bar{\mu}k)$. Then, there exists $\delta > 1$ such that*

$$\bar{\mu}\left\{\Gamma(b_{\mu^\star,\alpha} + \kappa) > \delta\mu^\star(\Gamma(b_{\mu^\star,\alpha} + \kappa))\right\} > 0 \,.$$

Proof. Note that (2.3) holds for any $\eta > 0$ and ζ (in particular $\zeta = \mu^\star$) by Lemma 11-(iii). As $\mu^\star$ is a fixed point of $\mathcal{I}_\alpha$, $g_{\mu^\star}$ is $\mu^\star$-almost all constant. Consequently, we have that $\mu^\star(g_{\mu^\star})^{\eta/1-\alpha} = \mu^\star(g_{\mu^\star}^{\eta/1-\alpha}) = \mu^\star(\Gamma(b_{\mu^\star,\alpha} + \kappa))$. For all $\delta > 1$, $\delta' := \delta^{(1-\alpha)/\eta} > 1$ and

$$\begin{aligned}\bar{\mu}\left\{\Gamma(b_{\mu^\star,\alpha} + \kappa) > \delta\mu^\star(\Gamma(b_{\mu^\star,\alpha} + \kappa))\right\} &= \bar{\mu}\left\{g_{\mu^\star} > \delta^{(1-\alpha)/\eta}[\mu^\star(g_{\mu^\star}^{\eta/(1-\alpha)})]^{(1-\alpha)/\eta}\right\} \\ &= \bar{\mu}(g_{\mu^\star} > \delta'\mu^\star(g_{\mu^\star})) \,.\end{aligned}$$

We conclude by applying Lemma 12. □

Lemma 14. *Assume (2.A1) and (2.A4). Let $\eta > 0$, let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ and define the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ according to (2.2). Suppose that $\mu_n \Rightarrow \mu^\star$ for some fixed point $\mu^\star \in \mathrm{M}_1(\mathsf{T})$ of $\mathcal{I}_\alpha$. Further assume there exists $\bar{\mu} \in \mathrm{M}_{1,\mu_1}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star k) > \Psi_\alpha(\bar{\mu}k)$. Then, there exist $\delta > 1$ and $n \in \mathbb{N}^*$ such that*

$$\bar{\mu}\left(\bigcap_{m\geqslant n}\left\{\Gamma(b_{\mu_m,\alpha} + \kappa) > \delta\mu_m(\Gamma(b_{\mu_m,\alpha} + \kappa))\right\}\right) > 0 \,.$$

Proof. First note that the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ is well-defined for any $\eta > 0$ by Lemma 11-(iii), which implies $\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa)) > 0$ for all $n \in \mathbb{N}^\star$. For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, set $h_\zeta(\theta) = \Gamma(b_{\zeta,\alpha}(\theta) + \kappa)$. We further have that

$$\begin{aligned}\lim_{n\to\infty}\bar{\mu}\left(\bigcap_{m\geqslant n}\{h_{\mu_m} > \delta\mu_m(h_{\mu_m})\}\right) &= \bar{\mu}\left(\bigcup_{n\geqslant 1}\bigcap_{m\geqslant n}\{h_{\mu_m} > \delta\mu_m(h_{\mu_m})\}\right) \\ &= \bar{\mu}\left(\left\{\theta \in \mathsf{T} \,:\, \liminf_{n\to\infty}\frac{h_{\mu_n}(\theta)}{\mu_n(h_{\mu_n})} > \delta\right\}\right) \,.\end{aligned}$$

Furthermore, applying (ii) and (iv) in Lemma 11, we have, $\lim_{n\to\infty} h_{\mu_n}(\theta) = h_{\mu^\star}(\theta)$ for all $\theta \in \mathsf{T}$ and $\lim_{n\to\infty} \mu_n(h_{\mu_n}) = \mu^\star(h_{\mu^\star})$. Hence, for all $\theta \in \mathsf{T}$,

$$\liminf_{n\to\infty} \frac{h_{\mu_n}(\theta)}{\mu_n(h_{\mu_n})} = \frac{h_{\mu^\star}(\theta)}{\mu^\star(h_{\mu^\star})} \, .$$

The proof is concluded by applying Lemma 13. □

Proof of Theorem 4. Assume (2.A1) and (2.A4).

Lemma 11-(iii) is exactly the first result we want to obtain, that is: for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, any $\eta > 0$ satisfies (2.3) for ζ. Furthermore, $|\Psi_\alpha(\zeta k)| < \infty$ by (2.A4)-(iii).

Assume that $(\mu_n)_{n\in\mathbb{N}^\star}$ weakly converges to $\mu^\star \in \mathrm{M}_1(\mathsf{T})$. First note that Lemma 11-(iii) implies that for any $\eta > 0$ the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ is well-defined and $\mu^\star$ satisfies (2.3). Using Theorem 1, we obtain that the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ is decreasing for all $\eta \in (0, 1]$, which gives Assertion (i).

We now prove Assertions (ii) and (iii) successively.

Proof of (ii). For all $\zeta \in \mathrm{M}_1(\mathsf{T})$ and all $y \in \mathsf{Y}$, set $a_\zeta(y) = f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) p(y)$, leading to

$$\Psi_\alpha(\zeta k) = \int_{\mathsf{Y}} a_\zeta(y)\nu(\mathrm{d}y) \, . \tag{2.30}$$

Then, for all $y \in \mathsf{Y}$,

$$|a_\zeta(y)| \leqslant \sup_{\theta\in\mathsf{T}} \left| f_\alpha\left(\frac{k(\theta, y)}{p(y)}\right)\right| p(y) \, , \tag{2.31}$$

which is integrable w.r.t $\nu(\mathrm{d}y)$ by (2.A4)-(iii). Furthermore, recall that for all $y \in \mathsf{Y}$,

$$[\mathcal{I}_\alpha(\mu_n)k](y) = \frac{\int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)\Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)k(\theta, y)}{\int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)\Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)} \, .$$

By applying twice Lemma 11-(iv) with $h(\theta) = 1$ and $h(\theta) = k(\theta, y)$, we have that for all $y \in \mathsf{Y}$,

$$\lim_{n\to\infty} [\mathcal{I}_\alpha(\mu_n)k](y) = [\mathcal{I}_\alpha(\mu^\star)k](y) \, . \tag{2.32}$$

Now, since f_α is C^1, we obtain from Lemma 11-(i) and (2.32) respectively that for all $y \in \mathsf{Y}$, $\lim_{n\to\infty} a_{\mu_n}(y) = a_{\mu^\star}(y)$ and $\lim_{n\to\infty} a_{\mathcal{I}_\alpha(\mu_n)}(y) = a_{\mathcal{I}_\alpha(\mu^\star)}(y)$. Combining with (2.31) and (2.30) we can thus apply the Dominated Convergence Theorem to obtain

$$\lim_{n\to\infty} \Psi_\alpha(\mu_n k) = \Psi_\alpha(\mu^\star k) \tag{2.33}$$

and

$$\lim_{n\to\infty} \Psi_\alpha(\mu_{n+1}k) = \lim_{n\to\infty} \Psi_\alpha(\mathcal{I}_\alpha(\mu_n)k) = \Psi_\alpha(\mathcal{I}_\alpha(\mu^\star)k) \, . \tag{2.34}$$

Finally, (2.33) and (2.34) together yield $\Psi_\alpha(\mu^\star) = \Psi_\alpha(\mathcal{I}_\alpha(\mu^\star)k)$, which in turn implies that $\mu^\star$ is a fixed point of $\mathcal{I}_\alpha$ according to Theorem 1-(ii).

Proof of (iii). We prove (iii) by contradiction. Suppose that $\mu_n \Rightarrow \mu^\star$, where $\mu^\star$ is a fixed point of $\mathcal{I}_\alpha$ that satisfies

$$\Psi_\alpha(\mu^\star k) > \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k) \ .$$

Then, there exists $\bar{\mu} \in \mathrm{M}_{1,\mu_1}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star k) > \Psi_\alpha(\bar{\mu} k)$. Now for all $n \in \mathbb{N}^\star$, set

$$B_n = \left\{ \theta \in \mathsf{T} \ : \ \bigcap_{m \geqslant n} \{ h_{\mu_m}(\theta) > \delta \mu_m(h_{\mu_m}) \} \right\} \ ,$$

where for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, for all $\theta \in \mathsf{T}$, $h_\zeta(\theta) := \Gamma(b_{\zeta,\alpha}(\theta) + \kappa)$. There exist, according to Lemma 14, a well chosen $\delta > 1$ and a sufficiently large n_0 such that $\bar{\mu}(B_{n_0}) > 0$.

Furthermore $\bar{\mu} \approx \mu_1$ by definition, where $\zeta \approx \mu_1$ if and only if for all $A \in \mathcal{T}$: $\zeta(A) > 0$ is equivalent to $\mu_1(A) > 0$. Since $0 < \Gamma(b_{\mu_1,\alpha}(\theta) + \kappa) < \infty$ for μ_1-almost all $\theta \in \mathsf{T}$ and $\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \propto \Gamma(b_{\mu_1,\alpha} + \kappa)$, we also have $\mu_2 \approx \mu_1$. Then by induction, $\mu_n \approx \mu_1$ for all $n \in \mathbb{N}^\star$. Finally, $\mu_{n_0}(B_{n_0}) > 0$. Moreover, for all $\theta \in B_{n_0}$ and all $m > n_0$, $\frac{h_{\mu_m}(\theta)}{\mu_m(h_{\mu_m})} > \delta$ and consequently

$$\mu_m(B_{n_0}) = \int_{B_{n_0}} \mu_{m-1}(\mathrm{d}\theta) \frac{h_{\mu_{m-1}}(\theta)}{\mu_{m-1}(h_{\mu_{m-1}})} \geqslant \delta \mu_{m-1}(B_{n_0}) \ .$$

By induction on m we get that, for all $m \geqslant n_0$, $\mu_m(B_{n_0}) \geqslant \delta^{m-n_0} \mu_{n_0}(B_{n_0})$. This contradicts the previously obtain facts that $\delta > 1$ and $\mu_{n_0}(B_{n_0}) > 0$. Therefore we get a contradiction and the proof is concluded. □

2.A.3 Adapting Theorem 3 in the stochastic case

Here, we want to adapt Theorem 3 to the stochastic case for the Entropic Mirror Descent. Given $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$ with $\Psi_\alpha(\hat{\mu}_1 k) < \infty$ and letting $(\hat{\mu}_n)_{n \in \mathbb{N}^\star}$ be defined by (2.16), we will need the following additionnal assumption on the sequence of iterates $(\mu_n)_{n \in \mathbb{N}^\star}$, which controls the difference $|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta) - b_{\hat{\mu}_n,\alpha}(\theta)|$ uniformly with respect to θ.

(2.A5) Assume that there exists $\sigma > 0$ such that for all $n \in \mathbb{N}^\star$,

$$\mathbb{E}\left[\sup_{\theta \in \mathsf{T}} \left| \hat{b}_{\hat{\mu}_n,\alpha,M}(\theta) - b_{\hat{\mu}_n,\alpha}(\theta) \right| \right] \leqslant \frac{\sigma}{\sqrt{M}} \ .$$

Before stating the result, let us first comment on the validity of this assumption.

Validity of Assumption (2.A5), an example. Set $g_n(\theta, y) = \frac{k(\theta,y)}{\hat{\mu}_n k(y)} f'_\alpha\left(\frac{\hat{\mu}_n k(y)}{p(y)}\right)$ for all $\theta \in \mathsf{T}$, all $n \in \mathbb{N}^\star$ and all $y \in \mathsf{Y}$. In the particular case of the Simplex framework (see Example 3), which is the case we use in practice, (2.A5) holds with $\sigma = 2C\sqrt{\frac{\log(2J)}{2}}$,

where C is a positive constant satisfying $|g_n(\theta_j, Y_{m,n+1})| \leqslant C$ almost-surely for all $j = 1 \dots J$, all $n \in \mathbb{N}^\star$ and all $m = 1 \dots M$.

Proof. For all $u > 0$, we have by Jensen's inequality that

$$e^{u\mathbb{E}\left[\max_{1\leqslant j\leqslant J} M\left|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)-b_{\hat{\mu}_n,\alpha}(\theta_j)\right|\right]} \leqslant \mathbb{E}\left[e^{u\max_{1\leqslant j\leqslant J} M\left|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)-b_{\hat{\mu}_n,\alpha}(\theta_j)\right|}\right] \quad (2.35)$$

Furthermore, Hoeffding's lemma implies

$$\mathbb{E}\left[e^{u\left\{g_n(\theta_j,Y_{m,n+1})-b_{\hat{\mu}_n,\alpha}(\theta_j)\right\}}\right] \leqslant e^{\frac{u^2C^2}{2}}$$

and consequently

$$\mathbb{E}\left[e^{uM\left\{\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)-b_{\hat{\mu}_n,\alpha}(\theta_j)\right\}}\right] \leqslant e^{\frac{Mu^2C^2}{2}} .$$

Similarly, we have

$$\mathbb{E}\left[e^{-uM\left\{\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)-b_{\hat{\mu}_n,\alpha}(\theta_j)\right\}}\right] \leqslant e^{\frac{Mu^2C^2}{2}} ,$$

which implies

$$\mathbb{E}\left[e^{uM\left|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)-b_{\hat{\mu}_n,\alpha}(\theta_j)\right|}\right] \leqslant 2e^{\frac{Mu^2C^2}{2}} .$$

Then, combining with (2.35), we have

$$e^{u\mathbb{E}\left[\max_{1\leqslant j\leqslant J} M\left|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)-b_{\hat{\mu}_n,\alpha}(\theta_j)\right|\right]} \leqslant 2Je^{\frac{Mu^2C^2}{2}}$$

and we obtain

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant J} M\left|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j) - b_{\hat{\mu}_n,\alpha}(\theta_j)\right|\right] \leqslant \frac{\log(2J)}{u} + \frac{MuC^2}{2} .$$

Setting $u = \sqrt{\frac{2\log(2J)}{MC^2}}$ yields the desired result, as we have

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant J} \left|\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j) - b_{\hat{\mu}_n,\alpha}(\theta_j)\right|\right] \leqslant 2C\sqrt{\frac{\log(2J)}{2M}}$$

□

We now state in the next Theorem an $O(1/\sqrt{N}+O(1/\sqrt{M}))$ bound on $\mathbb{E}[\Psi_\alpha(\hat{\mu}_n k)-\Psi_\alpha(\mu^\star)]$ in the particular case of the Stochastic Entropic Mirror Descent.

Theorem 6. *Assume (2.A1). Let $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\hat{\mu}_1 k) < \infty$, let $(\hat{\mu}_n)_{n\in\mathbb{N}^\star}$ be defined by (2.16) and assume that (2.A5) holds. Define $|\hat{b}|_{\infty,\alpha} :=$ $\sup_{n\in\mathbb{N}^\star,\theta\in\mathsf{T}} |\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta)|$, assume that $|\hat{b}|_{\infty,\alpha} < \infty$ and let $\Gamma(v) = e^{-\eta v}$. Finally, let (η, κ) belong to any of the following cases.*

(i) *Forward Kullback-Leibler divergence* ($\alpha = 1$)*:* $\eta \in (0, 1)$ *and* κ *is any real number;*

(ii) *Reverse Kullback-Leibler* ($\alpha = 0$) *and* α*-Divergence with* $\alpha \in \mathbb{R} \setminus \{0, 1\}$*:* $\eta \in (0, \frac{1}{|\alpha-1||\hat{b}|_{\infty,\alpha}+1})$ *and* κ *is any real number;*

Then, the sequence $(\hat{\mu}_n)_{n\in\mathbb{N}^\star}$ *is well-defined and for all* $N \in \mathbb{N}^\star$*, we have*

$$\mathbb{E}\left[\Psi_\alpha\left(\frac{1}{N}\sum_{n=1}^{N}\hat{\mu}_n k\right) - \Psi_\alpha(\mu^\star k)\right] \leqslant \frac{1}{N\eta}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right] + \frac{1}{\sqrt{M}}\frac{LL_{\alpha,3}L_{\alpha,4}\sigma}{\eta L_{\alpha,1}} ,$$

where $\mu^\star$ *is such that* $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\hat{\mu}_1}(\mathsf{T})}\Psi_\alpha(\zeta k)$ *and where we have defined* $\Delta_1 = \Psi_\alpha(\hat{\mu}_1 k) - \Psi_\alpha(\mu^\star k)$*,* $KL(\mu^\star||\hat{\mu}_1) = \int_{\mathsf{T}}\log\left(\frac{\mathrm{d}\mu^\star}{\mathrm{d}\hat{\mu}_1}\right)\mathrm{d}\mu^\star$ *and*

$$L_{\alpha,4} := \sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v)^{\alpha-1} \times \sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v)^{1-\alpha}\left[1 + \frac{\sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v)}{\inf_{v\in\mathrm{Dom}_\alpha}\Gamma(v)}\right] .$$

The first step to prove this result is to see what becomes of Lemma 7 in the stochastic framework, which we investigate in Lemma 15 below.

Lemma 15. *Assume (2.A1) and (2.A2). Let* $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$ *be such that* $\Psi_\alpha(\hat{\mu}_1 k) < \infty$*, let* $(\hat{\mu}_n)_{n\in\mathbb{N}^\star}$ *be defined by (2.16) and assume that (2.A5) holds. Further assume that* $L_{\alpha,4} < \infty$*. Then, for all* $n \in \mathbb{N}^\star$*,*

$$\frac{L_{\alpha,1}}{2}\mathbb{E}\left[\mathbb{V}\mathrm{ar}_{\hat{\mu}_n}\left(\hat{b}_{\hat{\mu}_n,\alpha,M}\right)\right] \leqslant \mathbb{E}\left[\Psi_\alpha(\hat{\mu}_n k) - \Psi_\alpha(\hat{\mu}_{n+1}k)\right] + L_{\alpha,4}\frac{\sigma}{\sqrt{M}} . \tag{2.36}$$

Proof. We consider the case $\kappa = 0$ for simplicity. Set $\hat{g}_n(\theta) = \tilde{\Gamma}(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta))$ for all $\theta \in \mathsf{T}$ and for all $n \in \mathbb{N}^\star$, where $\tilde{\Gamma}(u) = \Gamma(u)/\mathbb{E}_{\hat{\mu}_n}[\Gamma]$. Based on the proof of Theorem 1, we have

$$A_\alpha \leqslant \Psi_\alpha(\hat{\mu}_n k) - \Psi_\alpha(\hat{\mu}_{n+1}k) , \tag{2.37}$$

where

$$A_\alpha = \begin{cases} \int_{\mathsf{T}}\hat{\mu}_n(\mathrm{d}\theta)\left[\log\hat{g}_n(\theta) + b_{\hat{\mu}_n,1}(\theta) + \kappa\right]\left[1 - \hat{g}_n(\theta)\right] & \text{if } \alpha = 1 \\ \int_{\mathsf{T}}\hat{\mu}_n(\mathrm{d}\theta)\left[b_{\hat{\mu}_n,\alpha}(\theta) + \frac{1}{\alpha-1}\right]\hat{g}_n(\theta)^{\alpha-1}\left[1 - \hat{g}_n(\theta)\right], & \text{otherwise.} \end{cases}$$

Now defining

$$E_\alpha = \hat{\mu}_n\left(\left[\hat{b}_{\hat{\mu}_n,\alpha,M} - b_{\hat{\mu}_n,\alpha}\right]\hat{g}_n^{\alpha-1}\left[1 - \hat{g}_n\right]\right)$$

and based on the proof of Lemma 7 we can rewrite (2.37) as

$$\frac{L_{\alpha,1}}{2}\mathbb{V}\mathrm{ar}_{\hat{\mu}_n}\left(\hat{b}_{\hat{\mu}_n,\alpha,M}\right) \leqslant \Psi_\alpha(\hat{\mu}_n k) - \Psi_\alpha(\hat{\mu}_{n+1}k) + E_\alpha$$

and we deduce

$$\begin{aligned}\mathbb{E}\left[E_\alpha\right] &= \mathbb{E}\left[\hat{\mu}_n\left(\left[\hat{b}_{\hat{\mu}_n,\alpha,M} - b_{\hat{\mu}_n,\alpha}\right]\hat{g}_n^{\alpha-1}\left[1-\hat{g}_n\right]\right)\right] \\ &\leqslant L_{\alpha,4}\mathbb{E}\left[\hat{\mu}_n\left(\mathbb{E}\left[\left|\hat{b}_{\hat{\mu}_n,\alpha,M} - b_{\hat{\mu}_n,\alpha}\right| |\mathcal{F}_n\right]\right)\right] \\ &\leqslant L_{\alpha,4}\cdot\frac{\sigma}{\sqrt{M}}\,.\end{aligned}$$

□

Next, we derive the stochastic version of Theorem 2 in the particular case of the Entropic Mirror Descent.

Theorem 7. *Assume (2.A1). Set $\Gamma(v) = e^{-\eta v}$ and let η be such that (2.A2) and (2.A3) hold. Let $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\hat{\mu}_1 k) < \infty$, let $(\hat{\mu}_n)_{n\in\mathbb{N}^\star}$ be defined by (2.16) and assume that (2.A5) holds. Further assume that $L_{\alpha,1}$, $L_{\alpha,2} > 0$ and that $0 < \inf_{v\in\mathrm{Dom}_\alpha}\Gamma(v) \leqslant \sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v) < \infty$. Then, for all $N\in\mathbb{N}^\star$, we have*

$$\mathbb{E}\left[\Psi_\alpha\left(\frac{1}{N}\sum_{n=1}^N\hat{\mu}_n k\right) - \Psi_\alpha(\mu^\star k)\right] \leqslant \frac{1}{N\eta}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right] + \frac{1}{\sqrt{M}}\frac{LL_{\alpha,3}L_{\alpha,4}\sigma}{\eta L_{\alpha,1}}\,.$$

Proof. We consider the case $\kappa = 0$ for simplicity. Let $n\in\mathbb{N}^\star$ and set $\Delta_n = \Psi_\alpha(\hat{\mu}_n k) - \Psi_\alpha(\mu^\star k)$. Then,

$$\begin{aligned}\mathbb{E}[\Delta_n] &\leqslant \mathbb{E}\left[\int_{\mathsf{T}} b_{\hat{\mu}_n,\alpha}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star)\right] \\ &= \mathbb{E}\left[\int_{\mathsf{T}}\mathbb{E}\left[\hat{b}_{\hat{\mu}_n,\alpha,M}|\mathcal{F}_n\right](\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star)\right] \\ &= \mathbb{E}\left[\int_{\mathsf{T}}\hat{b}_{\hat{\mu}_n,\alpha,M}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star)\right] \\ &= \mathbb{E}\left[\int_{\mathsf{T}}(\hat{\mu}_n(\hat{b}_{\hat{\mu}_n,\alpha,M}) - \hat{b}_{\hat{\mu}_n,\alpha,M})\mathrm{d}\mu^\star\right]\,.\end{aligned} \tag{2.38}$$

By adapting the proof of Theorem 2, we deduce

$$\mathbb{E}\left[\Delta_n\right] \leqslant \mathbb{E}\left[\frac{1}{(-\log\Gamma)'(\hat{\mu}_n(\hat{b}_{\hat{\mu}_n,\alpha,M}))}\int_{\mathsf{T}}\left[\log\Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}) - \log\Gamma(\hat{\mu}_n(\hat{b}_{\hat{\mu}_n,\alpha,M}))\right]\mathrm{d}\mu^\star\right]\,.$$

In the particular case of the Entropic Mirror Descent (for which $\Gamma(v) = e^{-\eta v}$) we obtain that $-(\log\Gamma)' = \eta$, that is $L_{\alpha,2} = \frac{1}{\eta}$ and by following the proof of Theorem 2 we have

$$\mathbb{E}\left[\Delta_n\right] \leqslant \frac{1}{\eta}\mathbb{E}\left[\int_{\mathsf{T}}\log\left(\frac{\mathrm{d}\hat{\mu}_{n+1}}{\mathrm{d}\hat{\mu}_n}\right)\mathrm{d}\mu^\star + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\hat{\mu}_n}\left(\hat{b}_{\hat{\mu}_n,\alpha,M}\right)L_{\alpha,3}\right]\,. \tag{2.39}$$

By assumption on $\inf_{v\in \mathrm{Dom}_\alpha} \Gamma(v)$ and on $\sup_{v\in \mathrm{Dom}_\alpha} \Gamma(v)$, we have that $L_{\alpha,4} < \infty$ and combining with Lemma 15, we obtain

$$\mathbb{E}\left[\Delta_n\right] \leqslant \frac{1}{\eta}\mathbb{E}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_{n+1}}{\mathrm{d}\hat{\mu}_n}\right)\mathrm{d}\mu^\star + \frac{LL_{\alpha,3}}{L_{\alpha,1}}\left[\Psi_\alpha(\hat{\mu}_n k) - \Psi_\alpha(\hat{\mu}_{n+1}k)\right]\right] + \frac{LL_{\alpha,3}L_{\alpha,4}}{\eta L_{\alpha,1}}\frac{\sigma}{\sqrt{M}} \,.$$

As the r.h.s involves two telescopic sums, we deduce

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\Delta_n\right] \leqslant \frac{1}{\eta N}\left[KL(\mu^\star||\hat{\mu}_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right] + \frac{LL_{\alpha,3}L_{\alpha,4}}{\eta L_{\alpha,1}}\frac{\sigma}{\sqrt{M}} \,,$$

and we conclude using the convexity of the mapping $\mu \mapsto \Psi_\alpha(\mu k)$. □

With all these elements in hand, we can now prove Theorem 6.

Proof of Theorem 6. The proof follows from a straightforward adaptation of the proof of Theorem 3 for the Entropic Mirror Descent (we replace $|b|_{\infty,\alpha}$ by $|\hat{b}|_{\infty,\alpha}$) combined with Theorem 6. □

Proof of Theorem 5

Proof of Theorem 5. The proof of Theorem 5 can be adapted from the proof of Nemirovski et al., 2009, Section 2.3. We consider the case $\kappa = 0$ for simplicity. Note that the case $\kappa \neq 0$ unfolds similarly by replacing $b_{\hat{\mu}_n,\alpha}$ by $b_{\hat{\mu}_n,\alpha} + \kappa$ everywhere in the proof below. Let $n \in \mathbb{N}^\star$ and set $\Delta_n = \Psi_\alpha(\hat{\mu}_n k) - \Psi_\alpha(\mu^\star k)$. The convexity of f_α implies that

$$\Delta_n \leqslant \int_{\mathsf{T}} b_{\hat{\mu}_n,\alpha}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star) \,.$$

Now taking the expectation, we obtain that

$$\begin{aligned}\mathbb{E}[\Delta_n] &\leqslant \mathbb{E}\left[\int_{\mathsf{T}} b_{\hat{\mu}_n,\alpha}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star)\right] \\ &= \mathbb{E}\left[\int_{\mathsf{T}} \mathbb{E}[\hat{b}_{\hat{\mu}_n,\alpha,M} \mid \mathcal{F}_n](\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star)\right] \\ &= \mathbb{E}\left[\int_{\mathsf{T}} \hat{b}_{\hat{\mu}_n,\alpha,M}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star)\right] \,.\end{aligned} \tag{2.40}$$

In addition, using that $\frac{\mathrm{d}\hat{\mu}_{n+1}}{\mathrm{d}\hat{\mu}_n} \propto e^{-\eta_n \hat{b}_{\hat{\mu}_n,\alpha,M}}$ and noting that the integral of any constant w.r.t $\hat{\mu}_n - \mu^\star$ is null, we deduce

$$\begin{aligned}\int_{\mathsf{T}} \hat{b}_{\hat{\mu}_n,\alpha,M}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star) &= \frac{1}{\eta_n}\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\hat{\mu}_{n+1}}\right)(\mathrm{d}\hat{\mu}_n - \mathrm{d}\mu^\star) \\ &= \frac{1}{\eta_n}\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\hat{\mu}_{n+1}}\right)\mathrm{d}\hat{\mu}_n - \frac{1}{\eta_n}\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\hat{\mu}_{n+1}}\right)\mathrm{d}\mu^\star \\ &= \frac{1}{\eta_n}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\hat{\mu}_{n+1}}\right)(\mathrm{d}\hat{\mu}_n - \mathrm{d}\hat{\mu}_{n+1}) - KL(\hat{\mu}_{n+1}||\hat{\mu}_n)\right] \\ &\quad + \frac{1}{\eta_n}\left[KL(\mu^\star||\hat{\mu}_n) - KL(\mu^\star||\hat{\mu}_{n+1})\right]\end{aligned}$$

Let us first consider the term inside the first brackets. We have that

$$\begin{aligned}\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\hat{\mu}_{n+1}}\right)(\mathrm{d}\hat{\mu}_n - \mathrm{d}\hat{\mu}_{n+1}) &= \eta_n \int_{\mathsf{T}} \hat{b}_{\hat{\mu}_n,\alpha,M}(\mathrm{d}\hat{\mu}_n - \mathrm{d}\hat{\mu}_{n+1}) \\ &\leqslant \eta_n |b|_{\hat{\mu}_n,M,\alpha} \left\|\hat{\mu}_n - \hat{\mu}_{n+1}\right\|_{TV} ,\end{aligned}$$

where we have set

$$|b|_{\hat{\mu}_n,M,\alpha} = \frac{1}{M}\sum_{m=1}^{M} \sup_{\theta\in\mathsf{T}} \frac{k(\theta, Y_{m,n+1})}{\hat{\mu}_n k(Y_{m,n+1})}\left|f'_\alpha\left(\frac{\hat{\mu}_n k(Y_{m,n+1})}{p(Y_{m,n+1})}\right)\right|$$

and where we have used that $\frac{\mathrm{d}\hat{\mu}_{n+1}}{\mathrm{d}\hat{\mu}_n} \propto e^{-\eta_n \hat{b}_{\hat{\mu}_n,\alpha,M}}$ and that the integral of any constant w.r.t $\hat{\mu}_n - \hat{\mu}_{n+1}$ is null. Moreover, Pinsker's inequality yields

$$-KL(\hat{\mu}_{n+1}||\hat{\mu}_n) \leqslant -\frac{1}{2}\left\|\hat{\mu}_n - \hat{\mu}_{n+1}\right\|_{TV}^2 .$$

Now combining with the fact that $\eta_n|b|_{\hat{\mu}_n,M,\alpha}a - a^2/2 \leqslant (\eta_n|b|_{\hat{\mu}_n,M,\alpha})^2/2$ which is valid for all $a \geqslant 0$, we get:

$$\frac{1}{\eta_n}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\hat{\mu}_n}{\mathrm{d}\hat{\mu}_{n+1}}\right)(\mathrm{d}\hat{\mu}_n - \mathrm{d}\hat{\mu}_{n+1}) - KL(\hat{\mu}_{n+1}||\hat{\mu}_n)\right] \leqslant \frac{\eta_n |b|^2_{\hat{\mu}_n,M,\alpha}}{2}. \tag{2.41}$$

Furthermore, using Jensen's inequality and (2.18), we have that

$$\begin{aligned}\mathbb{E}\left[|b|^2_{\hat{\mu}_n,M,\alpha}\right] &\leqslant \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left(\sup_{\theta\in\mathsf{T}} \frac{k(\theta, Y_{m,n+1})}{\hat{\mu}_n k(Y_{m,n+1})}\left|f'_\alpha\left(\frac{\hat{\mu}_n k(Y_{m,n+1})}{p(Y_{m,n+1})}\right)\right|\right)^2\right] \\ &\leqslant B_\alpha^2\end{aligned}$$

and as a consequence, we obtain from (2.40) and (2.41) that

$$\eta_n \mathbb{E}[\Delta_n] \leqslant \eta_n^2 B_\alpha^2/2 + \mathbb{E}[(KL(\mu^\star||\hat{\mu}_n) - KL(\mu^\star||\hat{\mu}_{n+1}))] .$$

Finally, as we recognize a telescoping sum in the right-hand side, we have

$$\sum_{n=1}^{N} \eta_n \mathbb{E}[\Delta_n] \leqslant \sum_{n=1}^{N} \eta_n^2 B_\alpha^2/2 + KL(\mu^\star||\hat{\mu}_1)$$

that is we have, by convexity of the mapping $\mu \mapsto \Psi_\alpha(\mu k)$,

$$\mathbb{E}\left[\Psi_\alpha\left(\sum_{n=1}^{N} w_n \hat{\mu}_n k\right) - \Psi_\alpha(\mu^\star k)\right] \leqslant \frac{B_\alpha^2 \sum_{n=1}^{N} \eta_n^2/2}{\sum_{n=1}^{N} \eta_n} + \frac{KL(\mu^\star||\hat{\mu}_1)}{\sum_{n=1}^{N} \eta_n} . \tag{2.42}$$

Then,

- setting $\eta_n = \eta_0/\sqrt{n}$ for all $n \geqslant 1$ in (2.42) yields

$$\mathbb{E}\left[\Psi_\alpha\left(\sum_{n=1}^{N} w_n \hat{\mu}_n k\right) - \Psi_\alpha(\mu^\star k)\right] \leqslant \frac{(1+\log(N))B_\alpha^2 \eta_0^2/2 + KL(\mu^\star||\hat{\mu}_1)}{\eta_0\sqrt{N}} .$$

- setting $\eta_n = \eta_0/\sqrt{N}$ for all $n = 1 \dots N$ in (2.42) yields

$$\mathbb{E}\left[\Psi_\alpha\left(\frac{1}{N}\sum_{n=1}^{N} \hat{\mu}_n k\right) - \Psi_\alpha(\mu^\star k)\right] \leqslant \frac{B_\alpha^2 \eta_0^2/2 + KL(\mu^\star||\hat{\mu}_1)}{\eta_0\sqrt{N}} .$$

  Furthermore, the r.h.s is minimal for $\eta_0 = B_\alpha^{-1}\sqrt{2KL(\mu^\star||\hat{\mu}_1)}$ that is for $\eta_n = B_\alpha^{-1}\sqrt{\frac{2KL(\mu^\star||\hat{\mu}_1)}{N}}$ for all $n = 1 \dots N$.

□

Example 5 and Condition (2.18)

Proof that Condition (2.18) is satisfied in Example 5.

We have $k_h(\theta, y) = \frac{e^{-\|y-\theta\|^2/(2h^2)}}{(2\pi h^2)^{d/2}}$ and $p(y) = Z \times \left[0.5\frac{e^{-\|y-\theta_1^\star\|^2/2}}{(2\pi)^{d/2}} + 0.5\frac{e^{-\|y-\theta_2^\star\|^2/2}}{(2\pi)^{d/2}}\right]$ for all $\theta \in \mathsf{T}$ and all $y \in \mathsf{Y}$. Since we have chosen $\alpha = 1$, we have $f'_\alpha(u) = \log(u)$ for all $u > 0$ and we are interested in the following quantity

$$B_1^2 := \sup_{\mu \in \mathrm{M}_1(\mathsf{T})} \int_{\mathsf{Y}} \sup_{\theta,\theta' \in \mathsf{T}} \frac{k_h(\theta, y)^2}{k_h(\theta', y)} \left|\log\left(\frac{\mu k_h(y)}{p(y)}\right)\right|^2 \nu(\mathrm{d}y) .$$

For simplicity, we consider the case $Z = 1$. Recall that by assumption $\mathsf{T} = \mathcal{B}(0, r)$. Then, for all $\theta, \theta' \in \mathsf{T}$ and for all $y \in \mathsf{Y}$, we can write

$$\begin{aligned}\frac{k_h(\theta, y)}{k_h(\theta', y)} &= e^{\frac{-\|y-\theta\|^2+\|y-\theta'\|^2}{2h^2}} = e^{\frac{2<y,\theta-\theta'>-\|\theta\|^2+\|\theta'\|^2}{2h^2}} \\ &\leqslant e^{\frac{2|<y,\theta-\theta'>|+\|\theta\|^2+\|\theta'\|^2}{2h^2}} \\ &\leqslant e^{\frac{\|y\|\|\theta-\theta'\|+r^2}{h^2}} \\ &\leqslant e^{\frac{\|y\|2r+r^2}{h^2}}\end{aligned}$$

Furthermore, we also have for all $y \in \mathsf{Y}$

$$\begin{aligned}e^{-\sup_{\theta\in\mathsf{T}} \frac{\|y-\theta\|^2}{2h^2}} \leqslant (2\pi h^2)^{d/2} \mu k_h(y) \leqslant 1 \\ e^{-\max_{i\in\{1,2\}} \frac{\|y-\theta_i^\star\|^2}{2}} \leqslant (2\pi)^{d/2} p(y) \leqslant 1\end{aligned}$$

and we can deduce for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $y \in \mathsf{Y}$

$$\begin{aligned}\left|\log\left(\frac{\mu k_h(y)}{p(y)}\right)\right| &\leqslant \sup_{\theta\in\mathsf{T}} \frac{\|y-\theta\|^2}{2h^2} + \max_{i\in\{1,2\}} \frac{\|y-\theta_i^\star\|^2}{2} + d|\log h| \\ &\leqslant \frac{(\|y\|+r)^2}{2}\left[\frac{1}{h^2}+1\right] + d|\log h| \,.\end{aligned}$$

Consequently, we have

$$B_1^2 \leqslant \int_{\mathsf{Y}} \frac{e^{\frac{\|y\|2r+r^2}{h^2}}}{(2\pi h^2)^{d/2}} \underbrace{\sup_{\theta\in\mathsf{T}} e^{-\|y-\theta\|^2/(2h^2)}}_{\leqslant e^{-(\|y\|-r)_+^2/(2h^2)}} \left(\frac{(\|y\|+r)^2}{2}\left[\frac{1}{h^2}+1\right] + d|\log h|\right)^2 \nu(\mathrm{d}y)$$

that is $B_1 < \infty$. □

2.A.4 Lemma 16 : statement and proof

Recall that $Y_1, Y_2, \ldots$ are i.i.d random variables with common density μk w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we denote by $\mathbb{E}$ the associated expectation operator. Here, Γ is chosen as $\Gamma(v) = [(\alpha-1)\, v + 1]^{\eta/(1-\alpha)}$.

Lemma 16. *Assume (2.A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, $\eta > 0$ and κ be such that $(\alpha-1)\kappa \geqslant 0$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\mu(|b_{\mu,\alpha}|) < \infty$ and*

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta)\mathbb{E}\left[\left\{\frac{k(\theta, Y_1)}{\mu k(Y_1)}\left(\frac{\mu k(Y_1)}{p(Y_1)}\right)^{\alpha-1} + (\alpha-1)\kappa\right\}^{\frac{\eta}{1-\alpha}}\right] < \infty \,. \tag{2.43}$$

Then,

$$\lim_{M\to\infty} \mu(\Gamma(\hat{b}_{\mu,\alpha,M} + \kappa)) = \mu(\Gamma(b_{\mu,\alpha} + \kappa)) \,, \quad \mathbb{P}-\text{a.s.} \tag{2.44}$$

Proof. Set $g(\theta, y) = \frac{k(\theta,y)}{\mu k(y)}(\frac{\mu k(y)}{p(y)})^{\alpha-1} + (\alpha-1)\kappa$, $\phi = \frac{\eta}{1-\alpha}$ and $h(u) = (\alpha-1)u + (\alpha-1)\kappa + 1$. Note that $\mathbb{E}[g(\theta, Y_1)] = h(b_{\mu,\alpha}(\theta))$ and $h^\phi = \Gamma$.

(i) We start with the case $\phi \notin [0,1]$. Our goal is to apply Lemma 17, which is a generalized version of the Dominated Convergence Theorem. To do so, first note that $h(\hat{b}_{\mu,\alpha,M}(\theta))^\phi$ is positive and combining with the convexity of the mapping $u \mapsto u^\phi$, we have for all $M \in \mathbb{N}^\star$ and for all $\theta \in \mathsf{T}$,

$$0 \leqslant h(\hat{b}_{\mu,\alpha,M}(\theta))^\phi \leqslant M^{-1} \sum_{m=1}^{M} [g(\theta, Y_m)]^\phi \; . \tag{2.45}$$

Since $\mu(|b_{\mu,\alpha}|) < \infty$, the LLN for μ-almost all $\theta \in \mathsf{T}$ yields

$$\lim_{M\to\infty} \hat{b}_{\mu,\alpha,M}(\theta) = b_{\mu,\alpha}(\theta) \; . \tag{2.46}$$

Now applying successively (a) the LLN for μ-almost all $\theta \in \mathsf{T}$ (as stated in Lemma 18), which is valid under (2.43), (b) Fubini's Theorem and (c) again the LLN

$$\begin{aligned} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} \{g(\theta, Y_m)\}^\phi &\overset{(a)}{=} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}\left[\{g(\theta, Y_1)\}^\phi\right] \\ &\overset{(b)}{=} \mathbb{E}\left[\int_{\mathsf{T}} \mu(\mathrm{d}\theta)[g(\theta, Y_1)]^\phi\right] \overset{(c)}{=} \lim_{M\to\infty} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) M^{-1} \sum_{m=1}^{M} [g(\theta, Y_m)]^\phi \end{aligned} \tag{2.47}$$

that is

$$\mu\left(\lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} \{g(\cdot, Y_m)\}^\phi\right) = \lim_{M\to\infty} \mu\left(M^{-1} \sum_{m=1}^{M} [g(\cdot, Y_m)]^\phi\right) < \infty \; .$$

Combining with (2.45) and (2.46), we apply Lemma 17 and obtain

$$\mu\left(h(b_{\mu,\alpha})^\phi\right) = \mu\left(\lim_{M\to\infty} h(\hat{b}_{\mu,\alpha,M})^\phi\right) = \lim_{M\to\infty} \mu(h(\hat{b}_{\mu,\alpha,M})^\phi) \; ,$$

that is

$$\mu\left(\Gamma(b_{\mu,\alpha} + \kappa)\right) = \lim_{M\to\infty} \mu(\Gamma(\hat{b}_{\mu,\alpha,M} + \kappa)) \; .$$

(ii) We now turn to the case $\phi \in (0,1]$. Let $M' > 0$. Since

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(M^{-1} \sum_{m=1}^{M} g(\theta, Y_m) \mathbf{1}_{\{g(\theta,Y_m)\leqslant M'\}}\right)^\phi \leqslant \mu(h(\hat{b}_{\mu,\alpha,M})^\phi) \; ,$$

the LLN for μ-almost all $\theta \in \mathsf{T}$ (Lemma 18) and the Dominated Convergence Theorem yields

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\mathbb{E}[g(\theta, Y_1)\mathbf{1}_{\{g(\theta,Y_1)\leqslant M'\}}]\right)^\phi \leqslant \liminf_{M\to\infty} \mu(h(\hat{b}_{\mu,\alpha,M})^\phi) \; . \tag{2.48}$$

Using now $(u+v)^\phi \leqslant u^\phi + v^\phi$ and then Jensen's inequality for the concave mapping $u \mapsto u^\phi$,

$$\mu(h(\hat{b}_{\mu,\alpha,M})^\phi) \leqslant \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(M^{-1} \sum_{m=1}^{M} g(\theta, Y_m) \mathbf{1}_{\{g(\theta,Y_m) \leqslant M'\}} \right)^\phi + \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) M^{-1} \sum_{m=1}^{M} g(\theta, Y_m) \mathbf{1}_{\{g(\theta,Y_m) > M'\}} \right)^\phi$$

By invoking the LLN for μ-almost all $\theta \in \mathsf{T}$ (Lemma 18) and the Dominated Convergence Theorem for the first term of the rhs and the LLN combined with Fubini for the second term, we get

$$\limsup_{M\to\infty} \mu(h(\hat{b}_{\mu,\alpha,M})^\phi) \leqslant \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\mathbb{E}[g(\theta, Y_1) \mathbf{1}_{\{g(\theta,Y_1) \leqslant M'\}}] \right)^\phi + \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}[g(\theta, Y_1) \mathbf{1}_{\{g(\theta,Y_1) > M'\}}] \right)^\phi$$

Letting M' go to infinity both in this inequality and in (2.48) completes the proof of (2.44).

□

2.A.5 General Dominated Convergence Theorem

We state and prove a generalized version of the Dominated Convergence Theorem, adapted from Royden and Fitzpatrick, 2010, Theorem 19. We provide here a full proof for the sake of completeness.

Lemma 17 (General Dominated Convergence Theorem). *Let $\zeta \in \mathrm{M}_1(\mathsf{T})$. Assume there exist (a_M), (b_M), (c_M) three sequences of $(\mathcal{T}, \mathcal{B}(\mathbb{R}))$-measurable functions such that the limits $\lim_{M\to\infty} a_M(\theta)$, $\lim_{M\to\infty} b_M(\theta)$, $\lim_{M\to\infty} c_M(\theta)$ exist for ζ-almost all $\theta \in \mathsf{T}$ and*

$$\zeta|\lim_{M\to\infty} a_M| + \zeta|\lim_{M\to\infty} c_M| < \infty \,.$$

Assume moreover that for all $M \in \mathbb{N}^\star$ and for ζ-almost all $\theta \in \mathsf{T}$

$$a_M(\theta) \leqslant b_M(\theta) \leqslant c_M(\theta)$$

and

$$\zeta(\lim_{M\to\infty} a_M) = \lim_{M\to\infty} \zeta(a_M) \tag{2.49}$$

$$\zeta(\lim_{M\to\infty} c_M) = \lim_{M\to\infty} \zeta(c_M) \,. \tag{2.50}$$

Then,

$$\zeta(\lim_{M\to\infty} b_M) = \lim_{M\to\infty} \zeta(b_M)\ .$$

Proof. We apply Fatou's Lemma combined with (2.49) and (2.50) to the two non-negative, $(\mathcal{T}, \mathcal{B}(\mathbb{R}))$-measurable functions $\theta \mapsto b_M(\theta) - a_M(\theta)$ and $\theta \mapsto c_M(\theta) - b_M(\theta)$ and we obtain

$$\zeta(\liminf_{M\to\infty} b_M) \leqslant \liminf_{M\to\infty} \zeta(b_M)$$
$$\zeta(\liminf_{M\to\infty} -b_M) \leqslant \liminf_{M\to\infty} \zeta(-b_M)$$

which proves the lemma, as $\liminf_{M\to\infty} b_M(\theta) = \limsup_{M\to\infty} b_M(\theta)$ for ζ-almost all $\theta \in \mathsf{T}$. □

2.A.6 Integrated Law of Large Numbers

Let $Y_1, Y_2, \ldots$ be i.i.d. random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let f be a non-negative real-valued $(\mathcal{T} \otimes \mathcal{F}, \mathcal{B}(\mathbb{R}_{\geqslant 0}))$-measurable function. We are interested in showing

$$\int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m) = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] \tag{2.51}$$

for $\zeta \in \mathrm{M}_1(\mathsf{T})$ satisfying $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] < \infty$. While this result follows easily if we can show that

$$\mathbb{P}\left(\forall \theta \in \mathsf{T},\ \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m) = \mathbb{E}[f(\theta, Y_1)]\right) = 1 \tag{2.52}$$

unfortunately the LLN only yields

$$\mathbb{P}\left(\lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m) = \mathbb{E}[f(\theta, Y_1)]\right) = 1$$

for ζ-almost all $\theta \in \mathsf{T}$. The following lemma allows to show (2.51) without resorting to the much stronger identity (2.52).

Lemma 18. *Let $\zeta \in \mathrm{M}_1(\mathsf{T})$ and assume that $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] < \infty$. Then, $\mathbb{P}$ − a.s.*

$$\int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m) = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)]\ .$$

Proof. Set

$$B = \left\{(\theta, \omega) \in \mathsf{T} \times \Omega\ :\ \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m(\omega)) = \mathbb{E}[f(\theta, Y_1)]\right\}\ .$$

Let $\gamma_0 : (\theta, \omega) \mapsto \mathbf{1}_{B^c}(\theta, \omega)$ and $\gamma_1 = 1 - \gamma_0$. According to the Fubini Theorem and the LLN for $M^{-1} \sum_{m=1}^{M} f(\theta, Y_m)$ where θ is such that $\mathbb{E}[f(\theta, Y_1)] < \infty$ (which is satisfied for ζ-almost all $\theta \in \mathsf{T}$ by assumption),

$$\mathbb{E}\left[\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\gamma_0(\theta, \cdot)\right] = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}\left[\gamma_0(\theta, \cdot)\right] = 0 \,.$$

Therefore, $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\gamma_0(\theta, \cdot)$ is $\mathbb{P} - \text{a.s.}$ null that is, there exists Ω_1 such that $\mathbb{P}(\Omega_1) = 1$ and for all $\omega \in \Omega_1$, $A \mapsto \int_A \zeta(\mathrm{d}\theta)\gamma_0(\theta, \omega)$ is the null-measure on $(\mathsf{T}, \mathcal{T})$, which in turn implies that the measures ζ and $A \mapsto \int_A \zeta(\mathrm{d}\theta)\gamma_1(\theta, \omega)$ coincide. The latter property implies for all $\omega \in \Omega_1$,

$$\begin{aligned}
\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] &= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)]\gamma_1(\theta, \omega) \\
&= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \left[\lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m(\omega))\right] \gamma_1(\theta, \omega) \\
&= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m(\omega)) \,.
\end{aligned}$$

□

2.A.7 Proof of Proposition 10

Proof of Proposition 10. Recall that we have taken $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$. For the sake of readability, we only treat the case $\kappa = 0$ in the proof of Proposition 10. Note that the case $\kappa \neq 0$ unfolds similarly by replacing $b_{\mu,\alpha}$ by $b_{\mu,\alpha} + \kappa$ everywhere in the proof below.

A first remark is that $\Psi_\alpha(\mu k) < \infty$ implies $\mu(|b_{\mu,\alpha}|) < \infty$. This comes from the fact that for all $\alpha \in \mathbb{R}$ and for all $u \in \mathbb{R}_{>0}$, $u f'_\alpha(u) = \alpha f_\alpha(u) + (u - 1)$ and we can write

$$\mu(|b_{\mu,\alpha}|) \leqslant |\alpha| \int_{\mathsf{Y}} \left| f_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) + \int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) + 1 \,.$$

Under (2.A1), we have $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$, which settles the case $\alpha = 0$. As for the case $\alpha \in \mathbb{R} \setminus \{0\}$, we obtain from Lemma 4 that the r.h.s is finite if and only if $\Psi_\alpha(\mu k)$ is finite, which is implied by the assumption $\Psi_\alpha(\mu k) < \infty$.

By the triangular inequality, for all $M \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,

$$\begin{aligned}
\left|\frac{\Gamma(\hat{b}_{\mu,\alpha,M}(\theta))}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))} - \frac{\Gamma(b_{\mu,\alpha}(\theta))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| &\leqslant \frac{\Gamma(\hat{b}_{\mu,\alpha,M}(\theta))}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))} \left|1 - \frac{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| \\
&\quad + \frac{|\Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) - \Gamma(b_{\mu,\alpha}(\theta))|}{\mu(\Gamma(b_{\mu,\alpha}))}
\end{aligned}$$

Thus,

$$\left\|\hat{\mathcal{I}}_{\alpha,M}(\mu) - \mathcal{I}_\alpha(\mu)\right\|_{TV} = \mu\left(\left|\frac{\Gamma(\hat{b}_{\mu,\alpha,M})}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))} - \frac{\Gamma(b_{\mu,\alpha})}{\mu(\Gamma(b_{\mu,\alpha}))}\right|\right)$$
$$\leqslant \left|1 - \frac{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| + \frac{\mu(|\Gamma(\hat{b}_{\mu,\alpha,M}) - \Gamma(b_{\mu,\alpha})|)}{\mu(\Gamma(b_{\mu,\alpha}))}$$

For the first term of the rhs, Lemma 16 yields

$$\lim_{M\to\infty}\left|1 - \frac{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| = 0\ . \tag{2.53}$$

As for the second term of the rhs, first note that for all $M \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$

$$0 \leqslant |\Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) - \Gamma(b_{\mu,\alpha}(\theta))| \leqslant \Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) + \Gamma(b_{\mu,\alpha}(\theta))\ , \tag{2.54}$$

and since $\mu(\Gamma(b_{\mu,\alpha})) < \infty$ by assumption, the LLN for μ-almost all $\theta \in \mathsf{T}$ yields

$$\lim_{M\to\infty} \Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) = \Gamma(b_{\mu,\alpha}(\theta))\ . \tag{2.55}$$

Furthermore, since $\mu(\Gamma(b_{\mu,\alpha})) < \infty$, Lemma 16 and (2.55) imply

$$\lim_{M\to\infty} \mu\left[\Gamma(\hat{b}_{\mu,\alpha,M}) + \Gamma(b_{\mu,\alpha})\right] = \mu\left[\lim_{M\to\infty}\left(\Gamma(\hat{b}_{\mu,\alpha,M}) + \Gamma(b_{\mu,\alpha})\right)\right] < \infty$$

Combining with (2.54) and (2.55), we apply Lemma 17 and obtain

$$\lim_{M\to\infty} \frac{\mu(|\Gamma(\hat{b}_{\mu,\alpha,M}) - \Gamma(b_{\mu,\alpha})|)}{\mu(\Gamma(b_{\mu,\alpha}))} = 0$$

which, along with (2.53), finishes the proof. □

3

Mixture weights optimisation with the α-divergence

The work presented in this chapter corresponds to the paper entitled "Mixture weights optimisation for Alpha-divergence Variational Inference" (Daudel and Douc, 2021) that has been submitted as a conference paper at the time of writing.

3.1 Introduction

Chapter 2 introduced the (α, Γ)-descent, a general family of gradient-based algorithms that are able to optimise the mixture weights of a given mixture model by α-divergence minimisation, without any information on the underlying distribution of its mixture components parameters.

The benefit of these types of algorithms is that they allow to select the mixture components according to their overall importance in the set of components parameters and from there, one is able to optimise the weights and the components parameters alternatively.

The (α, Γ)-descent framework recovers the Entropic Mirror Descent algorithm (corresponding to $\Gamma(v) = e^{-\eta v}$ with $\eta > 0$) and includes the Power Descent, an algorithm defined for all $\alpha \in \mathbb{R}\backslash\{1\}$ and all $\eta > 0$ that sets $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$. Although these two algorithms are linked to one another from a theoretical perspective through the (α, Γ)-descent framework, numerical experiments in Chapter 2 showed that the Power Descent outperforms the Entropic Mirror Descent when $\alpha < 1$ as the dimension increases.

However, the global convergence of the Power Descent algorithm when $\alpha < 1$, as stated in Chapter 2, is subjected to the condition that the limit exists. Furthermore,

even though the convergence towards the global optimum is derived, there is no convergence rate available for the Power Descent when $\alpha < 1$.

While there is no general rule yet on how to select the value of α in practice, the case $\alpha < 1$ has the advantage that it enforces a *mass-covering* property, as opposed to the *mode-seeking* property exhibited when $\alpha \geqslant 1$ (Chapter 1 and 2) and which often may lead to posterior variance underestimation.

We are thus interested in studying Variational Inference methods for mixture weights optimisation via α-divergence minimisation when $\alpha < 1$. To do so, let us introduce some notation and state the problem we aim at solving in this chapter.

Notation and problem statement We retain the notation from Chapter 2. Letting p be any measurable positive function on $(\mathsf{Y}, \mathcal{Y})$, the optimisation problem we consider is then

$$\operatorname{arginf}_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; p) \ , \tag{3.1}$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$,

$$\Psi_\alpha(\mu k; p) = \int_{\mathsf{Y}} f_\alpha \left(\frac{\mu k(y)}{p(y)} \right) p(y) \nu(\mathrm{d}y)$$

and where we will be particularly interested in the case where $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$, $J \in \mathbb{N}^\star$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J) \in \mathcal{S}_J$ and $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$. We will yet again drop the dependency on p in the rest of the chapter and we now detail the organisation of Chapter 3.

Outline The chapter is organised as follows:

- In Section 3.2, we recall for clarity the basics of the Power Descent algorithm described in Chapter 2 alongside with the convergence result obtained in Chapter 2 when $\alpha < 1$.
- In Section 3.3, we derive the full convergence proof of the Power Descent algorithm towards the optimal mixture weights when $\alpha < 1$ (Theorem 10).
- Since the α-divergence becomes the traditional forward Kullback-Leibler when $\alpha \to 1$, we first bridge in Section 3.4 the gap between the two cases $\alpha < 1$ and $\alpha > 1$ of the Power Descent: we obtain that the extension of the Power Descent to the case $\alpha = 1$ is an Entropic Mirror Descent performing forward Kullback-Leibler minimisation (Proposition 19). We then keep on investigating the connections between the Power Descent and the Entropic Mirror Descent by considering first-order approximations. In doing so, we are able to go beyond the (α, Γ)-descent framework and to introduce an algorithm closely-related to the Power Descent. We call it *Renyi Descent* and we prove in Theorem 11 that it converges at an $O(1/N)$ rate towards its optimum for all $\alpha \in \mathbb{R}$.

- Finally, we run some numerical experiments in Section 3.5 to compare the behavior of the Power Descent and the Renyi Descent altogether. We conclude by discussing the potential benefits of one approach over the other.

3.2 Background on the Power Descent

The optimisation problem (3.1) can be solved for all $\alpha \in \mathbb{R} \setminus \{1\}$ by using the *Power Descent* algorithm introduced in Chapter 2 : given an initial measure $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi_\alpha(\mu_1 k) < \infty$, $\alpha \in \mathbb{R} \setminus \{1\}$, $\eta > 0$ and κ such that $(\alpha - 1)\kappa \geqslant 0$, the Power descent algorithm is an iterative scheme which builds the sequence of probability measures $(\mu_n)_{n \in \mathbb{N}^\star}$

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \;, \qquad n \in \mathbb{N}^\star \;, \tag{3.2}$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$, the one-step transition $\mu \mapsto \mathcal{I}_\alpha(\mu)$ is given by Algorithm 8 and where for all $v \in \mathrm{Dom}_\alpha$, $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ [and Dom_α denotes an interval of $\mathbb{R}$ such that for all $\theta \in \mathsf{T}$, all $\mu \in \mathrm{M}_1(\mathsf{T})$, $b_{\mu,\alpha}(\theta) + \kappa$ and $\mu(b_{\mu,\alpha}) + \kappa \in \mathrm{Dom}_\alpha$].

Algorithm 8: *Exact Power Descent transition;* $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$

1. Expectation step : $b_{\mu,\alpha}(\theta) = \displaystyle\int_{\mathsf{Y}} k(\theta, y) f'_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y)$
2. Iteration step : $\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \dfrac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))}$

A remarkable property of the Power Descent algorithm, which has been proven in Chapter 2 (it is a special case of Theorem 1 with $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$), is that under (3.A1) as defined below

(3.A1) The density kernel k on $\mathsf{T} \times \mathsf{Y}$, the function p on Y and the σ-finite measure ν on $(\mathsf{Y}, \mathcal{Y})$ satisfy, for all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

the Power Descent ensures a monotonic decrease in the α-divergence at each step for all $\eta \in (0, 1]$.

Theorem 8 (Theorem 1 applied to the Power Descent)**.** *Assume that p and k are as in (3.A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, set $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ for all $v \in \mathrm{Dom}_\alpha$, let κ be such that $(\alpha - 1)\kappa \geqslant 0$, let $\mu \in \mathrm{M}_1(\mathsf{T})$ and let $\eta \in (0, 1]$ be such that*

$$0 < \mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty \tag{3.3}$$

holds and $\Psi_\alpha(\mu k) < \infty$. Then, the two following assertions hold.

(i) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$.

(ii) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Under the additional assumptions that $\kappa > 0$ and

$$\sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}| < \infty \quad \text{and} \quad \Psi_\alpha(\mu_1 k) < \infty \,, \tag{3.4}$$

the Power Descent is also known to converge towards its optimal value at an $O(1/N)$ rate when $\alpha > 1$ (Theorem 3). On the other hand, when $\alpha < 1$, the convergence towards the optimum as written in Theorem 4 of Chapter 2 holds under different assumptions including

(3.A2) (i) T is a compact metric space and $\mathcal{T}$ is the associated Borel σ-field;

(ii) for all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;

(iii) we have $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} k(\theta, y) \times \sup_{\theta' \in \mathsf{T}} \left(\frac{k(\theta', y)}{p(y)} \right)^{\alpha - 1} \nu(\mathrm{d}y) < \infty$.

If $\alpha = 0$, assume in addition that $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} \left| \log \left(\frac{k(\theta, y)}{p(y)} \right) \right| p(y) \nu(\mathrm{d}y) < \infty$.

so that Theorem 4, that is recalled below under the form of Theorem 9, states the convergence of the Power Descent algorithm towards the global optimum.

Theorem 9 (Recalling Theorem 4)**.** *Assume (3.A1) and (3.A2). Let $\alpha < 1$ and let $\kappa \leqslant 0$. Then, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $\Psi_\alpha(\mu k) < \infty$ and any $\eta > 0$ satisfies $0 < \mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Further assume that $\eta \in (0, 1]$ and that there exist $\mu_1, \mu^\star \in \mathrm{M}_1(\mathsf{T})$ such that the (well-defined) sequence $(\mu_n)_{n \in \mathbb{N}^\star}$ defined by (3.2) weakly converges to $\mu^\star$ as $n \to \infty$. Finally, denote by $\mathrm{M}_{1,\mu_1}(\mathsf{T})$ the set of probability measures dominated by μ_1. Then the following assertions hold*

(i) $(\Psi_\alpha(\mu_n k))_{n \in \mathbb{N}^\star}$ *is nonincreasing,*

(ii) $\mu^\star$ *is a fixed point of* $\mathcal{I}_\alpha$*,*

(iii) $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$.

The above result assumes there must exist $\mu_1, \mu^\star \in \mathrm{M}_1(\mathsf{T})$ such that the sequence $(\mu_n)_{n \in \mathbb{N}^\star}$ defined by (3.2) weakly converges to $\mu^\star$ as $n \to \infty$, that is it assumes the limit already exists. Our first contribution consists in showing that this assumption can be alleviated when μ is chosen a weighted sum of Dirac measures, that is when we seek to perform mixture weights optimisation by α-divergence minimisation.

3.3 Convergence of the Power Descent algorithm in the mixture case

Before we state our convergence result, let us first make two comments on the assumptions from Theorem 9 that shall be retained in our upcoming convergence result.

A first comment is that (3.A1) is mild since the assumption that $p(y) > 0$ for all $y \in \mathsf{Y}$ can be discarded and is kept for convenience (see Remark 5 of Chapter 2). A

second comment is that (3.A2) is also mild and covers (3.4) as it amounts to assuming that $b_{\mu,\alpha}(\theta)$ and $\Psi_\alpha(\mu k)$ are uniformly bounded with respect to μ and θ. To see this, we give below an example for which (3.A2) is satisfied.

Example 6. *Consider the case* $\mathsf{Y} = \mathbb{R}^d$ *with* $\alpha \in [0,1)$. *Let* $r > 0$ *and let* $\mathsf{T} = \mathcal{B}(0,r) \subset \mathbb{R}^d$. *Furtheremore, let* K_h *be a Gaussian transition kernel with bandwidth* h *and denote by* k_h *its associated kernel density. Finally, let* p *be a mixture density of two* d*-dimensional Gaussian distributions multiplied by a positive constant* c *such that for all* $y \in \mathsf{Y}$, $p(y) = c \times [0.5\mathcal{N}(y;\theta_1^\star, \boldsymbol{I_d}) + 0.5\mathcal{N}(y;\theta_2^\star, \boldsymbol{I_d})]$ *where* $\theta_1^\star, \theta_2^\star \in \mathsf{T}$ *and* $\boldsymbol{I_d}$ *is the identity matrix. Then, (3.A2) holds (see Section 3.A.1).*

Next, we introduce some notation that are specific to the case of mixture models we aim at studying in this section. Given $J \in \mathbb{N}^\star$, we introduce the simplex of $\mathbb{R}^J$:

$$\mathcal{S}_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, \dots, J\}\,,\ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}$$

and we also define $\mathcal{S}_J^+ = \{\boldsymbol{\lambda} \in \mathcal{S}_J \ : \ \forall j \in \{1, \dots, J\}\,,\ \lambda_j > 0\}$. In addition, we let $\Theta = (\theta_1, \dots, \theta_J) \in \mathsf{T}^J$ be fixed and for all $\boldsymbol{\lambda} \in \mathcal{S}_J$, we define $\mu_{\boldsymbol{\lambda}} \in \mathrm{M}_1(\mathsf{T})$ by $\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$.

Consequently, $\mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y)$ corresponds to a mixture model and if we let $(\mu_n)_{n\in\mathbb{N}^\star}$ be defined by $\mu_1 = \mu_{\boldsymbol{\lambda}}$ and (3.2), an immediate induction yields that for every $n \in \mathbb{N}^\star$, μ_n can be expressed as $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ where $\boldsymbol{\lambda}_n = (\lambda_{1,n}, \dots, \lambda_{J,n}) \in \mathcal{S}_J$ satisfies the initialisation $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}$ and the update formula:

$$\boldsymbol{\lambda}_{n+1} = \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_n)\,,\ n \in \mathbb{N}^\star\,, \tag{3.5}$$

where for all $\boldsymbol{\lambda} \in \mathcal{S}_J$,

$$\mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}) := \left(\frac{\lambda_j \Gamma(b_{\mu_{\boldsymbol{\lambda}},\alpha}(\theta_j) + \kappa)}{\sum_{\ell=1}^J \lambda_\ell \Gamma(b_{\mu_{\boldsymbol{\lambda}},\alpha}(\theta_\ell) + \kappa)} \right)_{1 \leqslant j \leqslant J}$$

with $\Gamma(v) = [(\alpha-1)v + 1]^{\frac{\eta}{1-\alpha}}$ for all $v \in \mathrm{Dom}_\alpha$. Finally, let us rewrite (3.A2) in the simplified case where the initial measure μ_1 is a sum of Dirac measures, which gives (3.A3) below.

(3.A3) (i) For all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;

(ii) we have $\displaystyle\int_{\mathsf{Y}} \max_{1\leqslant j\leqslant J} k(\theta_j, y) \times \max_{1\leqslant j'\leqslant J} \left(\frac{k(\theta_{j'}, y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) < \infty$.

If $\alpha = 0$, we assume in addition that $\displaystyle\int_{\mathsf{Y}} \max_{1\leqslant j\leqslant J} \left|\log\left(\frac{k(\theta_j, y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) < \infty$.

We then have the following theorem, which establishes the full proof of the global convergence towards the optimum for the mixture weights under alleviated assumptions when $\alpha < 1$.

Theorem 10. *Assume (3.A1) and (3.A3). Let $\alpha < 1$, let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$ be fixed and let κ be such that $\kappa \leqslant 0$. Then for all $\boldsymbol{\lambda} \in \mathcal{S}_J$, $\Psi_\alpha(\mu_{\boldsymbol{\lambda}}) < \infty$ and for any $\eta > 0$ the sequence $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ defined by $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$ and (3.5) is well-defined. If in addition $(\boldsymbol{\lambda}_1, \eta) \in \mathcal{S}_J^+ \times (0, 1]$ and $\{K(\theta_1, \cdot), \ldots, K(\theta_J, \cdot)\}$ are linearly independent, then*

(i) $(\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k))_{n\in\mathbb{N}^\star}$ *is nonincreasing,*

(ii) the sequence $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ converges to some $\boldsymbol{\lambda}_\star \in \mathcal{S}_J$ which is a fixed point of $\mathcal{I}_\alpha^{\mathrm{mixt}}$,

(iii) $\Psi_\alpha(\mu_{\boldsymbol{\lambda}_\star} k) = \inf_{\boldsymbol{\lambda}'\in\mathcal{S}_J} \Psi_\alpha(\mu_{\boldsymbol{\lambda}'} k)$.

The proof of this result builds on Theorem 8 and 9 and is deferred to Section 3.A.2. Notice that since Ψ_α depends on $\boldsymbol{\lambda}$ through $\mu_{\boldsymbol{\lambda}} K$ in Theorem 10, an identifiably condition was to be expected in order to achieve the convergence of the sequence $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$. Following Example 6, this identifiably condition notably holds for $J \leqslant d$ under the assumption that the $\theta_1, ..., \theta_J$ are full-rank.

We thus have the convergence of the Power Descent under less stringent conditions when $\alpha < 1$ and when we consider the particular case of mixture models. This algorithm can easily become feasible for any choice of kernel K by resorting to an unbiased estimator of $(b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j))_{1\leqslant j\leqslant J}$ in the update formula (3.5) (as already seen in Algorithm 6 of Chapter 2).

Nevertheless, contrary to the case $\alpha > 1$ we still do not have a convergence rate for the Power Descent when $\alpha < 1$. Furthermore, the important case $\alpha = 1$ in (3.1), which corresponds to performing forward Kullback-Leibler minimisation, is not covered by the Power Descent algorithm. In the next section, we extend the Power Descent to the case $\alpha = 1$. As we shall see, this will lead us to investigate the connections between the Power Descent and the Entropic Mirror Descent beyond the (α, Γ)-descent framework. As a result, we will introduce a novel algorithm closely-related to the Power Descent that yields an $O(1/N)$ convergence rate when $\mu = \mu_{\boldsymbol{\lambda}}$ and $\alpha < 1$ (and more generally when $\mu \in \mathrm{M}_1(\mathsf{T})$ and $\alpha \in \mathbb{R}$).

3.4 Power Descent and Entropic Mirror Descent

Recall from Section 4.1 that the Power Descent is defined for all $\alpha \in \mathbb{R} \setminus \{1\}$. In this section, we first establish in Proposition 19 that the Power Descent can be extended to the case $\alpha = 1$ and that we recover an Entropic Mirror Descent, showing that a deeper connection exists between the two approaches beyond the one identified by the (α, Γ)-descent framework.

This result relies on typical convergence and differentiability assumptions summarised in (D1)

(D1) For some $\varepsilon > 0$: for all $\alpha \in [1 - \varepsilon, 1)$ or $\alpha \in (1, 1 + \varepsilon]$,

(i) there exists a function $N : \mathsf{Y} \to (0, +\infty)$ satisfying: $\int_{\mathsf{Y}} N(y)\nu(\mathrm{d}y) < \infty$ and

$$\sup_{\theta\in\mathsf{T}} k(\theta,\cdot) \times \sup_{\theta'\in\mathsf{T}} \left(\frac{k(\theta',\cdot)}{p(\cdot)}\right)^{\alpha-1} < N(\cdot) \; ;$$

(ii) there exists a function $M : \mathsf{Y} \to (0,+\infty)$ satisfying: $\int_{\mathsf{Y}} M(y)\nu(\mathrm{d}y) < \infty$ and

$$\sup_{\theta\in\mathsf{T}} k(\theta,\cdot) \times \sup_{\theta'\in\mathsf{T}} \left|\log\left(\frac{k(\theta',\cdot)}{p(\cdot)}\right)\right| \times \sup_{\theta''\in\mathsf{T}} \left(\frac{k(\theta'',\cdot)}{p(\cdot)}\right)^{\alpha-1} < M(\cdot) \; ;$$

(iii) for all $y \in \mathsf{Y}$, we have $\int_{\mathsf{Y}} \inf_{\theta\in\mathsf{T}} k(\theta,y) \times \inf_{\theta'\in\mathsf{T}} \left(\frac{k(\theta',y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) > 0.$

Note that these assumptions are mild if we assume that T is a compact metric space, which is generally the case. In addition, assumption (D1)-(iii) is only required when $\alpha > 1$ to ensure that the quantity $[(\alpha-1)(b_{\mu,\alpha}+\kappa)+1]^{\frac{\eta}{1-\alpha}}$ is bounded from above. This assumption could also be replaced by the assumption that κ is such that $(\alpha-1)\kappa > 0$. We then deduce Proposition 19 below.

Proposition 19 (Limiting case $\alpha \to 1$)**.** *Assume (3.A1) and (D1). Let $\eta > 0$ and κ be such that $(\alpha-1)\kappa \geqslant 0$. Then, for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all continuous and bounded real-valued functions h on T, we have that*

$$\lim_{\alpha\to 1} [\mathcal{I}_\alpha(\mu)](h) = [\mathcal{I}_1(\mu)](h) \; ,$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$, we have set

$$\mathcal{I}_1(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta)e^{-\eta b_{\mu,1}(\theta)}}{\mu\left(e^{-\eta b_{\mu,1}}\right)} \quad \textit{and} \quad b_{\mu,1}(\theta) = \int_{\mathsf{Y}} k(\theta,y)\log\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y) \; . \tag{3.6}$$

Proof. For all $\theta \in \mathsf{T}$, the Dominated Convergence Theorem and (D1)-(i) yield

$$\lim_{\alpha\to 1} (\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1 = \lim_{\alpha\to 1} \int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y) + 0 = 1 \; .$$

Then, using (D1)-(ii) we have that for all $\theta \in \mathsf{T}$,

$$\begin{aligned}
&\lim_{\alpha\to 1} [(\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1]^{\frac{\eta}{1-\alpha}} \\
&\qquad = \exp\left(\lim_{\alpha\to 1} -\eta\frac{\log[(\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1]}{\alpha-1}\right) \\
&\qquad = \exp\left(\lim_{\alpha\to 1} -\eta\frac{\int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1}\log\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y)+\kappa}{\int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)+(\alpha-1)\kappa}\right) \\
&\qquad = \exp\left[-\eta\int_{\mathsf{Y}} k(\theta,y)\log\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y)\right]\exp\left(-\eta\kappa\right)
\end{aligned}$$

In addition, by the Dominated Convergence Theorem (and (D1)-(iii) when $\alpha > 1$), we have

$$\lim_{\alpha \to 1} \mu \left([(\alpha - 1)(b_{\mu,\alpha} + \kappa) + 1]^{\frac{\eta}{1-\alpha}} \right) = \mu \left(\exp \left[-\eta \int_{\mathsf{Y}} k(\cdot, y) \log \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y) \right] \right) \exp \left(-\eta \kappa \right) .$$

Thus,

$$\lim_{\alpha \to 1} [\mathcal{I}_\alpha(\mu)](h) = \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta) h(\theta) e^{-\eta \int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y)}}{\mu \left(e^{-\eta \int_{\mathsf{Y}} k(\cdot, y) \log \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y)} \right)} = [\mathcal{I}_1(\mu)](h) .$$

□

We recognise the one-step transition associated to the Entropic Mirror Descent applied to $\mu \mapsto \Psi_1(\mu k)$ in (3.6). This algorithm is a special case of Chapter 2 with $\Gamma(v) = e^{-\eta v}$ and $\alpha = 1$ and as such, it is known to lead to a systematic decrease in Ψ_1 and to enjoy an $O(1/N)$ convergence rate under the assumptions that (3.4) holds and $\eta \in (0, 1)$ Theorem 3.

We have thus obtained that the Power Descent coincides exactly with the Entropic Mirror Descent applied to Ψ_1 when $\alpha = 1$ and we now focus on understanding the links between Power Descent and Entropic Mirror Descent when $\alpha \in \mathbb{R} \setminus \{1\}$. For this purpose, let κ be such that $(\alpha - 1)\kappa \geqslant 0$ and let us study first-order approximations of the Power Descent and the Entropic Mirror Descent applied to Ψ_α when $b_{\mu_n,\alpha}(\theta) \approx \mu_n(b_{\mu_n,\alpha})$ for all $\theta \in \mathsf{T}$.

Letting $\eta > 0$, we have that the update formula for the Power Descent is given by

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \left[(\alpha - 1)(b_{\mu_n,\alpha}(\theta) + \kappa) + 1 \right]^{\frac{\eta}{1-\alpha}}}{\mu_n([(\alpha - 1)(b_{\mu_n,\alpha} + \kappa) + 1]^{\frac{\eta}{1-\alpha}})} , \quad n \in \mathbb{N}^\star .$$

Now using the first-order approximation $u^{\frac{\eta}{1-\alpha}} \approx v^{\frac{\eta}{1-\alpha}} - \frac{\eta}{\alpha - 1} v^{\frac{\eta}{1-\alpha} - 1}(u - v)$ with $u = \frac{(\alpha-1)(b_{\mu_n,\alpha}(\theta)+\kappa)+1}{(\alpha-1)(\mu(b_{\mu_n,\alpha})+\kappa)+1}$ and $v = 1$, we can deduce the following approximated update formula

$$\mu_{n+1}(\mathrm{d}\theta) = \mu_n(\mathrm{d}\theta) \left[1 - \frac{\eta}{\alpha - 1} \frac{b_{\mu_n,\alpha}(\theta) - \mu_n(b_{\mu_n,\alpha})}{\mu_n(b_{\mu_n,\alpha}) + \kappa + 1/(\alpha - 1)} \right] , \quad n \in \mathbb{N}^\star .$$

Letting $\eta' > 0$, the update formula for the Entropic Mirror Descent applied to Ψ_α can be written as

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \exp \left[-\eta'(b_{\mu_n,\alpha}(\theta) + \kappa) \right]}{\mu_n(\exp \left[-\eta'(b_{\mu_n,\alpha} + \kappa) \right])} , \quad n \in \mathbb{N}^\star , \tag{3.7}$$

and we obtain in a similar fashion that an approximated version of this iterative scheme is

$$\mu_{n+1}(\mathrm{d}\theta) = \mu_n(\mathrm{d}\theta)\left[1 - \eta'\left(b_{\mu_n,\alpha}(\theta) - \mu_n(b_{\mu_n,\alpha})\right)\right] \ , \quad n \in \mathbb{N}^\star \ .$$

Thus, for the two approximated formulas above to coincide, we need to set $\eta' = \eta\left[(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1\right]^{-1}$. Now coming back to (3.7), we see that this leads us to consider the update formula given by

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta)\exp\left[-\eta\frac{b_{\mu_n,\alpha}(\theta)}{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1}\right]}{\mu_n\left(\exp\left[-\eta\frac{b_{\mu_n,\alpha}}{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1}\right]\right)} \ , \quad n \in \mathbb{N}^\star \ . \tag{3.8}$$

Observe then that (3.8) can again be seen as an Entropic Mirror Descent, but applied this time to the objective function $\mu \mapsto \Psi_\alpha^{AR}(\mu k)$, where for all $\alpha \in \mathbb{R}\setminus\{0,1\}$ and all probability density q with respect to ν on $(\mathsf{Y},\mathcal{Y})$ we have set

$$\Psi_\alpha^{AR}(q) := \frac{1}{\alpha(\alpha-1)}\log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha-1)\kappa\right) \ .$$

This means that we have applied the monotonic transformation

$$u \mapsto \frac{1}{\alpha(\alpha-1)}\log\left(\alpha(\alpha-1)u + \alpha + (1-\alpha)\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right)$$

to the initial objective function Ψ_α (see Section 3.A.3 for the derivation of (3.8) based on the objective function Ψ_α^{AR}).

Hence, in the spirit of Renyi's α-divergence gradient-based methods for Variational Inference (e.g. Hernandez-Lobato et al., 2016; Li and Turner, 2016), we can motivate the iterative scheme (3.8) by observing that we recover the VR bound introduced in Li and Turner, 2016 up to a constant $-\alpha^{-1}$ when we let $p = p(\cdot,\mathscr{D})$, $\kappa = 0$ and $\alpha > 0$ in $\Psi_\alpha^{AR}(\mu k)$. For this reason we call the algorithm given by (3.8) the *Renyi Descent* thereafter.

Contrary to the Entropic Mirror Descent applied to Ψ_α, the Renyi Descent now shares the same first-order approximation as the Power Descent. This might explain why the behavior of the Entropic Mirror Descent applied to Ψ_α and of the Power Descent differed greatly when $\alpha < 1$ in the numerical experiments from Chapter 2 despite their theoretical connection through the (α,Γ)-descent framework (the former performing poorly numerically compared to the later as the dimension increased).

Strikingly, we can prove an $O(1/N)$ convergence rate towards the global optimum for the Renyi Descent. Letting $\kappa' \in \mathbb{R}$, denoting by Dom_α^{AR} an interval of $\mathbb{R}$ such that for all $\theta \in \mathsf{T}$ and all $\mu \in \mathrm{M}_1(\mathsf{T})$,

$$\frac{b_{\mu,\alpha}(\theta) + 1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1} + \kappa' \quad \text{and} \quad \frac{\mu(b_{\mu,\alpha}) + 1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1} + \kappa' \in \mathrm{Dom}_\alpha^{AR}$$

and introducing the assumption on η

(3.A4) For all $v \in \mathrm{Dom}_\alpha^{AR}$, $1 - \eta(\alpha-1)(v-\kappa') \geqslant 0$.

we indeed have the following convergence result.

Theorem 11. *Assume (3.A1) and (3.A4). Let $\alpha \in \mathbb{R}\backslash\{1\}$ and let κ be such that $(\alpha-1)\kappa > 0$. Define $|b|_{\infty,\alpha} := \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta) + 1/(\alpha-1)|$ and assume that $|b|_{\infty,\alpha} < \infty$. Moreover, let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. Then, the following assertions hold.*

(i) The sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (3.8) is well-defined and $(\Psi_\alpha(\mu_n k))_{n\in\mathbb{N}^\star}$ is non-increasing.

(ii) For all $N \in \mathbb{N}^\star$, we have

$$\Psi_\alpha(\mu_N k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}(\alpha-1)\kappa}\Delta_1\right], \tag{3.9}$$

where $\mu^\star$ is such that $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})}\Psi_\alpha(\zeta k)$, $\mathrm{M}_{1,\mu_1}(\mathsf{T})$ denotes the set of probability measures dominated by μ_1, $KL(\mu^\star||\mu_1) = \int_\mathsf{T}\log\left(\mathrm{d}\mu^\star/\mathrm{d}\mu_1\right)\mathrm{d}\mu^\star$, $\Delta_1 = \Psi_\alpha(\mu_1 k) - \Psi_\alpha(\mu^\star k)$ and $L_{\alpha,2}$, L, $L_{\alpha,3}$, $L_{\alpha,1}$ are finite constants.

The proof of this result is deferred to Section 3.A.4 (alongside with the definition of the constants $L_{\alpha,2}$, L, $L_{\alpha,3}$, $L_{\alpha,1}$ in (3.18)) and we present in the next example an application of this theorem to the particular case of mixture models.

Example 7. *Let $\alpha \in \mathbb{R}\backslash\{1\}$, let $J \in \mathbb{N}^\star$, let $\Theta = (\theta_1,\ldots,\theta_J) \in \mathsf{T}^J$, let $\mu_1 = J^{-1}\sum_{j=1}^J \delta_{\theta_j}$ and let $\mathrm{Dom}_\alpha^{AR} = [-\frac{|b|_{\infty,\alpha}}{(\alpha-1)\kappa} + \kappa', \frac{|b|_{\infty,\alpha}}{(\alpha-1)\kappa} + \kappa']$ with $\kappa' \in \mathbb{R}$. In addition, assume that $1 - \eta|\kappa|^{-1}|b|_{\infty,\alpha} > 0$. Then, taking $\kappa' = -3\frac{|b|_{\infty,\alpha}}{(\alpha-1)\kappa}$, we obtain*

$$\Psi_\alpha(\mu_N k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{|\alpha-1|(|b|_{\infty,\alpha}+|\kappa|)}{N}\left[\frac{\log J}{\eta} + \frac{\sqrt{2\log(J)}|b|_{\infty,\alpha}}{(\alpha-1)\kappa(1-\eta|\kappa|^{-1}|b|_{\infty,\alpha})}\right],$$

where we have used that $KL(\mu^\star||\mu_1) \leqslant \log J$, $\Delta_1 \leqslant \sqrt{2\log J}|b|_{\infty,\alpha}$ and that the constants defined in (3.18) satisfy $L_{\alpha,2} = \eta^{-1}|\alpha-1|(|b|_{\infty,\alpha}+|\kappa|)$, $L = \eta^2 e^{\eta\frac{|b|_{\infty,\alpha}}{(\alpha-1)\kappa} - \eta\kappa'}$, $L_{\alpha,3} = e^{\eta\frac{|b|_{\infty,\alpha}}{(\alpha-1)\kappa}+\eta\kappa'}$ and $L_{\alpha,1} = (1-\eta|\kappa|^{-1}|b|_{\infty,\alpha})\eta e^{-\eta\frac{|b|_{\infty,\alpha}}{(\alpha-1)\kappa}-\eta\kappa'}$.

To put things into perspective, notice that under our assumptions the Renyi Descent enjoys an $O(1/\sqrt{N})$ convergence rate as a Entropic Mirror Descent algorithm for the sequence $(\Psi_\alpha(N^{-1}\sum_{n=1}^N \mu_n))_{N\in\mathbb{N}^\star}$ when η is proportional to $1/\sqrt{N}$, N being fixed (see Beck and Teboulle, 2003 or Bubeck, 2015, Theorem 4.2.).

The improvement thus lies in the fact that deriving an $O(1/N)$ convergence rate usually requires stronger smoothness assumptions on Ψ_α (Bubeck, 2015, Theorem 6.2) that we do not assume in Theorem 11. Furthermore, due to the monotonicity property, our result only involves the measure μ_N at time N while typical Entropic Mirror Result are expressed in terms of the average $N^{-1}\sum_{n=1}^N \mu_n$.

TABLE 3.1: Summary of the theoretical results obtained in this chapter compared to Chapter 2

	Power Descent	Renyi Descent
Chapter 2	$\alpha < 1$: convergence under restrictive assumptions; $\alpha > 1$: $O(1/N)$ convergence rate	not covered
This chapter	$\alpha < 1$: full proof of convergence for mixture weights; extension to $\alpha = 1$	$O(1/N)$ convergence rate

Finally, observe that the Renyi Descent becomes feasible in practice for any choice of kernel K by letting μ be a weighted sum of Dirac measures i.e. $\mu = \mu_{\boldsymbol{\lambda}}$ and by resorting to an unbiased estimate of $(b_{\mu,\alpha}(\theta_j))_{1\leqslant j\leqslant J}$ (e.g. Algorithm 9).

The theoretical results we have obtained compared to Chapter 2 are summarised in Table 3.1 for clarity and we next move on to numerical experiments.

Algorithm 9: *Practical version of the Renyi Descent for mixture models*

Input: p: measurable positive function, K: Markov transition kernel, M: number of samples, $\Theta = \{\theta_1, \dots, \theta_J\} \subset \mathsf{T}$: parameter set, $\Gamma(v) = e^{-\eta v}$ with η as in (3.A4), N: total number of iterations.

Output: Optimised weights $\boldsymbol{\lambda}$.

Set $\boldsymbol{\lambda} = [\lambda_{1,1}, \dots, \lambda_{J,1}]$.

for $n = 1 \dots N$ **do**

Sampling step : Draw independently M samples $Y_1, \dots, Y_M$ from $\mu_{\boldsymbol{\lambda}} k$.

Expectation step : Compute $\boldsymbol{B}_{\boldsymbol{\lambda}} = (b'_j)_{1\leqslant j\leqslant J}$ where for all $j = 1 \dots J$

$$b_j = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_m)}{\mu_{\boldsymbol{\lambda}} k(Y_m)} f'_\alpha \left(\frac{\mu_{\boldsymbol{\lambda}} k(Y_m)}{p(Y_m)} \right)$$

and for all $j = 1 \dots J$

$$b'_j = \frac{b_j}{(\alpha - 1)(\sum_{\ell=1}^{J} b_\ell + \kappa) + 1}$$

and deduce $\boldsymbol{W}_{\boldsymbol{\lambda}} = (\lambda_j \Gamma(b'_j + \kappa'))_{1\leqslant j\leqslant J}$ and $w_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j \Gamma(b'_j + \kappa')$.

Iteration step : Set

$$\boldsymbol{\lambda} \leftarrow \frac{1}{w_{\boldsymbol{\lambda}}} \boldsymbol{W}_{\boldsymbol{\lambda}}$$

end

3.5 Simulation study

Let the target p be a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant c so that for all $y \in \mathsf{Y}$, $p(y) = c \times [0.5\mathcal{N}(y; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(y; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, where $\boldsymbol{u_d}$ is the d-dimensional vector whose coordinates are all equal to 1, $s = 2$, $c = 2$ and $\boldsymbol{I_d}$ is the identity matrix. Given $J \in \mathbb{N}^\star$, the approximating family is described by

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \theta_1, \ldots, \theta_J \in \mathsf{T} \right\} ,$$

where K_h is a Gaussian transition kernel with bandwidth h and k_h denotes its associated kernel density.

Since the Power Descent and the Renyi Descent operate only on the mixture weights $\boldsymbol{\lambda}$ of $\mu_{\boldsymbol{\lambda}} k_h$ during the optimisation, as seen in Chapter 2 a fully adaptive algorithm is obtained by alternating T times between an *Exploitation step* where the mixture weights are optimised and an *Exploration step* where the $\theta_1, \ldots, \theta_J$ are updated, as written in Algorithm 10.

Algorithm 10: *Complete Exploitation-Exploration Algorithm*

Input: p: measurable positive function, α: α-divergence parameter, q_0: initial sampler, K_h: Gaussian transition kernel, T: total number of iterations, J: dimension of the parameter set.

Output: Optimised weights $\boldsymbol{\lambda}$ and parameter set Θ.

Draw $\theta_{1,1}, \ldots, \theta_{J,1}$ from q_0.

for $t = 1 \ldots T$ **do**

Exploitation step : Set $\Theta = \{\theta_{1,t}, \ldots, \theta_{J,t}\}$. Perform the Power Descent or Renyi Descent and obtain the optimised mixture weights $\boldsymbol{\lambda}$.

Exploration step : Perform any exploration step of our choice and obtain $\theta_{1,t+1}, \ldots, \theta_{J,t+1}$.

end

As mentioned in Chapter 2 many choices are possible for the Exploration step of Algorithm 10 since there is no constraint on $\{\theta_1, \ldots, \theta_J\}$. Here, we use the same Exploration step as the one used in Chapter 2. This means that h is set to be proportional to $J^{-1/(4+d)}$ and that the particles are updated by i.i.d sampling according to $\mu_{\boldsymbol{\lambda}} k_h$ in the Exploration step.

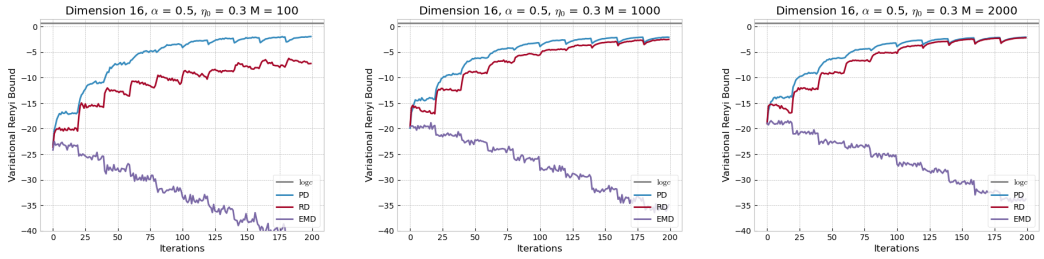
As for the Power Descent and Renyi Descent, we perform N transitions of these algorithms at each time $t = 1 \ldots T$ according to Algorithm 6 with $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/1-\alpha}$ and Algorithm 9, in which the initial weights are set to be $[1/J, \ldots, 1/J]$,

$\eta = \eta_0/\sqrt{N}$ with $\eta_0 > 0$ and M samples are used in the estimation of $(b_{\mu_{\boldsymbol{\lambda}},\alpha}(\theta_{j,t}))_{1\leqslant J}$ at each iteration $n = 1 \dots N$.

We take $J = 100$, $M \in \{100, 1000, 2000\}$, $\alpha = 0.5$, $\kappa = 0$, $\eta_0 = 0.3$ and the initial particles $\theta_1, \dots, \theta_J$ are sampled from a centered normal distribution q_0 with covariance matrix $5\boldsymbol{I_d}$. We let $T = 10$, $N = 20$ and we replicate the experiment 100 times independently in dimension $d = 16$ for each algorithm. The convergence is assessed using a Monte Carlo estimate of the VR bound introduced in Li and Turner, 2016 (which requires next to none additional computations).

The results for the Power Descent and the Renyi Descent are displayed on Figure 3.1 below and we add the Entropic Mirror Descent applied to Ψ_α as a reference.

FIGURE 3.1: Plotted is the average VR bound for the Power Descent (PD), the Renyi Descent (RD) and the Entropic Mirror Descent applied to Ψ_α (EMD) in dimension $d = 16$ computed over 100 replicates with $\eta_0 = 0.3$ and $\alpha = 0.5$ and an increasing number of samples M.



We then observe that the Renyi Descent is indeed better-behaved compared to the Entropic Mirror Descent applied to Ψ_α, which fails in dimension 16. Furthermore, it matches the performances of the Power Descent as M increases in our numerical experiment, which illustrates the link between the two algorithms we have established in the previous section.

Discussion From a theoretical standpoint, no convergence rate is yet available for the Power Descent algorithm when $\alpha < 1$. An advantage of the novel Renyi Descent algorithm is then that while being close to the Power Descent, it also benefits from the Entropic Mirror Descent optimisation literature and as such $O(1/\sqrt{N})$ convergence rates hold, which we have been able to improve to $O(1/N)$ convergence rates.

A practical use of the Power Descent and of the Renyi Descent algorithms requires approximations to handle intractable integrals appearing in the update formulas so that the Power Descent applies the function $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$ to an *unbiased* estimator of the translated gradient $b_{\mu,\alpha}(\theta) + \kappa$ before renormalising, while the the Renyi Descent applies the Entropic Mirror Descent function $\Gamma(v) = e^{-\eta v}$ to a *biased* estimator of $b_{\mu_n,\alpha}(\theta)/(\mu_n(b_{\mu_n,\alpha}) + \kappa + 1/(\alpha-1))$ before renormalising.

Finding which approach is best between biased and unbiased α-divergence minimisation is still an open issue in the literature, both theoretically and empirically

(Geffner and Domke, 2020a; Geffner and Domke, 2020b; Dhaka et al., 2021). Due to the exponentiation, considering the α-divergence instead of Renyi's α-divergence has for example been said to lead to high-variance gradients (Dieng et al., 2017; Li and Turner, 2016) and low Signal-to-Noise ratio when $\alpha \neq 0$ (Geffner and Domke, 2020b) during the Stochastic Gradient Descent optimization.

In that regard, our work sheds light on additional links between unbiased and biased α-divergence methods beyond the framework of Stochastic Gradient Descent algorithms, as both the unbiased Power Descent and the biased Renyi Descent share the same first-order approximation.

3.6 Conclusion and perspectives

We investigated algorithms that can be used to perform mixture weights optimisation for α-divergence minimisation regardless of how the mixture parameters are obtained. More precisely, we have established the full proof of the convergence of the Power Descent algorithm in the case $\alpha < 1$ when we consider mixture models and bridged the gap with the case $\alpha = 1$. We also introduced a closely-related algorithm called the Renyi Descent. We proved it enjoys an $O(1/N)$ convergence rate and illustrated in practice the proximity between these two algorithms when the number of samples M increases.

Further work could include establishing theoretical results regarding the stochastic version of these two algorithms, as well as providing complementary empirical results comparing the performances of the unbiased α-divergence-based Power Descent algorithm to those of the biased Renyi's α-divergence-based Renyi Descent.

These aspects are beyond the scope of this thesis and we now move on to Chapter 4, in which we focus on finding an appropriate Exploration step that can combined with the mixture weights optimisation framework we have developed during the course of Chapter 2 and Chapter 3.

3.A Deferred results

3.A.1 Proof that (3.A2) is satisfied in Example 6

Proof that (3.A2) is satisfied in Example 6.

We have $k_h(\theta, y) = \frac{e^{-\|y-\theta\|^2/(2h^2)}}{(2\pi h^2)^{d/2}}$ and $p(y) = c \times \left[0.5\frac{e^{-\|y-\theta_1^\star\|^2/2}}{(2\pi)^{d/2}} + 0.5\frac{e^{-\|y-\theta_2^\star\|^2/2}}{(2\pi)^{d/2}}\right]$ for all $\theta \in \mathsf{T}$ and all $y \in \mathsf{Y}$. Recall that by assumption $\mathsf{T} = \mathcal{B}(0, r) \subset \mathbb{R}^d$ with $r > 0$. Then, for all $\alpha \in [0, 1)$, we are interested in proving

$$\int_{\mathsf{Y}} \sup_{\theta\in\mathsf{T}} k(\theta, y) \times \sup_{\theta'\in\mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) < \infty \tag{3.10}$$

and

$$\int_{\mathsf{Y}} \sup_{\theta\in\mathsf{T}} \left|\log\left(\frac{k_h(\theta, y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) < \infty\,. \tag{3.11}$$

(i) We start by proving (3.10). First note that for all $\theta, \theta' \in \mathsf{T}$ and for all $y \in \mathsf{Y}$ we can write

$$\begin{aligned}\frac{k_h(\theta, y)}{k_h(\theta', y)} &= e^{\frac{-\|y-\theta\|^2+\|y-\theta'\|^2}{2h^2}} = e^{\frac{2<y,\theta-\theta'>-\|\theta\|^2+\|\theta'\|^2}{2h^2}} \\ &\leqslant e^{\frac{2|<y,\theta-\theta'>|+\|\theta\|^2+\|\theta'\|^2}{2h^2}} \leqslant e^{\frac{\|y\|\|\theta-\theta'\|+r^2}{h^2}}\,.\end{aligned}$$

from which we deduce that for all $\theta, \theta' \in \mathsf{T}$ and for all $y \in \mathsf{Y}$,

$$\frac{k_h(\theta, y)}{k_h(\theta', y)} \leqslant e^{\frac{\|y\|2r+r^2}{h^2}} \tag{3.12}$$

and that

$$\int_{\mathsf{Y}} \sup_{\theta\in\mathsf{T}} k(\theta, y) \times \sup_{\theta'\in\mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) \leqslant \int_{\mathsf{Y}} k(\theta, y) e^{\frac{\|y\|2r+r^2}{h^2}} \sup_{\theta'\in\mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y).$$

Additionally, Jensen's inequality applied to the concave function $u \mapsto u^{1-\alpha}$ implies

$$\begin{aligned}\int_{\mathsf{Y}} k(\theta, y) e^{\frac{\|y\|2r+r^2}{h^2}} \sup_{\theta'\in\mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) &\leqslant \left(\int_{\mathsf{Y}} k(\theta, y) e^{\frac{\|y\|2r+r^2}{(1-\alpha)h^2}} \sup_{\theta'\in\mathsf{T}} \frac{p(y)}{k(\theta', y)} \nu(\mathrm{d}y)\right)^{1-\alpha} \\ &\leqslant \left(\int_{\mathsf{Y}} \sup_{\theta,\theta'\in\mathsf{T}} \frac{k_h(\theta, y)}{k_h(\theta', y)} e^{\frac{\|y\|2r+r^2}{(1-\alpha)h^2}} p(y)\nu(\mathrm{d}y)\right)^{1-\alpha}\end{aligned}$$

Now using (3.12), we can deduce

$$\int_{\mathsf{Y}} \sup_{\theta,\theta'\in\mathsf{T}} \frac{k_h(\theta, y)}{k_h(\theta', y)} e^{\frac{\|y\|2r+r^2}{(1-\alpha)h^2}} p(y)\nu(\mathrm{d}y) \leqslant \int_{\mathsf{Y}} e^{\frac{\|y\|2r+r^2}{h^2}(1+\frac{1}{1-\alpha})} p(y)\nu(\mathrm{d}y) < \infty\,,$$

which yields the desired result.

(ii) We now prove (3.11). For all $y \in \mathsf{Y}$ and all $\theta \in \mathsf{T}$, we have

$$e^{-\sup_{\theta\in\mathsf{T}} \frac{\|y-\theta\|^2}{2h^2}} \leqslant (2\pi h^2)^{d/2} k_h(\theta, y) \leqslant 1$$
$$e^{-\max_{i\in\{1,2\}} \frac{\|y-\theta_i^\star\|^2}{2}} \leqslant c^{-1}(2\pi)^{d/2} p(y) \leqslant 1$$

and we can deduce for all $y \in \mathsf{Y}$ and all $\theta \in \mathsf{T}$

$$\begin{aligned}\left|\log\left(\frac{k_h(\theta, y)}{p(y)}\right)\right| &\leqslant \sup_{\theta\in\mathsf{T}} \frac{\|y-\theta\|^2}{2h^2} + \max_{i\in\{1,2\}} \frac{\|y-\theta_i^\star\|^2}{2} + d|\log h| + |\log c| \\ &\leqslant \frac{(\|y\|+r)^2}{2}\left[\frac{1}{h^2}+1\right] + d|\log h| + |\log c| \ . \end{aligned} \tag{3.13}$$

Since we have

$$\int_{\mathsf{Y}} \left(\frac{(\|y\|+r)^2}{2}\left[\frac{1}{h^2}+1\right] + d|\log h| + |\log c|\right) p(y)\nu(\mathrm{d}y) < \infty$$

we deduce that (3.11) holds.

□

3.A.2 Proof of Theorem 10

We start with some preliminary results. Let $\zeta, \zeta' \in \mathrm{M}_1(\mathsf{T})$. Recall that we say that $\zeta \mathcal{R} \zeta'$ if and only if $\zeta K = \zeta' K$ and that $\mathrm{M}_{1,\zeta}(\mathsf{T})$ denotes the set of probability measures dominated by ζ.

Lemma 20. *Assume (3.A1). Let* M *be a convex subset of* $\mathrm{M}_1(\mathsf{T})$ *and let* $\zeta_1, \zeta_2 \in \mathrm{M}_1(\mathsf{T})$ *be such that*

$$\Psi_\alpha(\zeta_1 k) = \Psi_\alpha(\zeta_2 k) = \inf_{\zeta\in\mathsf{M}} \Psi_\alpha(\zeta k).$$

Then, we have $\zeta_1 \mathcal{R} \zeta_2$.

Proof. For all $y \in \mathsf{Y}$, set $u_y = \zeta_1 k(y)/p(y)$ and $v_y = \zeta_2 k(y)/p(y)$. Then, for all $y \in \mathsf{Y}$ and for all $t \in (0,1)$, $f_\alpha(tu_y + (1-t)v_y) \leqslant t f_\alpha(u_y) + (1-t) f_\alpha(v_y)$ by convexity of f_α and we obtain

$$\Psi_\alpha(t\zeta_1 k + (1-t)\zeta_2 k) \leqslant t\Psi_\alpha(\zeta_1 k) + (1-t)\Psi_\alpha(\zeta_2 k) = \inf_{\zeta\in\mathsf{M}} \Psi_\alpha(\zeta k) \ . \tag{3.14}$$

Furthermore, $t\zeta_1 + (1-t)\zeta_2 \in \mathsf{M}$ which implies that we have equality in (3.14).

Consequently, for all $t \in (0,1)$:

$$\int_{\mathsf{Y}} \underbrace{[t f_\alpha(u_y) + (1-t) f_\alpha(v_y) - f_\alpha(tu_y + (1-t)v_y)]}_{\geqslant 0} p(y)\nu(\mathrm{d}y) = 0 \ .$$

Now using that f_α is strictly convex, we deduce that for p-almost all $y \in \mathsf{Y}$, $\zeta_1 k(y) = \zeta_2 k(y)$ that is $\zeta_1 \mathcal{R} \zeta_2$. □

Lemma 21. *Assume (3.A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, let κ be such that $(\alpha-1)\kappa \geqslant 0$ and let $\mu^\star \in \mathrm{M}_1(\mathsf{T})$ be a fixed point of $\mathcal{I}_\alpha$. Then,*

$$\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})} \Psi_\alpha(\zeta k) \; . \tag{3.15}$$

Furthermore, for all $\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})$, $\Psi_\alpha(\mu^\star k) = \Psi_\alpha(\zeta k)$ implies that $\mu^\star \mathcal{R} \zeta$.

Proof. Let $\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})$ be such that $\Psi_\alpha(\zeta k) \leqslant \Psi_\alpha(\mu^\star k)$. We have that

$$\zeta\left(b_{\mu^\star,\alpha} - \mu^\star(b_{\mu^\star,\alpha})\right) \leqslant \Psi_\alpha(\zeta k) - \Psi_\alpha(\mu^\star k) \leqslant 0 \; . \tag{3.16}$$

Furthermore, since $\mu^\star$ is a fixed point of $\mathcal{I}_\alpha$, $\Gamma(b_{\mu^\star,\alpha}+\kappa)$, hence $|b_{\mu^\star,\alpha}+\kappa+1/(\alpha-1)|$ is $\mu^\star$-almost all constant. In addition, $b_{\mu^\star,\alpha} + \kappa + 1/(\alpha-1)$ is of constant sign by assumption on κ. Since $\zeta \preceq \mu^\star$, we thus deduce that

$$\zeta\left(b_{\mu^\star,\alpha} - \mu^\star(b_{\mu^\star,\alpha})\right) = 0 \; .$$

Combining this result with (3.16) yields $\Psi_\alpha(\zeta k) = \Psi_\alpha(\mu^\star k)$ and we recover (3.15).

Finally, assume there exists $\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star k) = \Psi_\alpha(\zeta k)$. Then, since $\mathrm{M}_{1,\mu^\star}(\mathsf{T})$ is a convex set, we have by Lemma 20 that $\mu^\star \mathcal{R} \zeta$. □

We now move on to the proof of Theorem 10.

Proof of Theorem 10. For convenience, we define the notation $\Psi_\alpha(\boldsymbol{\lambda}) := \Psi_\alpha\left(\mu_{\boldsymbol{\lambda}} k\right)$ for all $\boldsymbol{\lambda} \in \mathcal{S}_J$. In this proof, we will use the equivalence relation $\mathcal{R}$ defined by: $\zeta \mathcal{R} \zeta'$ if and only if $\zeta K = \zeta' K$ and we write $\mathrm{M}_{1,\zeta}(\mathsf{T})$ the set of probability measures dominated by ζ.

(i) *Any possible limit of convergent subsequence of $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ is a fixed point of $\mathcal{I}_\alpha^{\mathrm{mixt}}$.*

First note that by (3.A3), we have that $|\Psi_\alpha(\boldsymbol{\lambda})| < \infty$ and that (3.3) is satisfied for all $\mu_{\boldsymbol{\lambda}}$ such that $\boldsymbol{\lambda} \in \mathcal{S}_J$. This means that the sequence $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ defined by (3.5) is well-defined, that the sequence $(\Psi_\alpha(\boldsymbol{\lambda}_n))_{n\in\mathbb{N}^\star}$ is lower-bounded and that $\Psi_\alpha(\boldsymbol{\lambda}_n)$ is finite for all $n \in \mathbb{N}^\star$. As $(\Psi_\alpha(\boldsymbol{\lambda}_n))_{n\in\mathbb{N}^\star}$ is nonincreasing by Theorem 8-(i), it converges in $\mathbb{R}$ and in particular we have

$$\lim_{n\to\infty} \Psi_\alpha \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_n) - \Psi_\alpha(\boldsymbol{\lambda}_n) = 0 \; .$$

Let $(\boldsymbol{\lambda}_{\varphi(n)})_{n\in\mathbb{N}^\star}$ be a convergent subsequence of $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ and denote by $\bar{\boldsymbol{\lambda}}$ its limit. Since the function $\boldsymbol{\lambda} \mapsto \Psi_\alpha \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}) - \Psi_\alpha(\boldsymbol{\lambda})$ is continuous we obtain that $\Psi_\alpha \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\bar{\boldsymbol{\lambda}}) = \Psi_\alpha(\bar{\boldsymbol{\lambda}})$ and hence by Theorem 8-(ii), $\bar{\boldsymbol{\lambda}}$ is a fixed point of $\mathcal{I}_\alpha^{\mathrm{mixt}}$.

(ii) *The set $F = \left\{\boldsymbol{\lambda} \in \mathcal{S}_J \; : \; \boldsymbol{\lambda} = \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda})\right\}$ of fixed points of $\mathcal{I}_\alpha^{\mathrm{mixt}}$ is finite.*

For any subset $R \subset \{1, \dots, J\}$, define

$$\mathcal{S}_{J,R} = \{\boldsymbol{\lambda} \in \mathcal{S}_J \ : \ \forall i \in R^c, \lambda_i = 0, \forall j \in R^c, \lambda_j \neq 0\} \ ,$$
$$\tilde{\mathcal{S}}_{J,R} = \{\boldsymbol{\lambda} \in \mathcal{S}_J \ : \ \forall i \in R^c, \lambda_i = 0\} \ ,$$

and write

$$F = \bigcup_{R \subset \{1,\dots,J\}} (S_{J,R} \cap F) \ .$$

In order to show that F is finite, we prove by contradiction that for any $R \subset \{1, \dots, J\}$, $S_{J,R} \cap F$ contains at most one element. Assume indeed the existence of two distinct elements $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}'$ belonging to $S_{J,R} \cap F$. Since $\mathrm{M}_{1,\mu_{\boldsymbol{\lambda}}}(\mathsf{T}) = \mathrm{M}_{1,\mu_{\boldsymbol{\lambda}'}}(\mathsf{T}) = \left\{\mu_{\boldsymbol{\lambda}''} \ : \boldsymbol{\lambda}'' \in \tilde{\mathcal{S}}_{J,R}\right\}$, Lemma 21 implies that

$$\Psi_\alpha(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\lambda}'' \in \tilde{\mathcal{S}}_{J,R}} \Psi_\alpha\left(\boldsymbol{\lambda}''\right) = \Psi_\alpha(\boldsymbol{\lambda}') \ .$$

Applying again Lemma 21, we get $\mu_{\boldsymbol{\lambda}} \mathcal{R} \mu_{\boldsymbol{\lambda}'}$, that is, $\mu_{\boldsymbol{\lambda}} K = \mu_{\boldsymbol{\lambda}'} K$. This means that $\sum_{j=1}^J (\lambda_j - \lambda'_j) K(\theta_j, \cdot)$ is the null measure, which in turns implies the identity $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$ since the family of measures $\{K(\theta_1, \cdot), \dots, K(\theta_J, \cdot)\}$ is assumed to be linearly independent.

(iii) *Conclusion.*

According to Lemma 20 applied to the convex subset of measures $\mathsf{M} = \mathcal{S}_J$, the function Ψ_α attains its global infimum at a unique $\boldsymbol{\lambda}_\star \in \mathcal{S}_J$. The uniqueness of $\boldsymbol{\lambda}_\star$ actually follows from the fact that, as shown above, $\mu_{\boldsymbol{\lambda}} \mathcal{R} \mu_{\boldsymbol{\lambda}'}$ if and only if $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$. Then, by Theorem 8-(i) and by definition of $\boldsymbol{\lambda}_\star$

$$\Psi_\alpha \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_\star) \leqslant \Psi_\alpha(\boldsymbol{\lambda}_\star) = \inf_{\boldsymbol{\lambda}' \in \mathcal{S}_J} \Psi_\alpha(\boldsymbol{\lambda}') \leqslant \Psi_\alpha \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_\star) \ ,$$

and hence, $\Psi_\alpha \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_\star) = \Psi_\alpha(\boldsymbol{\lambda}_\star)$, showing that $\boldsymbol{\lambda}_\star \in F$ by Theorem 8-(ii). Since by (ii), F is finite, there exists $L \geqslant 1$ such that $F = \left\{\boldsymbol{\lambda}^\ell \ : 1 \leqslant \ell \leqslant L\right\}$, where for $i \neq j$, $\boldsymbol{\lambda}^i \neq \boldsymbol{\lambda}^j$. Without any loss of generality, we set $\boldsymbol{\lambda}^1 = \boldsymbol{\lambda}_\star$ to simplify the notation.

We now introduce a sequence $(W_\ell)_{1 \leqslant \ell \leqslant L}$ of disjoint open neighborhoods of $(\boldsymbol{\lambda}^\ell)_{1 \leqslant \ell \leqslant L}$ such that for any $\ell \in \{1, \dots, L\}$,

$$\mathcal{I}_\alpha^{\mathrm{mixt}}(W_\ell) \cap \left(\bigcup_{j \neq \ell} W_j \right) = \emptyset \tag{3.17}$$

This is possible since $\mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}^\ell) = \boldsymbol{\lambda}^\ell$ and $\boldsymbol{\lambda} \mapsto \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda})$ is continuous.

By (i) , the set F contains all the possible limits of any subsequence of $(\boldsymbol{\lambda}_n)_{n \in \mathbb{N}^\star}$. As a consequence, there exists $N > 0$ such that for all $n \geqslant N$, $\boldsymbol{\lambda}_n \in \bigcup_{1 \leqslant \ell \leqslant L} W_\ell$. Combining with (3.17), there exists $\ell \in \{1, \dots, L\}$ such that for all $n \geqslant N$, $\boldsymbol{\lambda}_n \in W_\ell$.

Therefore $\boldsymbol{\lambda}^\ell$ is the only possible limit of any convergent subsequence of $(\boldsymbol{\lambda}_n)_{n\in\mathbb{N}^\star}$ and as a consequence, $\lim_{n\to\infty}\boldsymbol{\lambda}_n = \boldsymbol{\lambda}^\ell$.

Thus, the sequence $(\mu_{\boldsymbol{\lambda}_n})_{n\in\mathbb{N}^\star}$ weakly converges to $\mu_{\boldsymbol{\lambda}^\ell}$ as $n\to\infty$ and Theorem 9 can be applied. Since $\boldsymbol{\lambda}_1 \in \mathcal{S}_J^+$, we have $\mathrm{M}_{1,\mu_{\boldsymbol{\lambda}_1}}(\mathsf{T}) = \left\{\mu_{\boldsymbol{\lambda}'} \ : \boldsymbol{\lambda}' \in \mathcal{S}_J\right\}$ and Theorem 9-(iii) then shows that $\mu_{\boldsymbol{\lambda}^\ell}$ is the global arginf of Ψ_α over all $\left\{\mu_{\boldsymbol{\lambda}'} \ : \boldsymbol{\lambda}' \in \mathcal{S}_J\right\}$. Therefore, $\ell = 1$, i.e., $\boldsymbol{\lambda}^\ell = \boldsymbol{\lambda}^1 = \boldsymbol{\lambda}_\star$ and

$$\Psi_\alpha(\boldsymbol{\lambda}_\star) = \inf_{\boldsymbol{\lambda}'\in\mathcal{S}_J} \Psi_\alpha(\boldsymbol{\lambda}') \ .$$

□

3.A.3 Derivation of the update formula for the Renyi Descent

For all $\alpha \in \mathbb{R}\setminus\{0,1\}$ and κ such that $(\alpha-1)\kappa \geqslant 0$, we are interested applying the Entropic Mirror Descent algorithm to the following objective function

$$\Psi_\alpha^{AR}(\mu k) := \frac{1}{\alpha(\alpha-1)}\log\left(\int_{\mathsf{Y}} \mu k(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha-1)\kappa\right)$$

Lemma 22. *Assume (3.A1). The gradient of $\Psi_\alpha^{AR}(\mu k)$ is given by $\theta \mapsto \frac{b_{\mu,\alpha}(\theta)+1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}$.*

Proof. Let $\varepsilon > 0$ be small and let $\mu, \mu' \in \mathrm{M}_1(\mathsf{T})$. Then,

$$\begin{aligned}
&\Psi_\alpha^{AR}(\mu k + \varepsilon\mu' k) \\
&\quad= \frac{1}{\alpha(\alpha-1)}\log\left(\int_{\mathsf{Y}} [(\mu+\varepsilon\mu')k(y)]^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha-1)\kappa\right) \\
&\quad= \frac{1}{\alpha(\alpha-1)}\log\left(\int_{\mathsf{Y}} \mu k(y)^\alpha \left[1+\alpha\varepsilon\frac{\mu' k(y)}{\mu k(y)}\right] p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha-1)\kappa + o(\varepsilon)\right)
\end{aligned}$$

where we used that $(1+u)^\alpha = 1+\alpha u + o(u)$ as $u\to 0$. Thus,

$$\begin{aligned}
&\Psi_\alpha^{AR}(\mu k + \varepsilon\mu' k) \\
&\quad= \Psi_\alpha^{AR}(\mu k) + \frac{1}{\alpha(\alpha-1)}\log\left(1+\alpha\varepsilon\frac{\int_{\mathsf{Y}}\mu' k(y)\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)}{\int_{\mathsf{Y}}\mu k(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha-1)\kappa} + o(\varepsilon)\right) \\
&\quad= \Psi_\alpha^{AR}(\mu k) + \varepsilon\frac{1}{\alpha-1}\frac{\int_{\mathsf{Y}}\mu' k(y)\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)}{\int_{\mathsf{Y}}\mu k(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) + (\alpha-1)\kappa} + o(\varepsilon) \\
&\quad= \Psi_\alpha^{AR}(\mu k) + \varepsilon\int_{\mathsf{T}}\mu'(\mathrm{d}\theta)\frac{1}{\alpha-1}\frac{b_{\mu,\alpha}(\theta)+1/(\alpha-1)}{\mu(b_{\mu,\alpha})+\kappa+1/(\alpha-1)} + o(\varepsilon)
\end{aligned}$$

using that $\log(1+u) = u + o(u)$ as $u\to 0$. □

Consequently, the iterative update formula for the Entropic Mirror Descent applied to the objective function Ψ_α^{AR} is given by

$$\mu_{n+1}(\mathrm{d}\theta) = \mu_n(\mathrm{d}\theta)\frac{e^{-\frac{\eta}{\alpha-1}\frac{b_{\mu_n,\alpha}(\theta)}{\mu_n(b_{\mu_n,\alpha})+\kappa+1/(\alpha-1)}}}{\mu_n(e^{-\frac{\eta}{\alpha-1}\frac{b_{\mu_n,\alpha}}{\mu_n(b_{\mu_n,\alpha})+\kappa+1/(\alpha-1)}})}\ , \quad n\in\mathbb{N}^\star\ .$$

3.A.4 Proof of Theorem 11

As we shall see, the proof can be adapted from the proof of Theorem 2. For all $\mu\in\mathrm{M}_1(\mathsf{T})$, we will use the notation

$$\mathcal{I}_\alpha^{AR}(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta)\exp\left[-\eta\frac{b_{\mu,\alpha}(\theta)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right]}{\mu\left(\exp\left[-\eta\frac{b_{\mu,\alpha}}{(\alpha-1)(\mu_n(b_{\mu,\alpha})+\kappa)+1}\right]\right)}$$

to designate the one-step transition of the Renyi Descent algorithm. Note in passing that for all $\kappa'\in\mathbb{R}$, this definition can also be rewritten under the form

$$\mathcal{I}_\alpha^{AR}(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta)\exp\left[-\eta\frac{b_{\mu,\alpha}(\theta)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}+\kappa'\right]}{\mu\left(\exp\left[-\eta\frac{b_{\mu,\alpha}}{(\alpha-1)(\mu_n(b_{\mu,\alpha})+\kappa)+1}+\kappa'\right]\right)}\ .$$

We also define

$$\begin{aligned}
L_{\alpha,2} &= \eta^{-1}\sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})}[(\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1]\\
L &= \eta^2\sup_{v\in\mathrm{Dom}_\alpha^{AR}} e^{-\eta v}\\
L_{\alpha,3} &= \sup_{v\in\mathrm{Dom}_\alpha^{AR}} e^{\eta v}\\
L_{\alpha,1} &= \inf_{v\in\mathrm{Dom}_\alpha^{AR}}\left\{1-\eta(\alpha-1)(v-\kappa')\right\}\times\eta\inf_{v\in\mathrm{Dom}_\alpha^{AR}} e^{-\eta v}\ .
\end{aligned} \tag{3.18}$$

1. Recalling Lemma 6

Let (ζ,μ) be a couple of probability measures where ζ is dominated by μ which we denote by $\zeta\preceq\mu$ and define

$$A_\alpha := \int_{\mathsf{Y}}\nu(\mathrm{d}y)\int_{\mathsf{T}}\mu(\mathrm{d}\theta)k(\theta,y)f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1-g(\theta)]\ , \tag{3.19}$$

where g is the density of ζ w.r.t μ, i.e. $\zeta(\mathrm{d}\theta)=\mu(\mathrm{d}\theta)g(\theta)$. We recall Lemma 6 from Chapter 2 in Lemma 23 below.

Lemma 23 (Lemma 6). *Assume (3.A1). Then, for all $\mu,\zeta\in\mathrm{M}_1(\mathsf{T})$ such that $\zeta\preceq\mu$ and $\Psi_\alpha(\mu k)<\infty$, we have*

$$A_\alpha \leqslant \Psi_\alpha(\mu k)-\Psi_\alpha(\zeta k)\ . \tag{3.20}$$

Moreover, equality holds in (3.20) if and only if $\zeta=\mu$.

2. Adaptation of Theorem 1

Lemma 24. *Assume (3.A1) and (3.A4). Let $\alpha \in \mathbb{R} \setminus \{1\}$, let κ be such that $(\alpha-1)\kappa \geqslant 0$ and let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that*

$$0 < \mu\left\{\exp\left(-\eta \frac{b_{\mu,\alpha} + 1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right)\right\} < \infty \tag{3.21}$$

holds and $\Psi_\alpha(\mu k) < \infty$. Then, the two following assertions hold.

(i) We have $\Psi_\alpha(\mathcal{I}_\alpha^{AR}(\mu)k) \leqslant \Psi_\alpha(\mu k)$.

(ii) We have $\Psi_\alpha(\mathcal{I}_\alpha^{AR}(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha^{AR}(\mu)$.

Proof. The proof builds on the proof of Theorem 1 in the particular case $\alpha \in \mathbb{R} \setminus \{1\}$. Indeed, in this case,

$$\begin{aligned} A_\alpha &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha-1}\left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right][1-g(\theta)] \\ &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha-1}\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} g(\theta)^{\alpha-1}[1-g(\theta)] \\ &= \int_{\mathsf{T}} \mu(\mathrm{d}\theta)\left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha-1}\right] g(\theta)^{\alpha-1}[1-g(\theta)] \ . \end{aligned}$$

so that

$$A_\alpha = [(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1] \times \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{b_{\mu,\alpha}(\theta) + \frac{1}{\alpha-1}}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1} g(\theta)^{\alpha-1}[1-g(\theta)]$$

where $(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1 > 0$ under (3.A1). Set

$$g = \tilde{\Gamma} \circ \left(\frac{b_{\mu,\alpha} + 1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right)$$

where for all $v \in \mathrm{Dom}_\alpha^{AR}$,

$$\tilde{\Gamma}(v) = \frac{e^{-\eta v}}{\mu\left\{\exp\left(-\eta \frac{b_{\mu,\alpha}+1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1} - \eta\kappa'\right)\right\}} \ .$$

Finally, let us consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and let V be the random variable

$$V(\theta) = \frac{b_{\mu,\alpha}(\theta) + 1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1} + \kappa' \ .$$

Then, we have $\mathbb{E}[1-\tilde{\Gamma}(V)] = 0$ and we can write

$$\begin{aligned} A_\alpha &= [(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1] \times \mathbb{E}[(V-\kappa')\tilde{\Gamma}^{\alpha-1}(V)(1-\tilde{\Gamma}(V))] \\ &= [(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1] \times \mathbb{C}\mathrm{ov}((V-\kappa')\tilde{\Gamma}^{\alpha-1}(V), 1-\tilde{\Gamma}(V)) \ . \end{aligned} \tag{3.22}$$

Under (3.A4) with $\alpha \in \mathbb{R} \setminus \{1\}$, $v \mapsto (v - \kappa')\tilde{\Gamma}^{\alpha-1}(v)$ and $v \mapsto 1 - \tilde{\Gamma}(v)$ are increasing on Dom_α^{AR} which implies $\mathbb{C}\mathrm{ov}(V\tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V)) \geqslant 0$ and thus $A_\alpha \geqslant 0$ since $(\alpha - 1)(\mu(b_{\mu,\alpha}) + \kappa) + 1 > 0$. □

3. Adaptation of Lemma 7

Consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and denote by $\mathbb{V}\mathrm{ar}_\mu$ the associated variance operator.

Lemma 25. *Assume (3.A1) and (3.A4). Let $\alpha \in \mathbb{R} \setminus \{1\}$, let κ be such that $(\alpha - 1)\kappa > 0$, and let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that (3.21) holds and $\Psi_\alpha(\mu k) < \infty$. Then,*

$$\frac{(\alpha - 1)\kappa L_{\alpha,1}}{2} \mathbb{V}\mathrm{ar}_\mu \left(\frac{b_{\mu,\alpha} + 1/(\alpha - 1)}{(\alpha - 1)(\mu(b_{\mu,\alpha}) + \kappa) + 1} \right) \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha^{AR}(\mu)k) \,, \quad (3.23)$$

where

$$L_{\alpha,1} := \inf_{v \in \mathrm{Dom}_\alpha^{AR}} \left\{ 1 - \eta(\alpha - 1)(v - \kappa') \right\} \times \inf_{v \in \mathrm{Dom}_\alpha^{AR}} \eta e^{-\eta v} \,.$$

Proof. The proof of Lemma 25 builds on the proof of Lemma 7. Using (3.22) combined with the fact that under (3.A1), $(\alpha - 1)(\mu(b_{\mu,\alpha}) + \kappa) + 1 > (\alpha - 1)\kappa > 0$

$$\begin{aligned} A_\alpha &= [(\alpha - 1)(\mu(b_{\mu,\alpha}) + \kappa) + 1] \times \mathbb{C}\mathrm{ov}((V - \kappa')\tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V)) \\ &> (\alpha - 1)\kappa \times \mathbb{C}\mathrm{ov}((V - \kappa')\tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V)) \end{aligned}$$

Furthermore,

$$\begin{aligned} &\mathbb{C}\mathrm{ov}((V - \kappa')\tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V)) \\ &\qquad = \frac{1}{2}\mathbb{E}\left[((U - \kappa')\tilde{\Gamma}^{\alpha-1}(U) - (V - \kappa')\tilde{\Gamma}^{\alpha-1}(V))(-\tilde{\Gamma}(U) + \tilde{\Gamma}(V)) \right] \\ &\qquad = \frac{1}{2}\mathbb{E}\left[\frac{(U - \kappa')\tilde{\Gamma}^{\alpha-1}(U) - (V - \kappa')\tilde{\Gamma}^{\alpha-1}(V)}{U - V} \frac{-\tilde{\Gamma}(U) + \tilde{\Gamma}(V)}{U - V} (U - V)^2 \right] \\ &\qquad \geqslant \frac{L_{\alpha,1}}{2} \mathbb{V}\mathrm{ar}_\mu \left(\frac{b_{\mu,\alpha} + 1/(\alpha - 1)}{(\alpha - 1)(\mu(b_{\mu,\alpha}) + \kappa) + 1} \right) \end{aligned}$$

and we thus obtain (3.23). □

4. Adaptation of the proof of Theorem 2

Proof of Theorem 11. The proof of Theorem 11 builds on the proof of Theorem 2. We prove the assertions successively.

(i) The proof of (i) simply consists in verifying that we can apply Lemma 24. For all $\mu \in \mathrm{M}_1(\mathsf{T})$, (3.21) with $\mu = \mu_n$ holds for all $n \in \mathbb{N}^\star$ by assumption on $|b|_{\infty,\alpha}$ and since at each step $n \in \mathbb{N}^\star$, Lemma 24 combined with $\Psi_\alpha(\mu_n k) < \infty$ implies that $\Psi_\alpha(\mu_{n+1}k) \leqslant \Psi_\alpha(\mu_n k) < \infty$, we obtain by induction that $(\Psi_\alpha(\mu_n k))_{n \in \mathbb{N}^\star}$ is non-increasing.

(ii) Let $n \in \mathbb{N}^\star$, set $\Delta_n = \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu^\star k)$ and for all $\theta \in \mathsf{T}$, $V_n(\theta) = \frac{b_{\mu_n,\alpha}(\theta)+\frac{1}{\alpha-1}}{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1} + \kappa'$, such that $\mathrm{d}\mu_{n+1} \propto e^{-\eta V_n}\mathrm{d}\mu_n$.

We first show that

$$\Delta_n \leqslant L_{\alpha,2}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\mathrm{d}\mu^\star + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}(V_n)L_{\alpha,3}\right] . \tag{3.24}$$

The convexity of f_α implies that

$$\begin{aligned}\Delta_n &\leqslant \int_{\mathsf{T}} b_{\mu_n,\alpha}(\mathrm{d}\mu_n - \mathrm{d}\mu^\star)\\ &= \int_{\mathsf{T}}\left(b_{\mu_n,\alpha} + \frac{1}{\alpha-1}\right)(\mathrm{d}\mu_n - \mathrm{d}\mu^\star)\\ &= \frac{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1}{\eta}\int_{\mathsf{T}}(\mu_n(\eta V_n) - \eta V_n)\mathrm{d}\mu^\star .\end{aligned}$$

Then, noting that

$$-\eta V_n = \log\mu_n\left(e^{-\eta V_n}\right) + \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)$$

we deduce

$$\Delta_n \leqslant L_{\alpha,2}\int_{\mathsf{T}}\left[\mu_n(\eta V_n) + \log\mu_n\left(e^{-\eta V_n}\right) + \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\right]\mathrm{d}\mu^\star . \tag{3.25}$$

Since $v \mapsto e^{-\eta v}$ is L-smooth on Dom_α^{AR}, for all $\theta \in \mathsf{T}$ and for all $n \in \mathbb{N}^\star$ we can write

$$e^{-\eta V_n(\theta)} \leqslant e^{-\eta\mu_n(V_n)} + \eta e^{-\eta\mu_n(V_n)}(V_n(\theta) - \mu_n(V_n)) + \frac{L}{2}\left(V_n(\theta) - \mu_n(V_n)\right)^2$$

which in turn implies

$$\mu_n(e^{-\eta V_n}) \leqslant e^{-\eta\mu_n(V_n)} + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}(V_n) .$$

Finally, we obtain

$$\log\mu_n(e^{-\eta V_n}) \leqslant \log e^{-\eta\mu_n(V_n)} + \log\left(1 + \frac{L}{2}\frac{\mathbb{V}\mathrm{ar}_{\mu_n}(V_n)}{e^{-\eta\mu_n(V_n)}}\right) .$$

Using that $\log(1+u) \leqslant u$ when $u \geqslant 0$ and by definition of $L_{\alpha,3}$, we deduce

$$\log\mu_n(e^{-\eta V_n}) \leqslant -\eta\mu_n(V_n) + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}(V_n)L_{\alpha,3} ,$$

which combined with (3.25) implies (3.24). To conclude, we apply Lemma 25 to $g = \frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}$ and combining with (3.24), we obtain

$$\Delta_n \leqslant L_{\alpha,2}\left[\int_{\mathsf{T}}\log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\mathrm{d}\mu^\star + \frac{LL_{\alpha,3}}{L_{\alpha,1}(\alpha-1)\kappa}\left(\Delta_n - \Delta_{n+1}\right)\right] ,$$

where by assumption $L_{\alpha,1}$, $L_{\alpha,2}$ and $L_{\alpha,3} > 0$. As the r.h.s involves two telescopic sums, we deduce

$$\frac{1}{N}\sum_{n=1}^{N}\Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) - KL(\mu^\star||\mu_{N+1}) + L\frac{L_{\alpha,3}}{L_{\alpha,1}(\alpha-1)\kappa}(\Delta_1 - \Delta_{N+1})\right]$$

and we recover (3.9) using (i), that $KL(\mu^\star||\mu_{N+1}) \geqslant 0$ and that $\Delta_{N+1} \geqslant 0$.

□

4

Monotonic α-divergence minimisation

The work presented in this chapter corresponds to the paper entitled "Monotonic Alpha-divergence minimisation" (Daudel, Douc, and Roueff, 2021) submitted as a journal paper at the time of writing.

4.1 Introduction

In the two previous chapters, we have been interested in developing Variational Inference iterative procedures that ensure a monotonic decrease in the α-divergence at each step for an approximating family $\mathcal{Q}$ of the form

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\} .$$

This choice of approximating family allowed us to target the class of mixture models by letting the initial measure $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be a weighted sum of Dirac measures and as a result, we have enabled mixture weights optimisation by α-divergence minimisation.

Since our procedures maintain the components parameters fixed in order to carry out the mixture weights optimisation, we suggested to alternate between them and a suitable Exploration step in charge of updating the components parameters set. Yet, the underlying question of how to select an Exploration step remains unexplored.

In this chapter, we offer to derive an iterative algorithm for components parameters optimisation that systematically decreases the α-divergence at each step. As

we shall see, the particularity of our work in Chapter 4 will be that we are able to optimise both the weights and the components parameters of a given mixture model *simultaneously*, all the while maintaining the systematic decrease in the α-divergence.

The starting point of our approach will be to work within a parametric family of the form

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \ , \tag{4.1}$$

which as we have seen in Chapter 1 is the natural idea in Variational Inference. Before getting into the details of our work, let us first introduce some notation and specify the initial optimisation problem we consider in Chapter 4 in terms of the approximating family (4.1).

Notation and problem statement We retain the notation from earlier chapters. In particular recall that f_α is the convex function on $(0, +\infty)$ defined by $f_0(u) = u - 1 - \log(u)$, $f_1(u) = 1 - u + u\log(u)$ and $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$ for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$.

For any measurable positive function p on $(\mathsf{Y}, \mathcal{Y})$, the initial optimisation problem we consider in Chapter 4 is then

$$\inf_{\theta \in \mathsf{T}} \Psi_\alpha(k(\theta, \cdot); p) \ ,$$

where for all probability density q with respect to ν on $(\mathsf{Y}, \mathcal{Y})$,

$$\Psi_\alpha(q; p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ .$$

As usual, p will be dropped for notational ease unless we refer to the Bayesian case; in that case we shall use the notation $\Psi_\alpha(q; \mathscr{D})$ instead of $\Psi_\alpha(q; p(\cdot, \mathscr{D}))$.

For all $\alpha \in \mathbb{R} \setminus \{1\}$, we also let $\tilde{f}_\alpha$ be the *convex* function on $(0, +\infty)$ defined by $\tilde{f}_0(u) = -\log(u)$ and $\tilde{f}_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1\right]$ otherwise. Notice the subtle change here compared to the definition of f_α since $f_\alpha(u) = \tilde{f}_\alpha(u) + (u-1)/(1-\alpha)$ for all $\alpha \in \mathbb{R} \setminus \{1\}$. This change is for convenience in the proofs only as for all $\alpha \in \mathbb{R} \setminus \{1\}$ and all probability density q with respect to ν on $(\mathsf{Y}, \mathcal{Y})$ we can write

$$\begin{aligned}
\Psi_\alpha(q) &= \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} \left[\tilde{f}_\alpha\left(\frac{q(y)}{p(y)}\right) + \frac{1}{1-\alpha}\left(\frac{q(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y) \\
&= \tilde{\Psi}_\alpha(q) + \frac{1}{1-\alpha}\left(1 - \int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)\right) \ ,
\end{aligned}$$

where we have set

$$\tilde{\Psi}_\alpha(q) = \int_{\mathsf{Y}} \tilde{f}_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ .$$

This means that under the assumption that $\alpha \neq 1$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$, $\Psi_\alpha(q)$ and $\tilde{\Psi}_\alpha(q)$ differ solely by an additive constant. We now give below the outline of Chapter 4.

Outline The chapter is organised as follows:

- In Section 4.2, we consider the typical Variational Inference case where q belongs to a parametric family as in (4.1). In this particular case, we state in Theorem 12 conditions which ensure a systematic decrease in the α-divergence at each step for all $\alpha \in [0,1)$. We then show in Corollary 27 that these conditions are satisfied for a well-chosen iterative scheme.

The formulation of this iterative scheme is particularly convenient, a fact that we illustrate over several examples. Furthermore, we derive in Corollary 28 additional iterative schemes satisfying the conditions of Theorem 12, which we then use to underline the links between our approach and Gradient Descent schemes for α-divergence and Renyi's α-divergence minimisation.

- In Section 4.3, we further extend the results from Section 4.2 to the more general case of mixture models. We derive in Theorem 13 and 14 conditions to simultaneously optimise both the weights and the components parameters of a given mixture model, all the while maintaining the systematic decrease in the α-divergence initially enjoyed in Theorem 12.

These conditions are then met in Corollary 30 and 31, so that we can derive algorithms that are applicable to a wide range of mixture models. Furthermore, we connect our approach to the Power Descent algorithm from Chapter 2 and provide in Proposition 32 additional monotonicity results which go beyond the case $\alpha \in [0,1)$.

We also apply our results to the particular case of Gaussian Mixture Models before recovering the Mixture Population Monte Carlo (M-PMC) algorithm from Cappé et al., 2008 as a special case.

- Lastly, we show in Section 4.4 that having enhanced our framework beyond the particular example of the M-PMC algorithm also has practical benefits when we consider multimodal targets and we provide numerical experiments to compare our results to those obtained using a typical Adaptive Importance Sampling algorithm.

4.2 An iterative algorithm for optimising $\Psi_\alpha(k(\theta,\cdot))$

In this section, our goal is to define iterative procedures which optimise $\Psi_\alpha(k(\theta,\cdot))$ with respect to θ and which are such that they ensure a *systematic decrease* in Ψ_α at each step. For this purpose, we start by introducing some mild conditions on k, p and ν that will be used throughout the chapter.

(4.A1) The density kernel k on $\mathsf{T}\times\mathsf{Y}$, the function p on Y and the σ-finite measure ν on $(\mathsf{Y},\mathcal{Y})$ satisfy, for all $(\theta,y)\in\mathsf{T}\times\mathsf{Y}$, $k(\theta,y)>0$, $p(y)\geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)<\infty$.

Let us now construct a sequence $(\theta_n)_{n\geqslant 1}$ valued in T such that $(\Psi_\alpha(k(\theta_n,\cdot))_{n\geqslant 1}$ is decreasing. The core idea of our approach will rely on the following proposition.

Proposition 26. *Assume (4.A1). For all $\alpha \in [0,1)$ and all $\theta, \theta' \in \mathsf{T}$, it holds that*

$$\Psi_\alpha(k(\theta,\cdot)) \leqslant \int_{\mathsf{Y}} \frac{k(\theta',y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \log\left(\frac{k(\theta,y)}{k(\theta',y)}\right) \nu(\mathrm{d}y) + \Psi_\alpha(k(\theta',\cdot)) \,. \tag{4.2}$$

Proof. We treat the two cases $\alpha = 0$ and $\alpha \in (0,1)$ separately.

(a) Case $\alpha = 0$, with $\tilde{f}_0(u) = -\log(u)$ for all $u > 0$. This case is immediate since

$$\Psi_0(k(\theta,\cdot)) = -\int_{\mathsf{Y}} p(y) \log\left(\frac{k(\theta,y)}{k(\theta',y)}\right) \nu(\mathrm{d}y) + \Psi_0(k(\theta',\cdot)) \,.$$

(b) Case $\alpha \in (0,1)$ with $\tilde{f}_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1\right]$ for all $u > 0$. We have that

$$\begin{aligned}\tilde{\Psi}_\alpha(k(\theta,\cdot)) &= \int_{\mathsf{Y}} \frac{\left[\left(\frac{k(\theta,y)}{p(y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} \left(\frac{k(\theta',y)}{p(y)}\right)^\alpha \frac{\left[\left(\frac{k(\theta,y)}{k(\theta',y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) + \tilde{\Psi}_\alpha(k(\theta',\cdot))\end{aligned}$$

Furthermore, the concavity of the log function gives $\log(u^\alpha) \leqslant u^\alpha - 1$ for all $u > 0$ and since $\alpha \in (0,1)$, we can write

$$\frac{1}{\alpha-1}\log(u) = \frac{1}{\alpha(\alpha-1)}\log(u^\alpha) \geqslant \tilde{f}_\alpha(u) \,.$$

Thus,

$$\Psi_\alpha(k(\theta,\cdot)) \leqslant \int_{\mathsf{Y}} \frac{k(\theta',y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \log\left(\frac{k(\theta,y)}{k(\theta',y)}\right) \nu(\mathrm{d}y) + \Psi_\alpha(k(\theta',\cdot))$$

which is exactly (4.2).

□

This result then allows us to deduce Theorem 12 below.

Theorem 12. *Assume (4.A1). Let $\alpha \in [0,1)$ and starting from an initial $\theta_1 \in \mathsf{T}$, let $(\theta_n)_{n\geqslant 1}$ be defined iteratively such that for all $n \geqslant 1$,*

$$\int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \log\left(\frac{k(\theta_{n+1},y)}{k(\theta_n,y)}\right) \nu(\mathrm{d}y) \leqslant 0 \,. \tag{4.3}$$

Further assume that $\Psi_\alpha(k(\theta_1,\cdot)) < \infty$. Then, at time n, we have $\Psi_\alpha(k(\theta_{n+1},\cdot)) \leqslant \Psi_\alpha(k(\theta_n,\cdot))$.

Proof. The result follows by setting $\theta = \theta_{n+1}$ and $\theta' = \theta_n$ in (4.2) combined with (4.3). □

At this point, we seek to find iterative schemes satisfying (4.3). This leads us to our first corollary.

Corollary 27. *Assume (4.A1). Let $\alpha \in [0,1)$ and starting from an initial $\theta_1 \in \mathsf{T}$, let $(\theta_n)_{n\geqslant 1}$ be defined iteratively as follows*

$$\theta_{n+1} = \mathrm{argmax}_{\theta\in\mathsf{T}} \int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \log(k(\theta, y))\nu(\mathrm{d}y) \;, \quad n \geqslant 1 \;. \tag{4.4}$$

Then (4.3) holds and we can apply Theorem 12.

Proof. We have that (4.3) holds by definition of θ_{n+1} combined with the fact that $\alpha \in [0,1)$ and we can thus apply Theorem 12. □

Let us comment on Corollary 27. A remarkable aspect is that (4.4) is written as a maximisation problem involving the logarithm of the kernel k. This means that we can use (4.4) to derive simple update rules for $(\theta_n)_{n\geqslant 1}$ for some notable choices of kernel k, as illustrated in the following examples.

Example 8 (Gaussian distribution)**.** *We consider the case of a d-dimensional Gaussian density with $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ and where $\theta = (m, \Sigma) \in \mathsf{T}$ denotes the mean and covariance matrix of the Gaussian density. Then, starting from $\theta_1 = (m_1, \Sigma_1) \in \mathsf{T}$, solving (4.4) yields the following update formulas:*

$$\begin{aligned} \forall n \geqslant 1 \;, \quad m_{n+1} &= \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} y \; \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \\ \Sigma_{n+1} &= \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} (y - m_{n+1})(y - m_{n+1})^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \;. \end{aligned}$$

Example 9 (Student's distribution)**.** *We consider the case of a d-dimensional Student's density of the form $k(\theta, y) = \mathcal{T}(y; m, \Sigma, \nu)$, where $\theta = (m, \Sigma) \in \mathsf{T}$ denotes the mean and covariance matrix of the Student's density. Then, starting from $\theta_1 = (m_1, \Sigma_1) \in \mathsf{T}$, solving (4.4) yields the following update formulas:*

$$\begin{aligned} \forall n \geqslant 1 \;, \quad m_{n+1} &= \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) y \; \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) \nu(\mathrm{d}y)} \\ \Sigma_{n+1} &= \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) (y - m_{n+1})(y - m_{n+1})^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) \nu(\mathrm{d}y)} \;, \end{aligned}$$

where we have set $g^n(y) = (\nu + d)/(\nu + (y - m_n)^T (\Sigma_n)^{-1} (y - m_n))$ for all $y \in \mathsf{Y}$ and all $n \geqslant 1$.

Example 10 (Mean-field approximation)**.** *A generic member of the Mean-field variational family is $k(\theta, y) = \prod_{\ell=1}^L k^{(\ell)}(\theta^{(\ell)}, y^{(\ell)})$ with $\theta = (\theta^{(1)}, \ldots, \theta^{(L)}) \in \mathsf{T}$. Then, starting from $\theta_1 \in \mathsf{T}$, solving (4.4) yields the following update formulas: for all $n \geqslant 1$,*

$$\theta_{n+1}^{(\ell)} = \mathrm{argmax}_{\theta^{(\ell)}} \int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \log(k^{(\ell)}(\theta^{(\ell)}, y^{(\ell)}))\nu(\mathrm{d}y) \;, \quad 1 \leqslant \ell \leqslant L \;.$$

Interestingly, while Corollary 27 has a convenient formulation and corresponds to the intuitive choice so that (4.3) holds, it is also possible to derive alternative schemes satisfying (4.3) under additional smoothness conditions (see Section 4.A.1 for the definition of β-smoothness), as written in Corollary 28.

Corollary 28 ($\mathsf{T} = \mathbb{R}^d$)**.** *Assume (4.A1). Let $\alpha \in [0, 1)$, let $(\gamma_n)_{n\geqslant 1}$ be valued in $(0, 1]$ and let $(c_n)_{n\geqslant 1}$ be a positive sequence. Starting from an initial $\theta_1 \in \mathsf{T}$, let $(\theta_n)_{n\geqslant 1}$ be defined iteratively as follows*

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \nabla g_n(\theta)|_{\theta=\theta_n} , \quad n \geqslant 1 , \tag{4.5}$$

where $(g_n)_{n\geqslant 1}$ is the sequence of functions defined by: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$

$$g_n(\theta) = c_n \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \log\left(\frac{k(\theta, y)}{k(\theta_n, y)}\right) \nu(\mathrm{d}y) , \tag{4.6}$$

and g_n is assumed to be β_n-smooth. Then (4.3) holds and we can apply Theorem 12.

Proof. Since $\gamma_n \in (0, 1]$ and g_n is a β_n-smooth function by assumption, we can apply Lemma 37 and we obtain that for all $n \geqslant 1$,

$$g_n(\theta_n) - g_n\left(\theta_n - \frac{\gamma_n}{\beta_n} \nabla g_n(\theta)|_{\theta=\theta_n}\right) \geqslant \frac{\gamma_n}{2\beta_n} \|\nabla g_n(\theta)|_{\theta=\theta_n}\|^2 .$$

Thus, by definition of θ_{n+1} in (4.5), we have

$$0 = g_n(\theta_n) \geqslant g_n(\theta_{n+1}) ,$$

which in turn implies (4.3) and the proof is concluded. □

Let us now reflect on the implications of Corollary 28. Under common differentiability assumptions, we can write: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$

$$\nabla g_n(\theta) = c_n \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \nabla(\log k(\theta, y)) \nu(\mathrm{d}y) .$$

Then, considering the two cases where $c_n = 1$ and $c_n = (\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y))^{-1}$ at time n, (4.5) becomes respectively: for all $n \geqslant 1$,

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \nabla \log k(\theta, y)|_{\theta=\theta_n} \nu(\mathrm{d}y) , \tag{4.7}$$

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \left(\frac{1}{\alpha - 1} \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nabla \log k(\theta, y)|_{\theta=\theta_n} \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)}\right) . \tag{4.8}$$

Here, letting $p = p(\cdot, \mathscr{D})$, the iterative schemes (4.7) and (4.8) can both be seen as usual Gradient Descent iterations used to minimise $\theta \mapsto \Psi_\alpha(k(\theta, \cdot); \mathscr{D})$ and $\theta \mapsto -\mathcal{L}_\alpha(k(\theta, \cdot); \mathscr{D})$ with a learning policy proportional to $(\gamma_n \beta_n^{-1})_{n\geqslant 1}$.

This establishes the link between our approach and typical Gradient Descent algorithms for α-divergence and Renyi's α-divergence optimisation. Lastly, we give an example where the conditions on $(g_n)_{n\geqslant 1}$ from Corollary 28 are satisfied.

Example 11. *We consider the case of a d-dimensional Gaussian density with $k(\theta, y) = \mathcal{N}(y; \theta, \sigma^2 \boldsymbol{I_d})$ where $\theta \in \mathsf{T} = \mathbb{R}^d$ and $\sigma^2 > 0$ is assumed to be fixed. Then g_n as defined in (4.6) with $c_n = (\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y))^{-1}$ is convex and under usual differentiability assumptions*

$$\nabla g_n(\theta) = \frac{\sigma^{-2}}{\alpha - 1} \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} (y - \theta) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)}$$

so that by setting $\beta_n = \sigma^{-2}(1-\alpha)^{-1}$ and by denoting by $\|.\|$ the Euclidean norm, we can write for all $\theta, \theta' \in \mathsf{T}$ and all $n \geqslant 1$

$$\|\nabla g_n(\theta) - \nabla g_n(\theta')\| \leqslant \beta_n \|\theta - \theta'\| \, .$$

Hence, the conditions on $(g_n)_{n\geqslant 1}$ from Corollary 28 are satisfied and we obtain the iterative scheme given by: for all $n \geqslant 1$

$$\begin{aligned} \theta_{n+1} &= \theta_n + \gamma_n \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} (y - \theta_n) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \\ &= (1 - \gamma_n)\, \theta_n + \gamma_n \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha}\, y\, \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \, . \end{aligned}$$

The examples we have provided throughout the section underline the benefits of the approach we used in Theorem 12. However, the class of mixture models, which comes across as a very general and flexible parametric family, has yet to be included in our framework. In the next section we extend the monotonicity property to the case of mixture models.

4.3 Extension to mixture models

Given $J \in \mathbb{N}^\star$, we now consider the more general mixture model approximating family given by

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}, \Theta} k(y) = \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} \, ,$$

where we used the notation $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$ and $\mu_{\boldsymbol{\lambda}, \Theta} = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$ for all $\boldsymbol{\lambda} \in \mathcal{S}_J$ and all $\Theta \in \mathsf{T}^J$. We thus aim at solving

$$\inf_{\boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J} \Psi_\alpha(\mu_{\boldsymbol{\lambda}, \Theta} k) \, .$$

Notice in particular that the framework from Section 4.2 corresponds to having taken $J = 1$ in the optimisation problem above.

Let us denote $\boldsymbol{\lambda}_n = (\lambda_{j,n})_{1\leqslant j\leqslant J}$ and $\Theta_n = (\theta_{j,n})_{1\leqslant j\leqslant J}$ for all $n \geqslant 1$ and recall from Chapter 3 that we have also defined $\mathcal{S}_J^+ = \{\boldsymbol{\lambda} \in \mathcal{S}_J \,:\, \forall j \in \{1,\dots,J\},\ \lambda_j > 0\}$. For convenience, we also introduce the shorthand notation $\mu_n = \sum_{j=1}^J \lambda_{j,n}\delta_{\theta_{j,n}}$ and

$$\gamma_{j,\alpha}^n(y) = k(\theta_{j,n}, y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1} \tag{4.9}$$

for $\alpha \in [0,1)$, all $j = 1\dots J$, all $n \geqslant 1$ and all $y \in \mathsf{Y}$. The first step towards extending the approach of Section 4.2 to the case of mixture models is to generalise Proposition 26, which brings us to Proposition 29 below.

Proposition 29. *Assume (4.A1). For all $\alpha \in [0,1)$ and all $(\boldsymbol{\lambda}, \Theta), (\boldsymbol{\lambda}', \Theta') \in \mathcal{S}_J^+ \times \mathsf{T}^J$, it holds that*

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant \int_{\mathsf{Y}} \sum_{j=1}^J \frac{\lambda_j' k(\theta_j', y)}{\alpha - 1}\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^{\alpha-1} \log\left(\frac{\lambda_j}{\lambda_j'}\frac{k(\theta_j, y)}{k(\theta_j', y)}\right)\nu(\mathrm{d}y) + \Psi_\alpha(\mu_{\boldsymbol{\lambda}',\Theta'}k)\,. \tag{4.10}$$

Furthermore, equality holds in (4.10) if and only for all $j = 1\dots J$, $\lambda_j k(\theta_j, y) = \lambda_j' k(\theta_j', y)$ for ν-almost all $y \in \mathsf{Y}$.

Proof. By convexity of $\tilde{f}_\alpha$, Jensen's inequality implies

$$\begin{aligned}
\tilde{\Psi}_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) &= \int_{\mathsf{Y}} \tilde{f}_\alpha\left(\frac{\sum_{j=1}^J \lambda_j k(\theta_j, y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \\
&\leqslant \int_{\mathsf{Y}} \sum_{j=1}^J \frac{\lambda_j' k(\theta_j', y)}{\sum_{\ell=1}^J \lambda_\ell' k(\theta_\ell', y)} \tilde{f}_\alpha\left(\frac{\lambda_j k(\theta_j, y)}{p(y)\frac{\lambda_j' k(\theta_j', y)}{\sum_{\ell=1}^J \lambda_\ell' k(\theta_\ell', y)}}\right) p(y)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} \sum_{j=1}^J \frac{\lambda_j' k(\theta_j', y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \tilde{f}_\alpha\left(\frac{\lambda_j k(\theta_j, y)}{\lambda_j' k(\theta_j', y)}\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)\,.
\end{aligned} \tag{4.11}$$

We now treat the two cases $\alpha = 0$ and $\alpha \in (0,1)$ separately.

(a) Case $\alpha = 0$, with $\tilde{f}_0(u) = -\log(u)$ for all $u > 0$. In this case, (4.11) yields

$$\begin{aligned}
\tilde{\Psi}_0(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant &\int_{\mathsf{Y}} \sum_{j=1}^J \lambda_j' \times \frac{-k(\theta_j', y)p(y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \log\left(\frac{\lambda_j}{\lambda_j'}\frac{k(\theta_j, y)}{k(\theta_j', y)}\right)\nu(\mathrm{d}y) \\
&+ \int_{\mathsf{Y}} \sum_{j=1}^J \frac{\lambda_j' k(\theta_j', y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \times \left[-\log\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)\right] p(y)\nu(\mathrm{d}y)
\end{aligned}$$

which implies (4.10) since for all $y \in \mathsf{Y}$, $\sum_{j=1}^J \lambda_j' k(\theta_j', y)/\mu_{\boldsymbol{\lambda}',\Theta'}k(y) = 1$.

(b) Case $\alpha \in (0,1)$ with $\tilde{f}_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1\right]$ for all $u > 0$. In this setting, (4.11) gives

$$\begin{aligned}
\tilde{\Psi}_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant & \int_{\mathsf{Y}} \sum_{j=1}^{J} \frac{\lambda'_j k(\theta'_j, y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \frac{\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^\alpha \left[\left(\frac{\lambda_j}{\lambda'_j}\frac{k(\theta_j,y)}{k(\theta'_j,y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) \\
& + \int_{\mathsf{Y}} \sum_{j=1}^{J} \frac{\lambda'_j k(\theta'_j, y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \frac{\left[\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) \\
= & \int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda'_j k(\theta'_j, y) \left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^{\alpha-1} \tilde{f}_\alpha\left(\frac{\lambda_j}{\lambda'_j}\frac{k(\theta_j,y)}{k(\theta'_j,y)}\right) \nu(\mathrm{d}y) \\
& + \int_{\mathsf{Y}} \tilde{f}_\alpha\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \,, \qquad (4.12)
\end{aligned}$$

where we have used that for all $y \in \mathsf{Y}$, $\sum_{j=1}^{J} \lambda'_j k(\theta'_j, y)/\mu_{\boldsymbol{\lambda}',\Theta'}k(y) = 1$. Furthermore, recall from the proof of Proposition 29 that the concavity of the $\log$ function gives $\log(u^\alpha) \leqslant u^\alpha - 1$ for all $u > 0$ and since $\alpha \in (0,1)$, we can write

$$\frac{1}{\alpha-1}\log(u) = \frac{1}{\alpha(\alpha-1)}\log(u^\alpha) \geqslant \tilde{f}_\alpha(u) \,.$$

Thus, combining with (4.12) we deduce

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant \int_{\mathsf{Y}} \sum_{j=1}^{J} \frac{\lambda'_j k(\theta'_j, y)}{\alpha-1} \left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^{\alpha-1} \log\left(\frac{\lambda_j}{\lambda'_j}\frac{k(\theta_j,y)}{k(\theta'_j,y)}\right) \nu(\mathrm{d}y) + \Psi_\alpha(\mu_{\boldsymbol{\lambda}',\Theta'}k)$$

which establishes (4.10) for $\alpha \in (0,1)$.

As for the case of equality, equality in (4.10) implies equality in (4.11) which in turn by strict convexity of $\tilde{f}_\alpha$ implies the desired result and concludes the proof of Proposition 29. □

We can then state our second main theorem.

Theorem 13. *Assume (4.A1). Let $\alpha \in [0,1)$ and starting from an initial parameter set $(\boldsymbol{\lambda}_1, \Theta_1) \in \mathcal{S}_J^+ \times \mathsf{T}^J$, let $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$ be defined iteratively such that for all $n \geqslant 1$,*

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1} \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \leqslant 0 \qquad (4.13)$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1} \log\left(\frac{k(\theta_{j,n+1},y)}{k(\theta_{j,n},y)}\right) \nu(\mathrm{d}y) \leqslant 0 \,. \qquad (4.14)$$

Further assume that $\Psi_\alpha(\mu_1 k) < \infty$. Then, at time n, we have $\Psi_\alpha(\mu_{n+1}k) \leqslant \Psi_\alpha(\mu_n k)$.

Proof. The results follows immediately by setting $\theta = \theta_{n+1}$ and $\theta' = \theta_n$ in (4.10) combined with (4.13) and (4.14). □

We now plan on finding iterative schemes which satisfy (4.13) and (4.14). Strikingly, (4.13) does not depend on Θ_{n+1} nor does (4.14) depend on $\boldsymbol{\lambda}_{n+1}$. This means that we can treat these two inequalities separately and thus that the weights and components parameters of the mixture can be optimised simultaneously.

Observe also that the dependency in $\lambda_{j,n+1}$ appearing in (4.13) is simpler than the dependency in $\theta_{j,n+1}$ appearing in (4.14) and that is expressed through the kernel k. For this reason, we will first study (4.13). As we shall see, while the natural idea is to perform direct optimisation of the left-hand side of (4.13), a more general expression for the mixture weights can be derived, which will lead to numerical advantages later illustrated in Section 4.4.

4.3.1 Choice of $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$

In the following theorem, we identify an update formula which satisfies (4.13), regardless of the choice of the kernel k.

Theorem 14. *Assume (4.A1). Let $\alpha \in [0,1)$, let $(\eta_n)_{n\geqslant 1}$ be valued in $(0,1]$ and let κ be such that $(\alpha-1)\kappa \geqslant 0$. Starting from an initial parameter set $(\boldsymbol{\lambda}_1, \Theta_1) \in \mathcal{S}_J^+ \times \mathsf{T}^J$, let $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$ be defined iteratively such that for all $n \geqslant 1$*

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \gamma_{\ell,\alpha}^n(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}, \quad j = 1\dots J \tag{4.15}$$

and (4.14) is satisfied. Then (4.13) holds. Further assume that $\Psi_\alpha(\mu_1 k) < \infty$. Then, the two following assertions hold at iteration n.

(i) We have $\Psi_\alpha(\mu_{n+1}k) \leqslant \Psi_\alpha(\mu_n k)$.

(ii) Assuming that either $\{\eta_n = 1 \text{ and } \kappa < 0\}$ or $\{\eta_n \in (0,1)\}$, we have $\Psi_\alpha(\mu_{n+1}k) = \Psi_\alpha(\mu_n k)$ if and only if $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and for all $j = 1\dots J$, $k(\theta_{j,n+1}, y) = k(\theta_{j,n}, y)$ for ν-almost all $y \in \mathsf{Y}$.

Proof. Since (4.14) is assumed, it remains to show (4.13) so that we can apply Theorem 13, before characterising the case of equality. To prove (4.13), we treat the cases $\eta_n = 1$ and $\eta_n \in (0,1)$ separately.

(a) Case $\eta_n = 1$. Since $(\alpha-1)\kappa \geqslant 0$ with $\alpha \in (0,1)$, we have that

$$\kappa \sum_{j=1}^J \lambda_{j,n} \log(\lambda_j/\lambda_{j,n}) \geqslant 0$$

where we have used that $\sum_{j=1}^J \lambda_{j,n}\log(\lambda_j/\lambda_{j,n}) \leqslant \sum_{j=1}^J \lambda_{j,n}(\lambda_j/\lambda_{j,n} - 1) = 0$. In other words, to obtain (4.13) in the particular case $\eta_n = 1$, it is enough to show

$$\int_{\mathsf{Y}} \sum_{j=1}^J \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1} \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right)\nu(\mathrm{d}y) + \kappa \sum_{j=1}^J \lambda_{j,n}\log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \leqslant 0$$

that is

$$\sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \log \left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}} \right) \leqslant 0 \,. \tag{4.16}$$

Notice then that by definition of $(\lambda_{j,n+1})_{1 \leqslant j \leqslant J}$ when $\eta_n = 1$, we can write

$$\boldsymbol{\lambda}_{n+1} = \mathrm{argmin}_{\boldsymbol{\lambda} \in \mathcal{S}_J^+} \sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \log \left(\frac{\lambda_j}{\lambda_{j,n}} \right) \,.$$

[Indeed, setting $\beta_j = \lambda_{j,n} \left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^{n}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]$ and $\bar{\beta}_j = \beta_j / \sum_{\ell=1}^{J} \beta_\ell$ for all $j = 1 \dots J$, we have that $\sum_{j=1}^{J} \bar{\beta}_j \log \left(\bar{\beta}_j / \lambda_j \right) \geqslant 0$ and that this quantity is minimal when $\lambda_j = \bar{\beta}_j$ for $j = 1 \dots J$.] This implies (4.16) and settles the case $\eta_n = 1$.

(b) For the particular case $\eta_n \in (0, 1)$, we will use that for all $\epsilon > 0$ and all $u > 0$,

$$\log(u) = \frac{1}{\epsilon} \log(u^\epsilon) \geqslant \frac{1}{\epsilon} \left(1 - \frac{1}{u^\epsilon} \right) \,.$$

Indeed, since $\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \leqslant 0$ for all $j = 1 \dots J$, we can then write that for all $\epsilon > 0$,

$$\begin{aligned} &\sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \log \left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}} \right) \\ &\qquad \leqslant \frac{1}{\epsilon} \sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \left[1 - \left(\frac{\lambda_{j,n}}{\lambda_{j,n+1}} \right)^{\epsilon} \right] \,. \end{aligned} \tag{4.17}$$

Now notice that by definition of $(\lambda_{j,n+1})_{1 \leqslant j \leqslant J}$ we can write

$$\boldsymbol{\lambda}_{n+1} = \mathrm{argmin}_{\boldsymbol{\lambda} \in \mathcal{S}_J^+} \frac{1}{\epsilon} \sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \left[1 - \left(\frac{\lambda_{j,n}}{\lambda_j} \right)^{\epsilon} \right]$$

when ϵ satisfies $\eta_n = \frac{1}{1+\epsilon}$. [Indeed setting $\beta_j = \lambda_{j,n} \left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^{n}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\frac{1}{1+\epsilon}}$ and $\bar{\beta}_j = \beta_j / \sum_{\ell=1}^{J} \beta_\ell$ for all $j = 1 \dots J$, we have by convexity of the function $u \mapsto u^{1+\epsilon}$ that $\sum_{j=1}^{J} \left(\bar{\beta}_j / \lambda_j \right)^{1+\epsilon} \lambda_j \geqslant \left(\sum_{j=1}^{J} \bar{\beta}_j \right)^{1+\epsilon}$ and that this quantity is minimal when $\lambda_j = \bar{\beta}_j$ for $j = 1 \dots J$.] We then deduce that taking $\epsilon = \eta_n^{-1} - 1$ (it is always possible since $\eta_n \in (0, 1)$ by assumption) yields

$$\frac{1}{\epsilon} \sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^{n}(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \left[1 - \left(\frac{\lambda_{j,n}}{\lambda_{j,n+1}} \right)^{\epsilon} \right] \leqslant 0$$

which in turn yields (4.13) [since combined with (4.17) it implies (4.16) which itself implies (4.13) as seen in the case $\eta_n = 1$]. This settles the case $\eta_n \in (0, 1)$.

We can thus apply Theorem 13 and we obtain (i). As for the case of equality,

Theorem 13 implies that for all $j = 1 \dots J$, $\lambda_{j,n+1} k(\theta_{j,n+1}, y) = \lambda_{j,n} k(\theta_{j,n}, y)$ for ν-almost all $y \in \mathsf{Y}$. Since $\lambda_{j,1} > 0$ for all $j = 1 \dots J$, we also have $\lambda_{j,n} > 0$ for all $j = 1 \dots J$ under (4.A1). All that is left to do is thus to prove that $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ so that for all $j = 1 \dots J$, $k(\theta_{j,n+1}, y) = k(\theta_{j,n}, y)$ for ν-almost all $y \in \mathsf{Y}$.

Under the assumption that $\{\eta_n = 1 \text{ and } \kappa < 0\}$ equality in (4.16) implies that

$$\kappa \sum_{j=1}^{J} \lambda_{j,n} \log(\lambda_{j,n+1}/\lambda_{j,n}) = 0$$

i.e. that $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ by strict concavity of the $\log$ function. As for the case $\eta_n \in (0,1)$, equality in (4.17) and the strict concavity of the $\log$ function implies that $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$, which concludes the proof. □

Notice that as a byproduct of the proof of Theorem 14, the mixture weights update given by (4.15) can be rewritten under the form: for all $n \geqslant 1$

$$\boldsymbol{\lambda}_{n+1} = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathcal{S}_J^+} h_n(\boldsymbol{\lambda})$$

where, setting $\epsilon = \eta_n^{-1} - 1$, we have defined for all $\boldsymbol{\lambda} \in \mathcal{S}_J^+$,

$$h_n(\boldsymbol{\lambda}) = \begin{cases} \sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \log\left(\frac{\lambda_j}{\lambda_{j,n}} \right), & \text{if } \eta_n = 1 \ , \\ \frac{1}{\epsilon} \sum_{j=1}^{J} \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \nu(\mathrm{d}y) + \kappa \right] \left[1 - \left(\frac{\lambda_{j,n}}{\lambda_j} \right)^{\epsilon} \right], & \text{if } \eta_n \in (0,1) \ . \end{cases} \tag{4.18}$$

More specifically, $h_n(\boldsymbol{\lambda})$ acts as an upper bound of the left-hand side of (4.15) and we recover exactly the left-hand side of (4.15) in the particular case $\eta_n = 1$ and $\kappa = 0$.

Now that we have established Theorem 14, we are interested in deriving update formulas for the sequence $(\Theta_n)_{n \geqslant 1}$ satisfying (4.14).

4.3.2 Choice of $(\Theta_n)_{n \geqslant 1}$

We investigate three different approaches for choosing $(\Theta_n)_{n \geqslant 1}$.

4.3.2.1 A minimisation approach

The first idea is to consider the update for $(\Theta_n)_{n \geqslant 1}$ given by: for all $n \geqslant 1$,

$$\Theta_{n+1} = \operatorname{argmin}_{\Theta \in \mathsf{T}^J} g_n(\Theta)$$

where for all $\Theta \in \mathcal{S}_J^+ \times \mathsf{T}^J$,

$$g_n(\Theta) = \int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log\left(\frac{k(\theta_j, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \ . \tag{4.19}$$

In this case, the full update $(\boldsymbol{\lambda}_{n+1}, \Theta_{n+1})$ can be written as the following optimisation problem

$$(\boldsymbol{\lambda}_{n+1}, \Theta_{n+1}) = \operatorname{argmin}_{\boldsymbol{\lambda}\in\mathcal{S}_J^+, \Theta\in\mathsf{T}^J} \left(h_n(\boldsymbol{\lambda}) + g_n(\Theta)\right)$$

and we obtain Corollary 30.

Corollary 30. *Assume (4.A1). Let $\alpha \in [0,1)$, let $(\eta_n)_{n\geqslant 1}$ be valued in $(0,1]$ and let κ be such that $(\alpha-1)\kappa \geqslant 0$. Starting from an initial parameter set $(\boldsymbol{\lambda}_1, \Theta_1) \in \mathcal{S}_J^+ \times \mathsf{T}^J$, let $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$ be defined iteratively for all $n \geqslant 1$ by (4.15) and*

$$\theta_{j,n+1} = \operatorname{argmax}_{\theta_j\in\mathsf{T}} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \log(k(\theta_j, y))\nu(\mathrm{d}y) \;, \quad j = 1\dots J \;. \tag{4.20}$$

Then (4.14) holds and we can apply Theorem 14.

Proof. The result follows from the definition of Θ_{n+1} combined with the fact that $\alpha \in [0,1)$ and $\lambda_{j,n} > 0$ for all $j = 1\dots J$, so that (4.14) holds and we can apply Theorem 14. □

Consequently, under the assumptions of Corollary 30 we can define Algorithm 11, which leads to a systematic decrease in Ψ_α at each step and effectively generalises the monotonicity property from Corollary 27 to the case of mixture models. In line with Corollary 28, we next present another possible update formula for $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$.

Algorithm 11: Mixture models optimisation based on (4.20)

At iteration n**,**
For all $j = 1\dots J$, set

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \gamma_{\ell,\alpha}^n(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}$$

$$\theta_{j,n+1} = \operatorname{argmax}_{\theta_j\in\mathsf{T}} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \log(k(\theta_j, y))\nu(\mathrm{d}y) \;.$$

4.3.2.2 A Gradient Descent approach

We shall now resort to Gradient Descent steps to satisfy (4.13).

Corollary 31 ($\mathsf{T} = \mathbb{R}^d$)**.** *Assume (4.A1). Let $\alpha \in [0,1)$, let $(\eta_n)_{n\geqslant 1}$ be valued in $(0,1]$ and let κ be such that $(\alpha-1)\kappa \geqslant 0$. Furthermore, for all $j = 1\dots J$, let $(\gamma_{j,n})_{n\geqslant 1}$ be valued in $(0,1]$ and let $(c_{j,n})_{n\geqslant 1}$ be a positive sequence. Starting from an initial parameter set $(\boldsymbol{\lambda}_1, \Theta_1) \in \mathcal{S}_J^+ \times \mathsf{T}^J$, let $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$ be defined iteratively for all $n \geqslant 1$ by (4.15) and*

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} \;, \quad j = 1\dots J \;, \tag{4.21}$$

where for all $j = 1 \dots J$, $(g_{j,n})_{n \geqslant 1}$ is defined by: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$,

$$g_{j,n}(\theta) = c_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \; . \tag{4.22}$$

and $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth. Then (4.14) holds and we can apply Theorem 14.

Proof. Since $\gamma_{j,n} \in (0, 1]$ and $g_{j,n}$ is a $\beta_{j,n}$-smooth function by assumption, we can apply Lemma 37 and we obtain that for all $n \geqslant 1$ and all $j = 1 \dots J$,

$$g_{j,n}(\theta_{j,n}) - g_{j,n} \left(\theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} \right) \geqslant \frac{\gamma_{j,n}}{2\beta_{j,n}} \|\nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}\|^2 \; .$$

Thus, by definition of $\theta_{j,n+1}$ in (4.21), we have

$$0 = g_{j,n}(\theta_{j,n}) \geqslant g_{j,n}(\theta_{j,n+1}).$$

which in turn implies (4.14) so that we can apply Theorem 14. □

This gives us the monotonicity property for Algorithm 12 by Corollary 31 and we are now interested in possible choices for the constants $c_{j,n}$ appearing before $g_{j,n}$. Under common differentiability assumptions we can write: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$

$$\nabla g_{j,n}(\theta) = c_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \nabla \left(\log k(\theta, y) \right) \nu(\mathrm{d}y) \; , \quad j = 1 \dots J \; .$$

Algorithm 12: Mixture models optimisation based on (4.22)

At iteration n,
For all $j = 1 \dots J$, set

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \gamma_{\ell,\alpha}^n(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$
$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} \; .$$

As it turned out, the two most straightforward choices for $c_{j,n}$ correspond to taking $c_{j,n} = \lambda_{j,n}$ and $c_{j,n} = \lambda_{j,n} (\int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y))^{-1}$ for all $j = 1 \dots J$ and all $n \geqslant 1$. Indeed, letting $\gamma_{j,n} := \gamma_n \in (0, 1]$ and assuming that $\beta_{j,n}$ only depends on n for all $j = 1 \dots J$, that is $\beta_{j,n} := \beta_n$, the following update formulas ensue for Θ_{n+1} at iteration n:

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_n}{\beta_n} \lambda_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \nabla \log k(\theta, y)|_{\theta=\theta_{j,n}} \nu(\mathrm{d}y) \; , \quad j = 1 \dots J \; ,$$
$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_n}{\beta_n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nabla \log k(\theta, y)|_{\theta=\theta_{j,n}} \nu(\mathrm{d}y)}{(\alpha - 1) \int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \; , \quad j = 1 \dots J \; . \tag{4.23}$$

Letting $p = p(\cdot, \mathscr{D})$, we recognise usual Gradient Descent steps on Θ for minimising $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k; \mathscr{D})$ and $-\mathcal{L}_\alpha(\mu_{\boldsymbol{\lambda},\Theta}; \mathscr{D})$ using a learning policy proportional to $(\gamma_n\beta_n^{-1})_{n\geqslant 1}$.

An important point to take into consideration however is that by having performed a gradient step based on $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k; \mathscr{D})$ (resp. $-\mathcal{L}_\alpha(\mu_{\boldsymbol{\lambda},\Theta}; \mathscr{D})$), $\lambda_{j,n}$ now appears as a multiplicative factor by design in both updates. This is problematic since this could prevent learning in the algorithm for very small values of $\lambda_{j,n}$. Thankfully, we are able to circumvent this difficulty by choosing $c_{j,n} = (\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y))^{-1}$ so that we consider instead

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_n}{\beta_n} \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nabla \log k(\theta, y)|_{\theta=\theta_{j,n}} \nu(\mathrm{d}y)}{(\alpha-1)\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y)} , \quad j = 1 \dots J . \tag{4.24}$$

In this case, we are still in the framework of Corollary 31 and $\lambda_{j,n}$ only appears through $\mu_n k$, a property also shared with the update we introduced in Corollary 30. This further underlines the importance of having worked under the general conditions on $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$ stated in Theorem 13.

Finally, notice that the case where Θ_n is kept fixed at iteration n, that is, we solely optimise the mixture weights of a given mixture model, also maintains the monotonicity property. In fact, this particular case can be linked to the Power Descent (Daudel, Douc, and Portier, 2021) update formula for mixture models seen in Chapter 2.

4.3.2.3 A Power Descent approach

The Power Descent algorithm introduced in Chapter 2 is a gradient-based algorithm which operates on measures and performs α-divergence minimisation for all $\alpha \in \mathbb{R} \setminus \{1\}$. More precisely, denoting by $\mathrm{M}_1(\mathsf{T})$ the space of probability measures and letting $\mu \in \mathrm{M}_1(\mathsf{T})$, it seeks to optimise

$$\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

with respect to μ, where $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y)$ for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $y \in \mathsf{Y}$. Given an initial measure $\mu_1 \in \mathrm{M}_1(\mathsf{T})$, the optimisation is then done by applying several one-step transitions of the Power Descent algorithm:

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \quad n \geqslant 1 , \tag{4.25}$$

where, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, for all $\theta \in \mathsf{T}$,

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha - 1} - 1 \right] \nu(\mathrm{d}y)$$

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot [(\alpha - 1)(b_{\mu,\alpha}(\theta) + \kappa) + 1]^{\frac{\eta}{1-\alpha}}}{\mu([(\alpha - 1)(b_{\mu,\alpha} + \kappa) + 1]^{\frac{\eta}{1-\alpha}})} .$$

Observe then that by definition of $\gamma^n_{j,\alpha}$ in (4.9) and for $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_{j,n}}$ with $\Theta_n = \Theta$ and $\eta_n = \eta/(1-\alpha)$ at time n, (4.15) and (4.25) coincide.

Interestingly, the monotonicity property that has already been proved for the Power Descent algorithm in Theorem 1 of Chapter 2 uses a different proof technique compared to the one used in the proof of Theorem 14 to obtain that one transition of the Power Descent algorithm leads to a systematic decrease of Ψ_α for all $\alpha \in \mathbb{R} \setminus \{1\}$, for all $\eta \in (0, 1]$ and all κ such that $(\alpha - 1)\kappa \geqslant 0$.

This means that by maintaining Θ_n fixed and equal to a certain $\Theta \in \mathsf{T}$ in Theorem 14, it is possible to allow for a wider range of values of α and of $\eta_n = \eta/(1-\alpha)$ to be used while still preserving the monotonic decrease. In fact, we show a more general result in Proposition 32 below, where the results from Theorem 1 of Chapter 2 are further extended beyond the case $\eta > 1$ when $\alpha < 0$.

Proposition 32. *Assume that p and k are as in (4.A1). Let (α, η) belong to any of the following cases.*

(i) $\alpha \leqslant -1$ *and* $\eta \in (0, (\alpha - 1)/\alpha]$;

(ii) $\alpha \in (-1, 0)$ *and* $\eta \in (0, 1 - \alpha]$;

(iii) $\alpha \in [0, 1)$ *or* $\alpha > 1$ *and* $\eta \in (0, 1]$.

Moreover, let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu k) < \infty$ and let κ be such that $(\alpha - 1)\kappa \geqslant 0$. Then, the two following assertions hold.

(i) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$.

(ii) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

The proof of this result is deferred to Section 4.A.2 and we now make two comments. Firstly, while the results from Chapter 2 and Proposition 32 allow for a wider range of values for α and η to be used, the strong improvement of Chapter 4 is that by Theorem 14 we do not need to keep Θ_n constant anymore at each step of the algorithm. From there, extending Theorem 14 beyond the case $\alpha \in [0, 1)$ and $\eta_n \in (0, 1]$ is an interesting direction of research, which is left for future work.

Secondly, by connecting the Power Descent to (4.15), we now have a better understanding of the role of the parameter η_n appearing in (4.15). Indeed, as underlined in earlier chapters of this thesis, the Power Descent algorithm belongs to a more

general family of gradient-based algorithms which includes the Entropic Mirror Descent algorithm, a typical optimisation algorithm for optimisation under simplex constraints. Viewed from this angle, the parameter η_n can be understood as a learning rate applied to $b_{\mu_n,\alpha}$, the gradient of Ψ_α. This aspect will notably come in handy when interpreting our numerical experiments in Section 4.4.

We have derived several examples where the conditions of Theorem 14 are met and connected this theorem to the Power Descent algorithm. We will conclude this section by presenting relevant particular cases of Algorithm 11. We start by investigating the case where the kernel k belongs to the Gaussian family.

4.3.3 Algorithm 11 within the Gaussian family

We consider the case of d-dimensional Gaussian mixture densities with $k(\theta_j, y) = \mathcal{N}(y; m_j, \Sigma_j)$ and where $\theta_j = (m_j, \Sigma_j) \in \mathsf{T}$ denotes the mean and covariance matrix of the j-th Gaussian component density. Then, solving (4.20), that is

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta_j \in \mathsf{T}} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \log(k(\theta_j, y)) \nu(\mathrm{d}y) \ , \quad j = 1 \dots J$$

yields the following update formulas at time n for the means $(m_{j,n+1})_{1 \leqslant j \leqslant J}$ and covariances matrices $(\Sigma_{j,n+1})_{1 \leqslant j \leqslant J}$:

$$\forall j = 1 \dots J, \quad m_{j,n+1} = \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) y \ \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y)} \tag{4.26}$$

$$\Sigma_{j,n+1} = \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y)} \ . \tag{4.27}$$

Due to the intractable integrals appearing in (4.15), (4.26), and (4.27), we shall then use approximate update rules in practice. Many choices are possible here and for simplicity we will restrict ourselves to using a sequence of samplers $(q_n)_{n \geqslant 1}$ and performing typical Adaptive Importance Sampling estimation in order to approximate (4.15), (4.26), and (4.27). This leads to Algorithm 13 below, where based on (4.9) we have defined for all $j = 1 \dots J$, all $y \in \mathsf{Y}$ and all $n \geqslant 1$,

$$\hat{\gamma}_{j,\alpha}^n(y) = \frac{k(\theta_{j,n}, y)}{q_n(y)} \left(\frac{\mu_n k(y)}{p(y)} \right)^{\alpha - 1} \ .$$

Algorithm 13: Gaussian Mixture Models optimisation with (4.20)

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1\leqslant m\leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J}\lambda_{\ell,n}\left[\sum_{m=1}^{M}\hat{\gamma}_{\ell,\alpha}^{n}(Y_{m,n}) + (\alpha-1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = \frac{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})\cdot Y_{m,n}}{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})}$$

$$\Sigma_{j,n+1} = \frac{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})\cdot(Y_{m,n}-m_{j,n+1})(Y_{m,n}-m_{j,n+1})^{T}}{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \,.$$

We have thus obtained a tractable version of Algorithm 11 which allows us to iteratively update both the weights and components parameters of a Gaussian mixture model by optimising the α-divergence between the mixture distribution and the targeted distribution. We now make two remarks.

Remark 33. *A practical version of Algorithm 11 can be derived in the particular case of Student's distributions, which could be useful for robustification purposes (see Algorithm 15 in Section 4.A.3).*

Remark 34. *We can obtain practical versions of Algorithm 12 by considering the case of d-dimensional Gaussian mixture densities with $k(\theta_j, y) = \mathcal{N}(y; \theta_j, \sigma^2 \boldsymbol{I_d})$ where $\Theta \in \mathsf{T}^J$ with $\mathsf{T} = \mathbb{R}^d$ and $\sigma^2 > 0$ is assumed to be fixed. In this case, $g_{n,j}$ is convex for all $j = 1 \dots J$ and all $n \geqslant 1$.*

Following (4.23) and letting $c_{j,n} = \lambda_{j,n}(\int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y))^{-1}$ in the definition of $g_{j,n}$ permits to choose $\beta_{j,n} = \sigma^{-2}(1-\alpha)^{-1}$ [using that $\int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y) = \sum_{j=1}^{J}\int_{\mathsf{Y}} \lambda_{j,n}\gamma_{j,\alpha}^{n}(y)\nu(\mathrm{d}y)$]. This gives the update formula at iteration n below

$$\theta_{j,n+1} = \theta_{j,n} + \gamma_n \frac{\int_{\mathsf{Y}} \lambda_{j,n}\gamma_{j,\alpha}^{n}(y)(y-\theta_{j,n})\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)} \,, \quad j = 1 \dots J \,.$$

In addition, following (4.24) and letting $c_{j,n} = (\int_{\mathsf{Y}} \gamma_{j,\alpha}^{n}(y)\nu(\mathrm{d}y))^{-1}$ in the definition of $g_{j,n}$ also permits to choose $\beta_{j,n} = \sigma^{-2}(1-\alpha)^{-1}$ so that the update formula at iteration n is

$$\theta_{j,n+1} = (1-\gamma_n)\,\theta_{j,n} + \gamma_n \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^{n}(y)\; y\; \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^{n}(y)\nu(\mathrm{d}y)} \,, \quad j = 1 \dots J \,,$$

which coincides with (4.26) when $\gamma_n = 1$. Approximated versions of the two above iterative formulas are then given respectively by

$$\forall j = 1 \dots J\,, \quad \theta_{j,n+1} = \theta_{j,n} + \gamma_n \frac{\lambda_{j,n} \sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^{J} \sum_{m=1}^{M} \lambda_{j,n} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \tag{4.28}$$

$$\theta_{j,n+1} = (1 - \gamma_n)\, \theta_{j,n} + \gamma_n \frac{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \tag{4.29}$$

and tractable versions of Algorithm 12 for Gaussian mixture models can be deduced, as written in Algorithm 14.

Algorithm 14: Gaussian Mixture Models optimisation with (4.28)/(4.29)

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1 \leqslant m \leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\sum_{m=1}^{M} \hat{\gamma}_{\ell,\alpha}^{n}(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$\theta_{j,n+1} = \begin{cases} \theta_{j,n} + \gamma_n \frac{\lambda_{j,n} \sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^{J} \sum_{m=1}^{M} \lambda_{j,n} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} & (4.28) \\ (1 - \gamma_n)\theta_{j,n} + \gamma_n \frac{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} & (4.29)\,. \end{cases}$$

Lastly, we focus on the particular case $\alpha = 0$ in Algorithm 11 (and its application to the particular case of Gaussian Mixture Models as seen in Algorithm 13). As we shall see, this case can be linked to the M-PMC algorithm and it will be used to drive our numerical experiments.

4.3.4 The M-PMC algorithm as a particular case of Algorithm 11

We are interested in interpreting the results we have obtained thus far in the light of the M-PMC algorithm (Cappé et al., 2008). To do so, we first recall the basics of the M-PMC algorithm. For any measurable positive function p on $(\mathsf{Y}, \mathcal{Y})$, the M-PMC algorithm aims at solving the optimisation problem

$$\sup_{(\boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J)} \int_{\mathsf{Y}} \log \left(\sum_{j=1}^{J} \lambda_j k(\theta_j, y) \right) p(y) \nu(\mathrm{d}y)\,, \tag{4.30}$$

or equivalently, using a Variational Inference formulation, at minimising the Reverse Kullback-Leibler

$$\inf_{(\boldsymbol{\lambda}\in\mathcal{S}_J,\Theta\in\mathsf{T}^J)} D_0(\mu_{\boldsymbol{\lambda},\Theta}K||\mathbb{P}) \ ,$$

where for all $A \in \mathcal{Y}$, $\mathbb{P}(A) = \int_A p(y)\nu(\mathrm{d}y)/\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)$. This is done in Cappé et al., 2008, Section 2 by introducing the following iterative update formulas for all $j = 1 \ldots J$ and for all $n \geqslant 1$

$$\lambda_{j,n+1} = \int_{\mathsf{Y}} \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \frac{p(y)}{\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)}\nu(\mathrm{d}y) \tag{4.31}$$

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta_j\in\mathsf{T}} \int_{\mathsf{Y}} \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \log(k(\theta_j,y))p(y)\nu(\mathrm{d}y) \ . \tag{4.32}$$

Observing then that the two update formulas above correspond to having considered the particular case $\alpha = 0$, $\eta_n = 1$ and $\kappa = 0$ in Algorithm 11, it follows that the M-PMC algorithm can be seen as a particular example of our framework.

Remark 35. *Interestingly, equations* (4.31) *and* (4.32) *are presented in Cappé et al., 2008 as integrated versions under the target distribution of the update formulas for the Expectation-Maximisation (EM) algorithm applied to the mixture-density parameter estimation problem*

$$\sup_{(\boldsymbol{\lambda}\in\mathcal{S}_J,\Theta\in\mathsf{T}^J)} \sum_{m=1}^M \log\left(\sum_{j=1}^J \lambda_j k(\theta_j,Y_m)\right) \ .$$

Hence, we can interpret Algorithm 11 as a generalisation of an integrated EM algorithm preserving the monotonicity property and extending it to the case $\alpha \in [0,1)$.

A practical version of the M-PMC algorithm has been introduced in Cappé et al., 2008, Section 3 for the particular case of the Gaussian family, in which they use the sampler

$$q_n(y) = \mu_n k(y) = \sum_{j=1}^J \lambda_{j,n}k(\theta_{j,n},y) \ . \tag{4.33}$$

Thus, comparing Algorithm 13 to the original M-PMC algorithm for Gaussian Mixture Models from Cappé et al., 2008, Section 3, we do not yet specify the sequence of samplers $(q_n)_{n\geqslant 1}$ and now include additional choices for the sequence of learning rates $(\eta_n)_{n\geqslant 1}$, the parameter α and the constant κ. This has important practical consequences which we illustrate in our following numerical experiments.

4.4 Numerical Experiments: Multimodal Target

In our numerical experiments, we are interested in seeing how the choice of the sequence of samplers $(q_n)_{n\geqslant 1}$, the sequence of learning rates $(\eta_n)_{n\geqslant 1}$, the constant

κ and the choice of α influence the convergence of Algorithm 13. We use a similar setting to the one considered in Cappé et al., 2008. The target p is a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant c such that

$$p(y) = c \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})] \ ,$$

where $\boldsymbol{u_d}$ is the d-dimensional vector whose coordinates are all equal to 1, $s = 2$, $c = 2$ and $\boldsymbol{I_d}$ is the identity matrix. For all $A \in \mathcal{Y}$, we also denote $\mathbb{P}(A) = c^{-1} \int_A p(y)\nu(\mathrm{d}y)$.

Numerical Experiment 1: study of the particular case $\alpha = 0$. We take $J = 100$, $M = 200$, $d = 16$, $N = 100$ such that the total computational budget is $N \times M = 20000$ samples in Algorithm 13 with $\alpha = 0$ and we will vary the sequence of learning rates $(\eta_n)_{1 \leqslant n \leqslant N}$, the constant $\kappa \leqslant 0$ as well as the choice of the sampler.

We generate the initial parameter set for the means of the mixture distribution by sampling from a centered normal distribution with covariance matrix $5\boldsymbol{I_d}$ and we set their associated initial weights to $[1/J, \ldots, 1/J]$ (i.e. $\boldsymbol{\lambda}_1 = [1/J, \ldots, 1/J]$ at time $n = 1$). For simplicity, we chose to keep the covariance matrices fixed equal to $\sigma^2 \boldsymbol{I_d}$ with $\sigma^2 = 1$ and to only update the means and the mixture weights. Furthermore, we consider a constant policy for the sequence of learning rates $(\eta_n)_{1 \leqslant n \leqslant N}$ with $\eta_n := \eta$ for all $n = 1 \ldots N$.

As for the choice of sampler at time n, we are first interested in setting q_n as in (4.33), since this sampler is the best approximation to the targeted density we know of at time n (in terms of Reverse Kullback-Leibler) and it is also the one used in the M-PMC algorithm from Cappé et al., 2008. We denote the resulting algorithm M-PMC(η, κ), the case $(\eta, \kappa) = (1, 0)$ corresponding to the initial M-PMC algorithm of Cappé et al., 2008.

We let $\eta \in \{1, 0.5, 0.2, 0.1\}$, $-\kappa \in \{0, 0.1, 1\}$ and we replicate the experiment 200 times independently for the M-PMC(η, κ) algorithm. To assess the convergence, note that since we have sampled M samples from q_n at time n, these samples can readily be used to obtain an estimate $\hat{c} = M^{-1} \sum_{m=1}^{M} p(Y_{m,n})/q_n(Y_{m,n})$ of the normalising constant $c = \int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)$ with no additional computational cost.

Then, as we can see on Figure 4.1, the choice of η and of κ does impact the convergence of the algorithm. Notably, for a fixed κ, choosing $\eta < 1$ results in improved numerical results in the estimation of the normalising constant c.

This can be explained by the stochastic nature of the approximation that appears in the update formula for the mixture weights of Algorithm 13. Recall from Section 4.3.2.3 that performing our mixture weights update corresponds to applying one transition of the Power Descent algorithm: since this algorithm is known to share similarities with gradient-based algorithms, choosing $\eta_n = 1$ might not be the

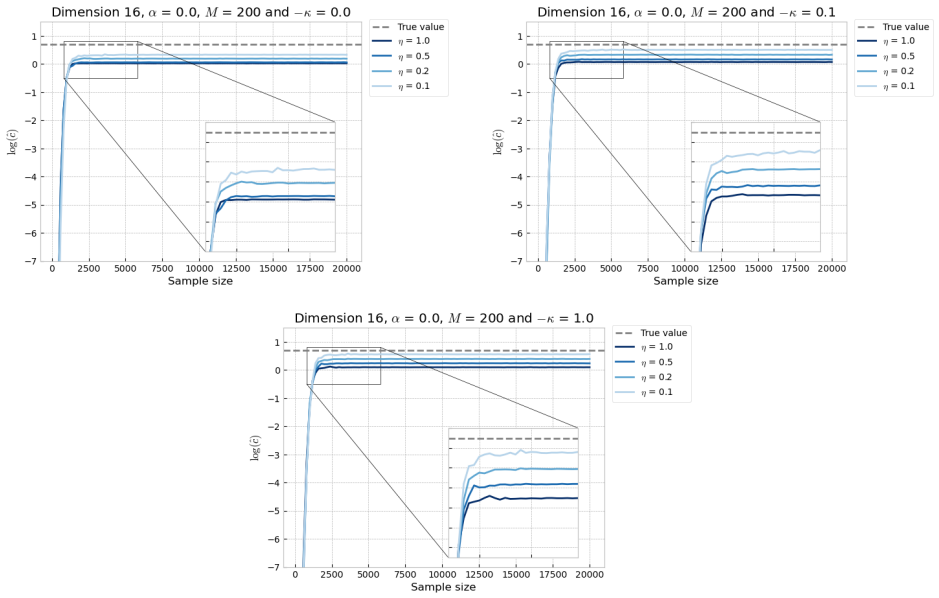


FIGURE 4.1: Normalisation constant estimation by the M-PMC(η, κ) algorithm in dimension $d = 16$ for $\eta \in \{1, 0.5, 0.2, 0.1\}$ and $-\kappa \in \{0, 0.1, 1\}$.

best course of action in practice when we resort to approximations [much like choosing a learning rate equal to 1 in a Stochastic Gradient Descent scheme might not be the best choice in general].

Similarly, for a fixed $\eta < 1$, choosing $-\kappa > 0$ leads to improved numerical results. The idea behind this is that by adding a positive constant $-\kappa$, we enforce the positivity of the mixture weights throughout the algorithm. This is handy in practice to avoid setting some mixture weights to zero, which could for example be an unfortunate consequence of having taken a learning large that is too large or having used a sampler q_n which is very different from the targeted density in the early stages.

We have thus seen that by changing the values of η and of κ, we are able to improve on the initial M-PMC algorithm of Cappé et al., 2008 for which $(\eta, \kappa) = (1, 0)$. Next, we are interested in using at time n a uniform sampler of the form

$$q_n(y) = J^{-1} \sum_{j=1}^{J} k(\theta_{j,n}, y) \; .$$

This is motivated by the fact that based on the form of the integrals appearing in (4.15), (4.26), and (4.27), we would like to sample according to $k(\theta_{j,n}, y)$ when updating the parameters λ_j, m_j and Σ_j. This could easily become computationally expensive as J increases, which is why we consider a uniform sampler as a cheaper alternative.

We call the resulting algorithm UM-PMC(η, κ) and we now want to compare it to the M-PMC(η, κ). To do so, we will use the Mean-Squared Error at time n for each algorithm denoted MSE, which is computed as the average of $\|m_{\text{approx},n} - m_{\text{true}}\|^2$

over 200 independent runs of the algorithm.

Here, $\|.\|$ stands for the Euclidian norm, $m_{\text{true}} = \mathbb{E}_{\mathbb{P}}[Y]$ for the mean of the targeted density and $m_{\text{approx},n}$ for the mean of the approximating density at time n (in our setting $m_{\text{true}} = 0.\boldsymbol{u_d}$ and $m_{\text{approx},n} = \sum_{j=1}^{J} \lambda_{j,n} m_{j,n}$). The logMSE (logarithm of the MSE) can be visualised on Figure 4.2 below.

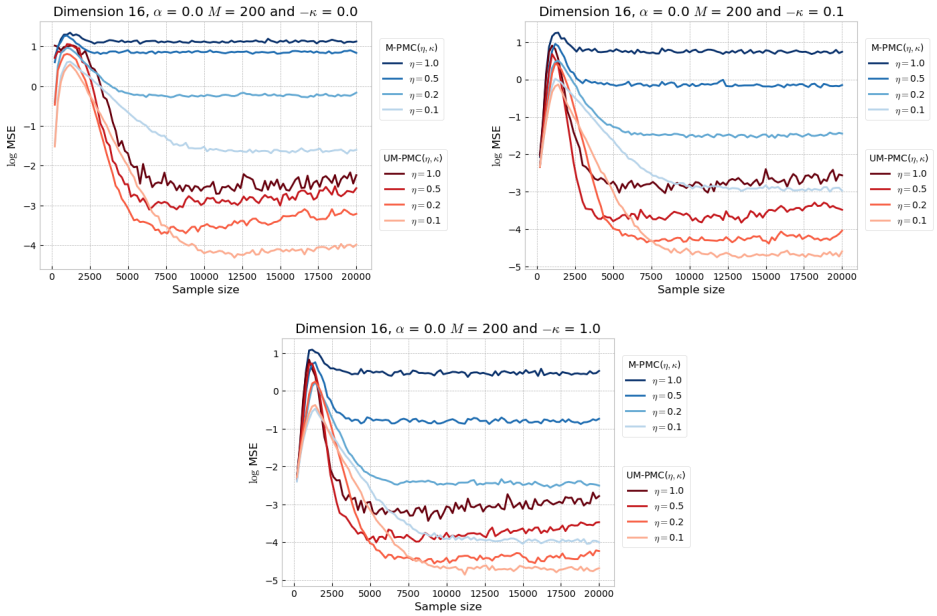


FIGURE 4.2: LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms in dimension $d = 16$ for $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$.

Notice then that for a relatively small number of samples M at each time n (here $M = 200$), the UM-PMC(η, κ) algorithm generally outperforms the M-PMC(η, κ) algorithm in terms of Mean-Squared Error, the latter one being more prone to missing one of the two modes, especially for larger values of η.

This means that the results of the M-PMC(η, κ) algorithm are more sensitive to the number of samples M used. As we increase the number of samples M, it can however be observed that the performances of the M-PMC(η, κ) algorithm in terms of Mean-Squared Error are improved and become comparable to those of the UM-PMC(η, κ) algorithm (see Section 4.A.4 for additional plots when $M = \{500, 1000\}$).

We now move on to our second numerical experiment in which we are interested in varying the parameter α.

Numerical Experiment 2: effect of α. We let $\alpha \in \{0, 0.5\}$ and our goal in this numerical experiment will be to estimate $m_{\text{true}} = \mathbb{E}_{\mathbb{P}}[Y]$, which is a typical Bayesian Inference task.

We take $J = 100$, $d = 16$, $M = \{200, 500\}$ and N such that the total computational budget is $N \times M = 20000$ samples in Algorithm 13. The initial parameter set is

generated exactly like in *Numerical Experiment 1*. Based on our previous numerical results, we focus mainly on the UM-PMC(η, κ) algorithm, even though we will in addition run the experiment for the M-PMC$(1., 0.)$ algorithm, which corresponds to the M-PMC algorithm from Cappé et al., 2008.

As for the covariance matrices, they are kept fixed equal to $\sigma^2 \boldsymbol{I_d}$ so that we only update the means and the mixture weights and this time we let $\sigma^2 \in \{1, 4\}$ to investigate how the variance of the kernel impacts the convergence according to the value of α. We consider yet again a constant policy for all $1 \leqslant n \leqslant N$ with $\eta_n := \eta = 0.1$ and we let $-\kappa = 0.1$, as it appears to be a good tradeoff in terms of hyperparameters.

Note that the results from Remark 34 apply for this choice of covariance matrices, that is it is also possible to perform gradient-descent steps for Renyi's α divergence minimisation when updating the means, as defined in (4.28) (see Algorithm 14 for the description of the full algorithm). We will then run the experiment with $\gamma_n := \gamma = 1$ at iteration n. For a fair comparison, we will use a uniform sampler and take the same hyperparameter as those used the UM-PMC(η, κ) algorithm. The resulting algorithm is denoted RGD(η, κ).

We use the Parallel Interacting Markov AIS (PIMAIS) algorithm from Martino et al., 2017 as a reference algorithm to compare our results with. Indeed, this algorithm also approximate the targeted density by a mixture model. More precisely, it alternates between two steps: (1) a parameter update step where the means of each kernel is updated via several MH transitions (2) an Importance Sampling step providing weighted particles which are then used to estimate the desired quantity (in our case $\mathbb{E}_{\mathbb{P}}[Y]$).

In the PIMAIS algorithm, we then employ the MH algorithm with a Gaussian proposal with covariance matrix $\sigma^2_{\text{MH}} \boldsymbol{I_d}$ with $\sigma^2_{\text{MH}} \in \{1, 25\}$ to construct the Markov chains. We consider a mixture of J Gaussians with covariance matrices $\sigma^2 \boldsymbol{I_d}$ and a deterministic number of samples M/J is drawn from each mixand at time n, so that this algorithm uses the same computational power as those we present.

Finally, M additional samples are generated at time n to estimate $\mathbb{E}_{\mathbb{P}}[Y]$ following the PIMAIS methodology which gives the estimator $\hat{m}^{\text{PIMAIS}}_{\text{approx},n}$ (we refer to Martino et al., 2017 for more details on how this estimator is obtained). As for the UM-PMC(η, κ) algorithm (resp. the M-PMC$(1., 0.)$ and the RGD(η, κ) algorithms), we too generate M additional samples and we consider at time $n = 1 \dots N$ the Importance Sampling estimator of $\mathbb{E}_{\mathbb{P}}[Y]$ given by

$$\hat{m}_{\text{approx},n} = \sum_{n'=1}^{n} \sum_{m=1}^{M} w_{m,n'} Y'_{m,n'}$$

where $(Y'_{m,n'})_{1\leqslant m\leqslant M}$ have been generated independently from $\mu_{n'}k$ at time $n' = 1\dots n$ and where for all $n' = 1\dots n$ and all $m = 1\dots M$, we have defined

$$w_{m,n'} \propto \frac{p(Y'_{m,n'})}{\mu_{n'}k(Y'_{m,n'})} \quad \text{such that} \quad \sum_{n'=1}^{n}\sum_{m=1}^{M} w_{m,n'} = 1\,.$$

We replicate the experiment 200 times independently for all the algorithms. To assess the performance of the different algorithms, we consider the Mean-Squared Error at time n denoted $\widehat{\mathrm{MSE}}$, which is computed as the average of $\|\hat{m}_{\mathrm{approx},n} - m_{\mathrm{true}}\|^2$ over 200 independent runs of our algorithms (resp. $\|\hat{m}^{\mathrm{PIMAIS}}_{\mathrm{approx},n} - m_{\mathrm{true}}\|^2$ for the PIMAIS algorithm). The Log$\widehat{\mathrm{MSE}}$ (logarithm of the $\widehat{\mathrm{MSE}}$) can then be visualised on Figure 4.3 below.

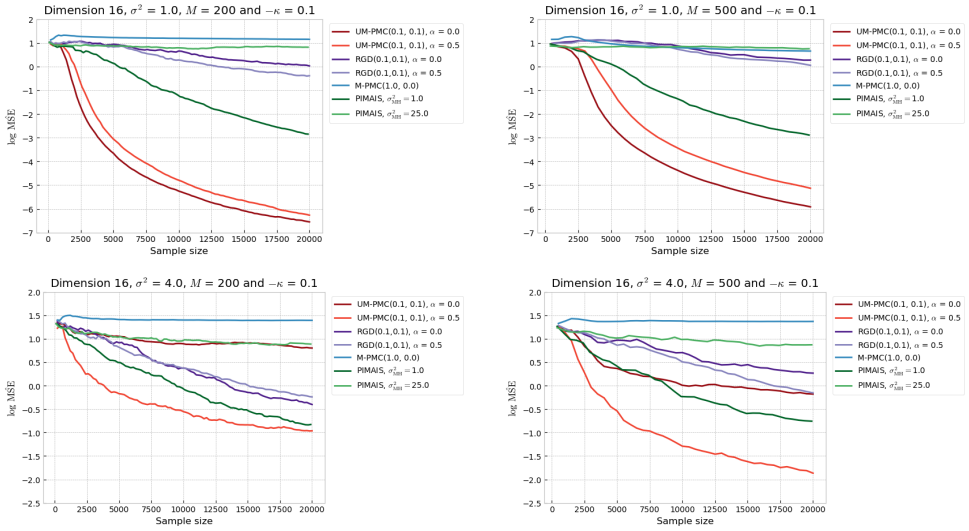


FIGURE 4.3: Log$\widehat{\mathrm{MSE}}$ for UM-PMC(η,κ) and RGD(η,κ) in dimension $d = 16$ for $\alpha \in \{0., 0.5\}$, $\sigma^2 \in \{1, 4\}$, $\eta = 0.1$ and $-\kappa = 0.1$ compared with the PIMAIS algorithm and the M-PMC(1., 0.) algorithm.

Observe that for $\sigma^2 = 1$, all the versions of the UM-PMC(η,κ) algorithm considered outperform the PIMAIS algorithm in terms of Log$\widehat{\mathrm{MSE}}$ and that the case $\alpha = 0$ yields the best result. Notice also that since in this case the covariance matrix is well-tailored to the problem, increasing the number of samples from $M = 200$ to $M = 500$ slows down the UM-PMC(η,κ) algorithm.

As for the case $\sigma^2 = 4$, we obtain this time that the case $\alpha = 0.5$ performs the best and that the case $\alpha = 0$ underperforms compared to the PIMAIS algorithm with $\sigma^2_{\mathrm{MH}} = 1$ (even though it still outperforms the PIMAIS algorithm with $\sigma^2_{\mathrm{MH}} = 25$). This underlines the importance of having provided a framework which goes beyond the typical case of the reverse Kullback-Leibler with $\alpha = 0$. Unsurprisingly, since we have now considered a less favourable value for σ^2 with $\sigma^2 = 4$, increasing the sample size results in improved results.

Moreover, observe that the RGD(η, κ) algorithm underperforms in this numerical experiment. As already mentioned in Section 4.3.2.2, this is due to the fact that $\lambda_{j,n}$ appears by design as a multiplicative factor in the update formula for the means. This prevents learning when the algorithm produces small values for $\lambda_{j,n}$, a pitfall avoided by the UM-PMC(η, κ) algorithm. Finally, note that the M-PMC($1., 0.$) algorithm performs poorly in all four cases considered in Figure 4.3, which further illustrates how we were able to successfully improve on this algorithm introduced in Cappé et al., 2008 by including it into a wider framework.

4.5 Conclusion and perspectives

We introduced a novel methodology to carry out α-divergence minimisation via an iterative algorithm ensuring a monotonic decrease in the α-divergence at each step. Notably, our framework allows us to perform simultaneous updates for both the weights and components parameters of a given mixture model for all $\alpha \in [0, 1)$.

We then underlined the links between our approach and Gradient Descent algorithms for α-divergence minimisation and connected our results to the Power Descent algorithm. We also presented practical algorithms for Gaussian mixture models parameters optimisation and recovered the M-PMC algorithm as a particular case of our framework. Finally, we provided empirical evidence that our methodology can be used to enhance the M-PMC algorithm and Gradient Descent-based algorithms. As a result we achieved better performances compared to the PIMAIS algorithm and shed light on the importance of having some flexibility in the choice of α.

To conclude, we state several directions to extend our work in Chapter 4. First of all, now that we have established a systematic decrease for our iterative schemes, the next step is to derive convergence rates and to compare them with those obtained using typical Gradient Descent schemes. Based on the results from Proposition 32, another interesting direction consists in generalising the monotonicity property from Theorem 14 beyond the case $\alpha \in [0, 1)$. Lastly, we also expect that resorting to more advanced Monte Carlo methods in the estimation of the intractable integrals appearing in (4.15), (4.26), and (4.27) will result in further improved numerical results.

4.A Deferred results

4.A.1 Quantifying the improvement in one step of Gradient Descent

Here, $\langle \cdot, \cdot \rangle$ denotes an inner product defined on $\mathsf{T} \times \mathsf{T}$ with corresponding norm $\|.\|$. Typically, we will consider $\mathsf{T} = \mathbb{R}^d$ with $d \geqslant 1$, so that $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^d$ and $\|.\|$ is the Euclidian norm.

Definition 36. *A continuously differentiable function g defined on T is said to be β-smooth if for all $\theta, \theta' \in \mathsf{T}$,*

$$\|g(\theta) - g(\theta')\| \leqslant \beta \|\theta - \theta'\| \, .$$

Lemma 37. *Let $\gamma \in (0, 1]$, let g be a β-smooth function defined on $\mathsf{T} = \mathbb{R}^d$. Then, for all $\theta \in \mathsf{T}$ it holds that*

$$g(\theta) - g\left(\theta - \frac{\gamma}{\beta} \nabla g(\theta)\right) \geqslant \frac{\gamma}{2\beta} \|\nabla g(\theta)\|^2 \, .$$

Proof. By assumption on g, we have that for all $\theta, \theta' \in \mathsf{T}$

$$g(\theta') - g(\theta) - \langle \nabla g(\theta), \theta' - \theta \rangle \leqslant \frac{\beta}{2} \|\theta' - \theta\|^2 \, .$$

In particular, setting $\theta' = \theta - \frac{\gamma}{\beta} \nabla g(\theta)$ yields

$$\begin{aligned} g(\theta) - g\left(\theta - \frac{\gamma}{\beta} \nabla g(\theta)\right) &\geqslant \frac{\gamma}{\beta} \|\nabla g(\theta)\|^2 - \frac{\gamma^2}{2\beta} \|\nabla g(\theta)\|^2 \\ &\geqslant \frac{\gamma}{\beta} \left(1 - \frac{\gamma}{2}\right) \|\nabla g(\theta)\|^2 \, . \end{aligned}$$

Since $\gamma \in (0, 1]$, we can deduce the desired result, that is

$$g(\theta) - g\left(\theta - \frac{\gamma}{\beta} \nabla g(\theta)\right) \geqslant \frac{\gamma}{2\beta} \|\nabla g(\theta)\|^2 \, .$$

□

4.A.2 Monotonicity property for the Power Descent

Preliminary remarks First note that for convenience in the proofs, we redefine in this section the function $b_{\mu,\alpha}(\theta)$ for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$ by

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \tilde{f}'_\alpha \left(\frac{\mu k(y)}{p(y)}\right)^{\alpha - 1} \nu(\mathrm{d}y)$$

Then, for all $\eta > 0$, the iteration $\mu \mapsto \mathcal{I}_\alpha(\mu)$ is well-defined if we have

$$0 < \mu(|b_{\mu,\alpha} + \kappa|^{\frac{\eta}{1-\alpha}}) < \infty \, . \tag{4.34}$$

Furthermore, Theorem 1 of Chapter 2 already established that one transition of the Power Descent algorithm ensures a monotonic decrease in the α-divergence at each step for all $\eta \in (0, 1]$ and all κ such that $(\alpha - 1)\kappa \geqslant 0$ under the assumption of Proposition 32, which settles the case (iii).

Finally, while we establish our results for (i) and (ii) in the general case where $\mu \in \mathrm{M}_1(\mathsf{T})$, the particular case of mixture models follows immediately by choosing μ as a weighted sum of dirac measures.

Extending the monotonicity

Let (ζ, μ) be a couple of probability measures where ζ is dominated by μ, which we denote by $\zeta \preceq \mu$. A first lower-bound for the difference $\Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$ was derived in Chapter 2 and was used to establish that the Power Descent algorithm diminishes Ψ_α for all $\eta \in (0, 1]$.

We now prove a novel lower-bound for the difference $\Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$ which will allow us to extend the monotonicity results from Chapter 2 beyond the case $\eta \in (0, 1]$ when $\alpha < 0$. This result relies on the existence of an exponent ϱ satisfying condition (4.A2) below, which will later on be used to specify a range of values for η ensuring that Ψ_α is decreasing after having applied one transition $\mu \mapsto \mathcal{I}_\alpha(\mu)$

(4.A2) We have $\varrho \in \mathbb{R} \setminus [0, 1]$ and the function $f_{\alpha,\varrho} : u \mapsto \tilde{f}_\alpha(u^{1/\varrho})$ is non-decreasing and concave on $\mathbb{R}_{>0}$.

Proposition 38. *Assume (4.A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, assume that ϱ satisfies (4.A2) and let κ be such that $(\alpha - 1)\kappa \geqslant 0$. Then, for all $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_{\mu,\alpha}|) < \infty$ and $\zeta \preceq \mu$,*

$$|\varrho|^{-1} \left\{ \mu(|b_{\mu,\alpha} + \kappa|) - \mu\left(|b_{\mu,\alpha} + \kappa| g^\varrho\right) \right\} \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k) \,, \tag{4.35}$$

where g is the density of ζ wrt μ, i.e. $\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta)$. Moreover, equality holds in (4.35) if and only if $\zeta = \mu$.

Proof. First note that for all $\alpha \in \mathbb{R} \setminus \{1\}$, we have by (4.A2) that $f'_{\alpha,\varrho}(u) \geqslant 0$ for all $u > 0$, and thus that $sg(\varrho) = sg(\alpha - 1)$ where $sg(v) = 1$ if $v \geqslant 0$ and -1 otherwise. Since $sg(f'_\alpha(u)) = sg(\alpha - 1) = sg(\kappa)$ for all $u > 0$, this implies that $\varrho^{-1} \tilde{f}'_\alpha(u) = |\varrho|^{-1} |\tilde{f}'_\alpha(u)|$, $\varrho^{-1}\kappa = |\varrho^{-1}\kappa|$ and finally that $\varrho^{-1}(b_{\mu,\alpha}(\theta) + \kappa) = |\varrho^{-1}||b_{\mu,\alpha}(\theta) + \kappa|$ for all $\theta \in \mathsf{T}$, which will be used later in the proof.

Write by definition of $f_{\alpha,\varrho}$ in (4.A2) and ζ,

$$\begin{aligned}
\tilde{\Psi}_\alpha(\zeta k) &= \int_{\mathsf{Y}} \tilde{f}_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} f_{\alpha,\varrho}\left(\left[\frac{\zeta k(y)}{p(y)}\right]^\varrho\right) p(y)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} f_{\alpha,\varrho}\left(\left[\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta, y)}{\mu k(y)} \left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\right]^\varrho\right) p(y)\nu(\mathrm{d}y) \\
&\leqslant \int_{\mathsf{Y}} f_{\alpha,\varrho}\left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta, y)}{\mu k(y)} \left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^\varrho\right) p(y)\nu(\mathrm{d}y)
\end{aligned} \tag{4.36}$$

where the last inequality follows from Jensen's inequality applied to the convex function $u \mapsto u^\varrho$ (since $\varrho \in \mathbb{R} \setminus [0,1]$) and the fact that $f_{\alpha,\varrho}$ is non-decreasing. Now set

$$u_y = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} \left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^\varrho$$
$$v_y = \left(\frac{\mu k(y)}{p(y)} \right)^\varrho$$

and note that

$$u_y - v_y = \left(\frac{\mu k(y)}{p(y)} \right)^\varrho \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} g^\varrho(\theta) - 1 \right) \tag{4.37}$$

Since $f_{\alpha,\varrho}$ is concave, $f_{\alpha,\varrho}(u_y) \leqslant f_{\alpha,\varrho}(v_y) + f'_{\alpha,\varrho}(v_y)(u_y - v_y)$. Then, combining with (4.36), we get

$$\begin{aligned} \tilde{\Psi}_\alpha(\zeta k) &\leqslant \int_{\mathsf{Y}} f_{\alpha,\varrho}(u_y) p(y) \nu(\mathrm{d}y) \\ &\leqslant \int_{\mathsf{Y}} f_{\alpha,\varrho}(v_y) p(y) \nu(\mathrm{d}y) + \int_{\mathsf{Y}} f'_{\alpha,\varrho}(v_y)(u_y - v_y) p(y) \nu(\mathrm{d}y) \end{aligned} \tag{4.38}$$

Note that the first term of the rhs can be written as

$$\int_{\mathsf{Y}} f_{\alpha,\varrho}(v_y) p(y) \nu(\mathrm{d}y) = \int_{\mathsf{Y}} \tilde{f}_\alpha \left(\frac{\mu k(y)}{p(y)} \right) p(y) \nu(\mathrm{d}y) = \tilde{\Psi}_\alpha(\mu k) \tag{4.39}$$

Using now $f'_{\alpha,\varrho}(v_y) = \varrho^{-1} v_y^{1/\varrho - 1} \tilde{f}'_\alpha(v_y^{1/\varrho})$ and (4.37), the second term of the rhs of (4.38) may be expressed as

$$\begin{aligned} &\int_{\mathsf{Y}} f'_{\alpha,\varrho}(v_y)(u_y - v_y) p(y) \nu(\mathrm{d}y) \\ &\quad = \varrho^{-1} \int_{\mathsf{Y}} \left(\frac{\mu k(y)}{p(y)} \right)^{1-\varrho} \tilde{f}'_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \\ &\qquad\qquad \left(\frac{\mu k(y)}{p(y)} \right)^\varrho \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} g^\varrho(\theta) - 1 \right) p(y) \nu(\mathrm{d}y) \\ &\quad = \varrho^{-1} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta,y) \tilde{f}'_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y) \right) g^\varrho(\theta) \\ &\qquad\qquad - \varrho^{-1} \int_{\mathsf{Y}} \mu k(y) \tilde{f}'_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y) \\ &\quad = \varrho^{-1} \left\{ \mu\left(b_{\mu,\alpha} \cdot g^\varrho \right) - \mu(b_{\mu,\alpha}) \right\} \\ &\quad = |\varrho|^{-1} \left\{ \mu\left(|b_{\mu,\alpha} + \kappa| g^\varrho \right) - \mu(|b_{\mu,\alpha} + \kappa|) \right\} + |\varrho^{-1}\kappa|(1 - \mu(g^\varrho)) \,, \end{aligned}$$

where we have used that $\varrho^{-1}(b_{\mu,\alpha}(\theta) + \kappa) = |\varrho^{-1}||b_{\mu,\alpha}(\theta) + \kappa|$ for all $\theta \in \mathsf{T}$ and that $\varrho^{-1}\kappa = |\varrho^{-1}\kappa|$. In addition, Jensen's inequality applied to the convex function

$u \mapsto u^\varrho$ implies that $\mu(g^\varrho) \geqslant 1$ and thus

$$\int_{\mathsf{Y}} f'_{\alpha,\varrho}(v_y)(u_y - v_y)p(y)\nu(\mathrm{d}y) \leqslant |\varrho|^{-1}\left\{\mu\left(|b_{\mu,\alpha}+\kappa|g^\varrho\right) - \mu(|b_{\mu,\alpha}+\kappa|)\right\} . \tag{4.40}$$

Combining this inequality with (4.38) and (4.39) finishes the proof of the inequality. Furthermore, if the equality holds in (4.35), then the equality in Jensen's inequality (4.40) shows that g is constant μ-a.e. so that $\zeta = \mu$, and the proof is completed. □

Remark 39. *The proof of Proposition 38 relies on $\tilde{f}'_\alpha$ being of constant sign, which is why we used $\tilde{f}_\alpha$ in the proof instead of f_α.*

We now plan on setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Proposition 38 and obtain that one iteration of the Power Descent yields $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$. For this purpose and based on the upper bound obtained in Proposition 38, we strengthen the condition (4.34) as follows to take into account the exponent ϱ

$$0 < \mu(|b_{\mu,\alpha}+\kappa|^{\frac{\eta}{1-\alpha}}) < \infty \text{ and } \mu(|b_{\mu,\alpha}+\kappa|g^\varrho) \leqslant \mu(|b_{\mu,\alpha}+\kappa|) \quad \text{with } g = \frac{|b_{\mu,\alpha}+\kappa|^{\frac{\eta}{1-\alpha}}}{\mu(|b_{\mu,\alpha}+\kappa|^{\frac{\eta}{1-\alpha}})} . \tag{4.41}$$

This leads to the following result.

Proposition 40. *Assume (4.A1). Let $\alpha \in \mathbb{R}\setminus\{1\}$, assume that ϱ satisfies (4.A2) and let κ be such that $(\alpha-1)\kappa \geqslant 0$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\mu(|b_{\mu,\alpha}|) < \infty$ and assume that η satisfies (4.41). Then, the two following assertions hold.*

(i) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$.

(ii) We have $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Proof. We treat the case $\kappa = 0$ in the proof below (the case $\kappa \neq 0$ unfolds similarly). We apply Proposition 38 with $\zeta = \mathcal{I}_\alpha(\mu)$ so that $\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta)g(\theta)$ with $g = |b_{\mu,\alpha}|^{\eta/(1-\alpha)}/\mu(|b_{\mu,\alpha}|^{\eta/(1-\alpha)})$. Then,

$$\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k) + |\varrho|^{-1}\left\{\mu\left(|b_{\mu,\alpha}|g^\varrho\right) - \mu(|b_{\mu,\alpha}|)\right\} \leqslant \Psi_\alpha(\mu k) \tag{4.42}$$

where the last inequality follows from condition (4.41).

Let us now show (ii). The *if* part is obvious. As for the *only if* part, $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ combined with (4.42) yields

$$\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k) + |\varrho|^{-1}\left\{\mu\left(|b_{\mu,\alpha}|g^\varrho\right) - \mu(|b_{\mu,\alpha}|)\right\} ,$$

which is the case of equality in Proposition 38. Therefore, $\mathcal{I}_\alpha(\mu) = \mu$. □

While Proposition 40 resembles Theorem 1 of Chapter 2 in its formulation and in the properties on the iteration $\mu \mapsto \mathcal{I}_\alpha(\mu)$ it establishes, it is important to note that the proof techniques used, and thus the conditions on η obtained, are different.

This brings us to the proof of Proposition 32. The proof of this theorem requires intermediate results, which are derived in Section 4.A.2 alongside with the proof of Proposition 32.

Proof of Proposition 32

For the sake of readability, we only treat the case $\kappa = 0$ in the proofs below (and the case $\kappa \neq 0$ unfolds similarly). In Proposition 38, the difference $\Psi_\alpha(\zeta k) - \Psi_\alpha(\mu k)$ is split into two terms

$$\Psi_\alpha(\zeta k) - \Psi_\alpha(\mu k) = A(\mu, \zeta) + |\varrho|^{-1} \left\{ \mu \left(|b_{\mu,\alpha}| g^\varrho \right) - \mu(|b_{\mu,\alpha}|) \right\} \ ,$$

where $g = \mathrm{d}\zeta/\mathrm{d}\mu$. Moreover, Proposition 38 states that $A(\mu, \zeta)$ is always non-positive.

It turns out that the second term is minimal over all positive probability densities g when it is proportional to $|b_{\mu,\alpha}|^{1/(1-\varrho)}$, as we show in Lemma 41 below.

Lemma 41. *Let $\varrho \in \mathbb{R} \setminus [0, 1]$. Then, for any positive probability density g w.r.t μ, we have*

$$\mu \left(|b_{\mu,\alpha}| g^\varrho \right) \geqslant \left[\mu \left(|b_{\mu,\alpha}|^{1/(1-\varrho)} \right) \right]^{1-\varrho} \ ,$$

with equality if and only if $g \propto |b_{\mu,\alpha}|^{1/(1-\varrho)}$.

Proof. The function $x \mapsto x^{1-\varrho}$ is strictly convex for $\varrho \in \mathbb{R} \setminus [0, 1]$. Thus Jensen's inequality yields, for any positive probability density g w.r.t. μ,

$$\mu \left(|b_{\mu,\alpha}| g^\varrho \right) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\frac{|b_{\mu,\alpha}(\theta)|^{1/(1-\varrho)}}{g(\theta)} \right)^{1-\varrho} g(\theta) \geqslant \left[\mu \left(|b_{\mu,\alpha}|^{1/(1-\varrho)} \right) \right]^{1-\varrho} \tag{4.43}$$

which finishes the proof of the inequality. The next statement follows from the case of equality in Jensen's inequality: g must be proportional to $|b_{\mu,\alpha}|^{1/(1-\varrho)}$. □

The next lemma shows that this choice leads to a non-positive second term, thus implying that $\Psi_\alpha(\zeta k) \leqslant \Psi_\alpha(\mu k)$.

Lemma 42. *Assume (4.A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$ and assume that ϱ satisfies (4.A2). Then $\eta = (1 - \alpha)/(1 - \varrho)$ satisfies (4.41) for any $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_{\mu,\alpha}|) < \infty$.*

Proof. We apply (4.43) with $g = 1$ and get that

$$\left[\mu \left(|b_{\mu,\alpha}|^{1/(1-\varrho)} \right) \right]^{1-\varrho} \leqslant \mu(|b_{\mu,\alpha}|) < \infty \ . \tag{4.44}$$

Then (4.41) can be readily checked with $\eta = (1-\alpha)/(1-\varrho)$. Set $\phi = \eta/(1-\alpha)$. Using that $\mu(|b_{\mu,\alpha}|) < \infty$ when $\phi < 0$ and (4.A1) for $\phi > 0$, we obtain $\mu(|b_{\mu,\alpha}|^\phi) > 0$, which concludes the proof. □

While Lemma 42 seems to advocate for $g = \mathrm{d}\zeta/\mathrm{d}\mu$ to be proportional to $|b_{\mu,\alpha}|^{1/(1-\varrho)}$, notice that this choice of g might not be optimal to minimize $\Psi_\alpha(\zeta k) - \Psi_\alpha(\mu k)$, as $A(\mu, \zeta)$ also depends on g through ζ. In the next lemma, we thus propose another choice of the tuning parameter η, which also satisfies (4.41) for any $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_{\mu,\alpha}|) < \infty$.

Lemma 43. *Assume (4.A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$ and assume that ϱ satisfies (4.A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\mu(|b_{\mu,\alpha}|) < \infty$. Assume in addition that $|\varrho| \geqslant 1$, then $\eta = (\alpha-1)/\varrho$ satisfies (4.41).*

Proof. Setting $g \propto |b_{\mu,\alpha}|^{-1/\varrho}$, we get

$$\mu(|b_{\mu,\alpha}|g^\varrho) = \mu(|b_{\mu,\alpha}|^{1-\varrho/\varrho})[\mu(|b_{\mu,\alpha}|^{-1/\varrho})]^{-\varrho} = [\mu(|b_{\mu,\alpha}|^{-1/\varrho})]^{-\varrho} \leqslant \mu(|b_{\mu,\alpha}|)$$

where the last inequality follows from Jensen's inequality applied to the convex function $u \mapsto u^{-\varrho}$ (since $|\varrho| \geqslant 1$). Since $\mu(|b_{\mu,\alpha}|) < \infty$, the parameter $\eta = (\alpha - 1)/\varrho$ satisfies (4.41). Set $\phi = \eta/(1 - \alpha)$. Using that $\mu(|b_{\mu,\alpha}|) < \infty$ when $\phi < 0$ and (4.A1) for $\phi > 0$, we obtain $\mu(|b_{\mu,\alpha}|^\phi) > 0$, which concludes the proof. □

Lemma 42 and Lemma 43 allow us to define a range of values for η that decreases Ψ_α after one transition of the Power Descent, under the assumption that ϱ satisfies (4.A2). Now, in order to prove Proposition 32 and given $\alpha \in \mathbb{R} \setminus \{1\}$, we need to check which values of ϱ satisfy the conditions expressed in (4.A2).

Proof of Proposition 32. The proof consists in verifying that we can apply Proposition 40, that is, given $\alpha \in \mathbb{R} \setminus \{1\}$, we must find a range of constants ϱ which satisfy (4.A2). We then use Lemma 42 or Lemma 43 to deduce that, for the provided constants η, (4.41) holds.

(i) Assumption (4.A2) holds for all $\varrho < 0$, with $f_{\alpha,\varrho}(u) = -\log(u)/\varrho$. Moreover, by definition of $b_{\mu,\alpha}$, we get for all $n \geqslant 1$,

$$\mu(|b_{\mu,\alpha}|) = \int_{\mathsf{Y}} \mu k(y) \frac{p(y)}{\mu k(y)} \nu(\mathrm{d}y) = \int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty \,.$$

Combining with Lemma 42 and Lemma 43, (4.41) holds for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and for any $\eta \in (0, 1]$.

(ii) Observing that for $\alpha \notin \{0, 1\}$,

$$f_{\alpha,\varrho}(u) = \frac{1}{\alpha(\alpha-1)} \left(u^{\alpha/\varrho} - 1\right) \,,$$

we get that (4.A2) holds for $\varrho \leqslant \alpha$ if $\alpha < 0$ Lemmas 42 and 43 provide the corresponding ranges for η in Cases (i) and (ii). To finish the proof, we now show that for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $\mu(|b_{\mu,\alpha}|)$ is finite, so that Lemmas 42 and 43 can indeed be applied.

Since $u\tilde{f}'_\alpha(u) = \alpha\tilde{f}_\alpha(u) + 1/(\alpha-1)$, we have, for all $n \geqslant 1$,

$$\begin{aligned}\mu(|b_{\mu,\alpha}|) &= \int_{\mathsf{Y}} \left|\left(\frac{\mu k(y)}{p(y)}\right)\tilde{f}'_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) \\ &\leqslant |\alpha| \int_{\mathsf{Y}} \left|\tilde{f}_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) + \frac{1}{|\alpha-1|}\end{aligned} \tag{4.45}$$

Using that $\Psi_\alpha(\mu k) > -\infty$ (which is a consequence of (4.A1) and of Jensen's inequality applied to the convex function $u \mapsto uf_\alpha(1/u)$), the r.h.s is finite if and only if $\Psi_\alpha(\mu k)$ is finite, which is what we have assumed and thus the proof is finished.

□

4.A.3 Algorithm 11 updates for the Student's family

We consider the case of d-dimensional Student's mixture densities of the form $k(\theta_j, y) = \mathcal{T}(y; m_j, \Sigma_j, \nu_j)$, where $\theta_j = (m_j, \Sigma_j)$ denotes the mean and covariance matrix of the j-th Student's component density. Then, based on Example 9, solving

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta_j \in \mathsf{T}} \int_{\mathsf{Y}} \lambda_{j,n}\gamma^n_{j,\alpha}(y)\log(k(\theta_j, y))\nu(\mathrm{d}y)\ , \quad j = 1\dots J$$

yields the following update formulas

$$\begin{aligned}\forall j = 1\dots J, \quad m_{j,n+1} &= \frac{\int_{\mathsf{Y}} \gamma^n_{j,\alpha}(y) g^n_j(y) y\ \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma^n_{j,\alpha}(y) g^n_j(y)\nu(\mathrm{d}y)} \\ \Sigma_{j,n+1} &= \frac{\int_{\mathsf{Y}} \gamma^n_{j,\alpha}(y) g^n_j(y)(y - m_{j,n+1})(y - m_{j,n+1})^T\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma^n_{j,\alpha}(y) g^n_j(y)\nu(\mathrm{d}y)}\ ,\end{aligned}$$

where for all $y \in \mathsf{Y}$ and for all $j = 1\dots J$, we have set

$$g^n_j(y) = \frac{\nu_j + d}{\nu_j + (y - m_{j,n})^T(\Sigma_{j,n})^{-1}(y - m_{j,n})}\ .$$

Based on Algorithm 11 and given a sequence of samplers $(q_n)_{n\geqslant 1}$, one may consider in practice Algorithm 15 below.

Algorithm 15: α-divergence minimisation for Student's Mixture Models

At iteration n**,**

1. Draw independently M samples $(Y_{m,n})_{1\leqslant m\leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J}\lambda_{l,n}\left[\sum_{m=1}^{M}\hat{\gamma}_{\ell,\alpha}^{n}(Y_{m,n}) + (\alpha-1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = \frac{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})g_j^n(Y_{m,n})\cdot Y_{m,n}}{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})g_j^n(Y_{m,n})}$$

$$\Sigma_{j,n+1} = \frac{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})g_j^n(Y_{m,n})\cdot(Y_{m,n}-m_{j,n+1})(Y_{m,n}-m_{j,n+1})^T}{\sum_{m=1}^{M}\hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})g_j^n(Y_{m,n})}.$$

4.A.4 Additional numerical experiments

In this section we provide further plots based on the numerical experiments from Section 4.4.

Numerical Experiment 1 when $M \in \{500, 1000\}$**.**
We work within the same framework as the one from Numerical Experiment 1 except that we now take $M \in \{500, 1000\}$ samples at each step n while keeping the total computational budget equal to $N \times M = 20000$ samples. The experiment is repeated 200 times independently for each algorithm considered and the results are plotted on Figure 4.4 and Figure 4.5 below.

Observe that as M increases, the performances of the UM-PMC(η, κ) algorithm are improved and become comparable to the one of the M-PMC(η, κ) algorithm, especially for smaller values of η.

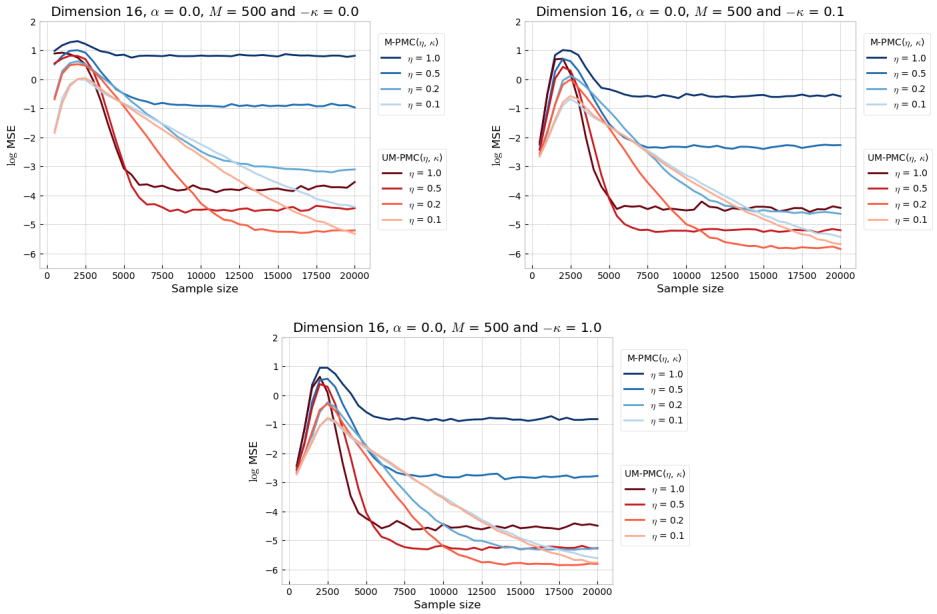


FIGURE 4.4: $\underline{M = 500.}$ LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms with $d = 16$, $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$.

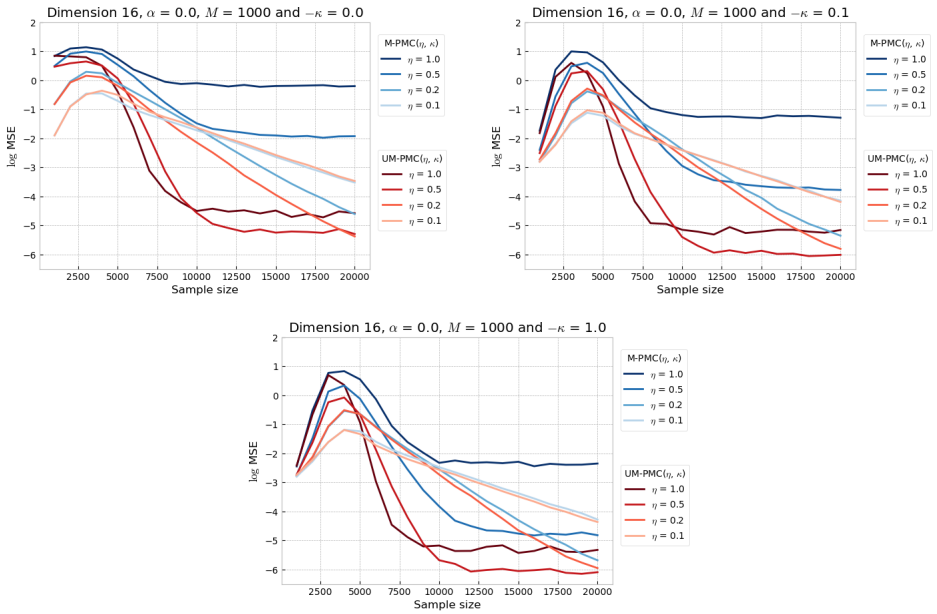


FIGURE 4.5: $\underline{M = 1000.}$ LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms with $d = 16$, $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$.

5

Conclusion

In this thesis, we have been interested in pairing up Monte Carlo and Variational Inference methods in order to develop improved adaptive procedures that enable complex modelling of large-scale datasets in a Bayesian setting.

Since Variational Inference methods are commonly limited by the choice of the measure of dissimilarity and of the approximating family, we proposed in our work to go beyond the traditional framework of forward Kullback-Leibler divergence minimisation with a parametric family.

More specifically, we considered the α-divergence as a wider class of objective functions. From there, we designed efficient algorithms that ensure a monotonic decrease in the α-divergence at each step and that permit us to extend the approximating family in two main ways: one, by putting a prior on the variational parameter in the form of a measure and two, by working within the broad family of mixture models.

Our approach is very general, in the sense that it recovers several algorithms from the optimisation literature. In particular, it improves on Entropic Mirror Descent schemes in terms of convergence rates and makes Gradient Descent steps compatible with mixture weights updates for mixture models optimisation. Furthermore, it includes an integrated Expectation-Maximisation algorithm from the Monte Carlo community, further shedding light on the connections between optimisation-based and sampling-based methodologies. Finally, it allows us to propose novel, and theoretically-sound, algorithms that empirically yield better results compared to these aforementioned algorithms.

To conclude, we will share some possible future directions of research based on the work carried out in this thesis.

A first direction would be to establish additional theoretical results regarding our algorithms. For example, one may seek to prove a convergence rate for the Power Descent when $\alpha < 1$. One could also investigate convergence results for our mixture components parameters optimisation procedures, before attempting to study the case where the mixture weights and the mixture components parameters are optimised together. As our algorithms rely on Monte Carlo approximations in practice, another possibility is to look into convergence results in the stochastic case. Technical challenges that must be overcome to advance in that first direction will then notably include (i) the uncommon form of the Power Descent updates in comparison with usual optimisation techniques and (ii) the non-convexity of the objective function once we start taking into account the mixture components parameters updates.

A second direction would focus on numerical aspects. In order to improve on our empirical results and for variance reduction purposes, one may want to resort to more advanced Monte Carlo methods in our integral estimations. In addition, figuring out which approach is most successful between unbiased (e.g. Power Descent) and biased approaches (e.g. Renyi Descent) for α-divergence minimisation remains an important question in the literature and so is the choice of the hyperparameter α. On a more open-ended note, one could try to derive new well-chosen functions Γ that satisfy our assumptions for the (α, Γ)-descent and to propose supplementary choices of exploration steps.

A

Appendices for Chapter 1

A.1 Detailed derivations for Example 2

• Letting $\alpha \in \mathbb{R} \setminus \{0, 1\}$, we first show that the optimal update for a certain $\ell \in \{1, \dots, L\}$ while keeping the other components fixed is given by

$$q_\ell^\star(y_\ell) \propto \left(\mathbb{E}_{-\ell} \left[p(y, \mathscr{D})^{1-\alpha} \prod_{k=1, k\neq\ell}^{L} q_k(y_k)^{\alpha-1} \right] \right)^{\frac{1}{1-\alpha}} , \quad \text{for } \nu_\ell \text{ almost all } y \in \mathsf{Y}_\ell .$$

To see this, observe that optimising (1.10) is equivalent to optimising

$$\frac{1}{\alpha(\alpha-1)} \int_{\mathsf{Y}} q(y)^\alpha p(y, \mathscr{D})^{1-\alpha} \nu(\mathrm{d}y) .$$

Since q is assumed to belong to the Mean-field family, we can write

$$\frac{1}{\alpha(\alpha-1)} \int_{\mathsf{Y}} q(y)^\alpha p(y, \mathscr{D})^{1-\alpha} \nu(\mathrm{d}y) = \frac{1}{\alpha(\alpha-1)} \int_{\mathsf{Y}_\ell} q_l(y_l)^\alpha \mathbb{E}_{-\ell} \left[\left(\frac{p(y, \mathscr{D})}{\prod_{k=1, k\neq\ell}^{L} q_k(y_k)} \right)^{1-\alpha} \right] \nu_\ell(\mathrm{d}y_\ell)$$

Now maintaining all other components fixed except ℓ, we deduce by Jensen's inequality combined with the strict convexity of the convex function $u \mapsto \frac{1}{\alpha(\alpha-1)} u^{1-\alpha}$ when $u > 0$ that the optimal update is given by $q_\ell^\star$.

• Letting $\alpha \in (0, 1)$, plugging in the model from Example 2 and assuming that q_ℓ follows a Gaussian distribution with parameters m_ℓ and Λ_ℓ, i.e. $q(y) = \mathcal{N}(y; m_\ell, \Lambda_\ell)$

with $\ell \in \{1, 2\}$ (an assumption we did not need to make for the forward Kullback-Leibler), we thus aim at finding

$$\begin{aligned} q_1(y_1) &\propto \left(\int_{\mathsf{Y}_2} \exp\left[-\frac{(1-\alpha)}{2}(y-\mu)^T \Lambda (y-\mu) - \frac{\alpha}{2}(y_2 - m_2)^2 \Lambda_2 \right] \nu_2(y_2) \right)^{\frac{1}{1-\alpha}} \\ &\propto \left(\int_{\mathsf{Y}_2} \exp\left[-\frac{(1-\alpha)}{2}(y-\mu)^T \Lambda (y-\mu) - \frac{\alpha}{2}(y_2^2 - 2y_2 m_2) \Lambda_2 \right] \nu_2(y_2) \right)^{\frac{1}{1-\alpha}} \end{aligned}$$

Since $(y-\mu)^T \Lambda (y-\mu) = (y_1 - \mu_1)^2 \Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2 \Lambda_{2,2}$, by grouping each term we have that

$$\begin{aligned} &(y-\mu)^T \Lambda (y-\mu) \\ &\quad = y_1^2 \Lambda_{1,1} - 2y_1 \left[\mu_1 \Lambda_{1,1} + \mu_2 \Lambda_{1,2} \right] + y_2^2 \Lambda_{2,2} - 2y_2 \left[\mu_2 \Lambda_{2,2} - (y_1 - \mu_1)\Lambda_{1,2} \right] + \text{cte} \end{aligned}$$

where cte does not depend on y_1 nor y_2. Consequently, we have

$$q_1(y_1) \propto \exp\left[-\frac{1}{2}\left(y_1^2 \Lambda_{1,1} - 2y_1 \left[\mu_1 \Lambda_{1,1} + \mu_2 \Lambda_{1,2} \right] \right) \right] \times \left(\int_{\mathsf{Y}_2} A(y_1, y_2) \nu_2(\mathrm{d}y_2) \right)^{\frac{1}{1-\alpha}} \tag{A.1}$$

where

$$\begin{aligned} A(y_1, y_2) &:= \exp\left[-\frac{(1-\alpha)}{2}\left(y_2^2 \Lambda_{2,2} - 2y_2 \left[\mu_2 \Lambda_{2,2} - (y_1 - \mu_1)\Lambda_{1,2} \right] \right) - \frac{\alpha}{2}(y_2^2 - 2y_2 m_2)\Lambda_2 \right] \\ &= \exp\left[-\frac{1}{2}\Lambda_{inter}\left(y_2^2 - 2y_2 \mu_{inter} \right) \right] \end{aligned}$$

with

$$\begin{aligned} \Lambda_{inter} &= (1-\alpha)\Lambda_{2,2} + \alpha\Lambda_2 \\ \Lambda_{inter}\mu_{inter} &= (1-\alpha)\left[\mu_2 \Lambda_{2,2} - (y_1 - \mu_1)\Lambda_{1,2} \right] + \alpha m_2 \Lambda_2 \, . \end{aligned} \tag{A.2}$$

Thus,

$$\left(\int_{\mathsf{Y}_2} A(y_1, y_2) \nu_2(\mathrm{d}y_2) \right)^{\frac{1}{1-\alpha}} \propto \exp\left[\frac{1}{2(1-\alpha)} \Lambda_{inter} \mu_{inter}^2 \right]$$

Using that

$$\begin{aligned} (\Lambda_{inter}\mu_{inter})^2 &= \left((1-\alpha)\left[\mu_2 \Lambda_{2,2} + \mu_1 \Lambda_{1,2} \right] + \alpha m_2 \Lambda_2 - y_1(1-\alpha)\Lambda_{1,2} \right)^2 \\ &= y_1^2 (1-\alpha)^2 \Lambda_{1,2}^2 - 2y_1(1-\alpha)\Lambda_{1,2}\left((1-\alpha)\left[\mu_2 \Lambda_{2,2} + \mu_1 \Lambda_{1,2} \right] + \alpha m_2 \Lambda_2 \right) \\ &\quad + \text{cte} \, , \end{aligned}$$

where cte does not depend on y_1 nor y_2, and plugging into (A.1), we deduce

$$q_1(y_1) \propto \exp\left[-\frac{1}{2}\left(y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} + \mu_2\Lambda_{1,2}\right]\right)\right] \times \exp\left[\frac{1}{2\Lambda_{inter}}\left(y_1^2(1-\alpha)\Lambda_{1,2}^2 - 2y_1\Lambda_{1,2}\left((1-\alpha)\left[\mu_2\Lambda_{2,2} + \mu_1\Lambda_{1,2}\right] + \alpha m_2\Lambda_2\right)\right)\right] .$$

As a consequence,

$$\Lambda_1 = \Lambda_{1,1} - \frac{1}{\Lambda_{inter}}(1-\alpha)\Lambda_{1,2}^2$$

and

$$\begin{aligned}
\Lambda_1 m_1 &= \mu_1\Lambda_{1,1} + \mu_2\Lambda_{1,2} - \frac{1}{\Lambda_{inter}}\Lambda_{1,2}\left((1-\alpha)\left[\mu_2\Lambda_{2,2} + \mu_1\Lambda_{1,2}\right] + \alpha m_2\Lambda_2\right) \\
&= \mu_1\left[\Lambda_{1,1} - \frac{1}{\Lambda_{inter}}(1-\alpha)\Lambda_{1,2}^2\right] + \mu_2\left[\Lambda_{1,2} - \frac{1}{\Lambda_{inter}}\Lambda_{1,2}(1-\alpha)\Lambda_{2,2}\right] \\
&\quad - \frac{1}{\Lambda_{inter}}\Lambda_{1,2}\alpha m_2\Lambda_2 \\
&= \Lambda_1\mu_1 + \mu_2\left[\Lambda_{1,2} - \frac{1}{\Lambda_{inter}}\Lambda_{1,2}(1-\alpha)\Lambda_{2,2}\right] - \frac{1}{\Lambda_{inter}}\Lambda_{1,2}\alpha m_2\Lambda_2 \\
&= \Lambda_1\mu_1 + \frac{\Lambda_{1,2}}{\Lambda_{inter}}\left[\mu_2\left(\Lambda_{inter} - (1-\alpha)\Lambda_{2,2}\right) - \alpha m_2\Lambda_2\right] .
\end{aligned}$$

By definition of Λ_{inter} in (A.2) we get

$$\begin{aligned}
\Lambda_1 &= \Lambda_{1,1} - \frac{1}{(1-\alpha)\Lambda_{2,2} + \alpha\Lambda_2}(1-\alpha)\Lambda_{1,2}^2 \\
m_1 &= \mu_1 - \frac{\Lambda_{1,2}\Lambda_2\alpha}{\Lambda_1\left[(1-\alpha)\Lambda_{2,2} + \alpha\Lambda_2\right]}\left[m_2 - \mu_2\right]
\end{aligned}$$

and by symmetry we also obtain

$$\begin{aligned}
\Lambda_2 &= \Lambda_{2,2} - \frac{1}{(1-\alpha)\Lambda_{1,1} + \alpha\Lambda_1}(1-\alpha)\Lambda_{1,2}^2 \\
m_2 &= \mu_2 - \frac{\Lambda_{1,2}\Lambda_1\alpha}{\Lambda_2\left[(1-\alpha)\Lambda_{1,1} + \alpha\Lambda_1\right]}\left[m_1 - \mu_1\right] .
\end{aligned}$$

- From there, we can deduce that the only possible stable fixed point for m_1 and m_2 are $m_1 = \mu_1$ and $m_2 = \mu_2$. As for Λ_1 and Λ_2, the fixed point conditions give for Λ_1:

$$\begin{aligned}
\Lambda_1 &= \Lambda_{1,1} - \frac{(1-\alpha)\Lambda_{1,2}^2}{(1-\alpha)\Lambda_{2,2} + \alpha\left(\Lambda_{2,2} - \frac{1}{(1-\alpha)\Lambda_{1,1}+\alpha\Lambda_1}(1-\alpha)\Lambda_{1,2}^2\right)} \\
&= \Lambda_{1,1} - \frac{(1-\alpha)\Lambda_{1,2}^2}{\Lambda_{2,2} - \frac{\alpha(1-\alpha)\Lambda_{1,2}^2}{(1-\alpha)\Lambda_{1,1}+\alpha\Lambda_1}} .
\end{aligned}$$

Our goal is thus to rewrite

$$\Lambda_1 = \Lambda_{1,1} - \frac{1}{a_\alpha - \frac{\alpha}{(1-\alpha)\Lambda_{1,1}+\alpha\Lambda_1}} \tag{A.3}$$

as a second-order equation, where we have set $a_\alpha = \frac{\Lambda_{2,2}}{\Lambda_{1,2}^2(1-\alpha)}$. Since,

$$\Lambda_1 \left(a_\alpha - \frac{\alpha}{(1-\alpha)\Lambda_{1,1} + \alpha\Lambda_1} \right) = \Lambda_{1,1} a_\alpha - \frac{\alpha\Lambda_{1,1}}{(1-\alpha)\Lambda_{1,1} + \alpha\Lambda_1} - 1$$

that is

$$\begin{aligned} \Lambda_1 \left(\left[(1-\alpha)\Lambda_{1,1} + \alpha\Lambda_1 \right] a_\alpha - \alpha \right) &= \left[(1-\alpha)\Lambda_{1,1} + \alpha\Lambda_1 \right] \left(\Lambda_{1,1} a_\alpha - 1 \right) - \alpha\Lambda_{1,1} \\ &= \Lambda_{1,1} \left((1-\alpha)\Lambda_{1,1} a_\alpha - 1 \right) + \alpha\Lambda_1 \left(\Lambda_{1,1} a_\alpha - 1 \right) \ , \end{aligned}$$

we deduce that (A.3) is equivalent to

$$\alpha a_\alpha \Lambda_1^2 + a_\alpha (1-2\alpha)\Lambda_{1,1}\Lambda_1 - \Lambda_{1,1} \left((1-\alpha)\Lambda_{1,1} a_\alpha - 1 \right) = 0$$

whose solutions are given by

$$\begin{aligned} \Lambda_1 &= \Lambda_{1,1} \times \frac{1}{2\alpha} \left((2\alpha - 1) \pm \sqrt{1 - \frac{4\alpha}{a_\alpha \Lambda_{1,1}}} \right) \\ &= \Lambda_{1,1} \times \frac{1}{2\alpha} \left((2\alpha - 1) \pm \sqrt{1 - \frac{4\alpha\Lambda_{1,2}^2(1-\alpha)}{\Lambda_{2,2}\Lambda_{1,1}}} \right) \ . \end{aligned}$$

- For the numerical application, recall that we have taken $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = \Lambda_{2,1} = -2$ in Example 2. In that case, since we need to ensure $\Lambda_1 > 0$, we must satisfy

$$1 - \frac{4\alpha\Lambda_{1,2}^2(1-\alpha)}{\Lambda_{2,2}\Lambda_{1,1}} > 0$$

and $\Lambda_1 > 0$. As $\frac{4\alpha\Lambda_{1,2}^2(1-\alpha)}{\Lambda_{2,2}\Lambda_{1,1}} \leqslant 4/9$ for $\alpha \in (0,1)$, the first condition is satisfied. Now defining

$$\varrho_\alpha := \frac{1}{2\alpha} \left((2\alpha - 1) + \sqrt{1 - \frac{4\alpha\Lambda_{1,2}^2(1-\alpha)}{\Lambda_{2,2}\Lambda_{1,1}}} \right) \ .$$

the only solution for $\alpha < 0.87$ is given by $\Lambda_1 = \varrho_\alpha \Lambda_{1,1}$ and (by symmetry) by $\Lambda_2 = \varrho_\alpha \Lambda_{2,2}$, as $1/2(1 + \sqrt{1 - 4/9}) > 0.87$. We have thus recovered the results announced in Hernandez-Lobato et al., 2016.

Observe that letting $\alpha \to 1$ in these expressions gives back the solutions $\Lambda_1 = \Lambda_{1,1}$ and $\Lambda_2 = \Lambda_{2,2}$ found in the forward Kullback-Leibler case. As for the case $\alpha \to 0$, de l'Hopital's rule gives $\Lambda_1 = \Lambda_{1,1} - \Lambda_{1,2}^2\Lambda_{2,2}$ and $\Lambda_2 = \Lambda_{2,2} - \Lambda_{1,2}^2\Lambda_{1,1}$.

A.2 Equivalence between optimising $D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ and $\Psi_\alpha(q;\mathscr{D})$

- Case $\alpha = 1$ with $f_1(u) = 1 - u + u\log(u)$ for all $u > 0$. Then,

$$\begin{aligned} D_1(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) &= \int_{\mathsf{Y}} f_1\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} q(y)\log\left(\frac{q(y)}{p(y|\mathscr{D})}\right)\nu(\mathrm{d}y) + 0 \\ &= \int_{\mathsf{Y}} q(y)\log\left(\frac{q(y)}{p(y,\mathscr{D})}\right)\nu(\mathrm{d}y) + \log p(\mathscr{D}) \\ &= \int_{\mathsf{Y}} f_1\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) + 1 - p(\mathscr{D}) + \log p(\mathscr{D}) \end{aligned}$$

Thus,

$$\mathrm{arginf}_{q\in\mathcal{Q}} D_1(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \mathrm{arginf}_{q\in\mathcal{Q}} \Psi_1(q;\mathscr{D})$$

- Case $\alpha = 0$ with $f_0(u) = u - 1 - \log(u)$ for all $u > 0$.

$$\begin{aligned} D_0(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) &= \int_{\mathsf{Y}} f_0\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) - \log p(\mathscr{D}) \\ &= \frac{1}{p(\mathscr{D})}\left[\int_{\mathsf{Y}} f_1\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) + p(\mathscr{D}) - 1 - p(\mathscr{D})\log p(\mathscr{D})\right] \end{aligned}$$

Thus

$$\mathrm{arginf}_{q\in\mathcal{Q}} D_0(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \mathrm{arginf}_{q\in\mathcal{Q}} \Psi_0(q;\mathscr{D})$$

- Case $\alpha \in \mathbb{R}\setminus\{1\}$ with $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$ for all $u > 0$.

$$\begin{aligned} &D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \\ &= \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y|\mathscr{D})}\right)^\alpha - 1\right] p(y|\mathscr{D})\nu(\mathrm{d}y) \\ &= p(\mathscr{D})^{\alpha-1}\int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y,\mathscr{D})}\right)^\alpha - 1\right] p(y,\mathscr{D})\nu(\mathrm{d}y) + \frac{p(\mathscr{D})^\alpha - 1}{\alpha(\alpha-1)} \\ &= p(\mathscr{D})^{\alpha-1}\int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) + \frac{\alpha p(\mathscr{D})^{\alpha-1} + (1-\alpha)p(\mathscr{D})^\alpha - 1}{\alpha(\alpha-1)} \end{aligned}$$

Thus,

$$\mathrm{arginf}_{q\in\mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \mathrm{arginf}_{q\in\mathcal{Q}} \Psi_\alpha(q;\mathscr{D})$$

B

Introduction (en Français)

L'objectif de ce chapitre est d'introduire les concepts fondamentaux apparaissant dans la thèse. Nous commençons par rappeler les bases de l'Inférence Bayésienne en soulignant les principales difficultés rencontrées dans ce domaine dès lors que l'on travaille avec des modèles bayésiens complexes. Nous expliquons par la suite comment les méthodes de Monte Carlo et les méthodes d'Inférence Variationnelle permettent de dépasser certaines de ces difficultés, avant de résumer les contributions que nous apportons dans ces deux domaines.

B.1 Inférence Bayésienne

L'Inférence Statistique regroupe les méthodes visant à modéliser un phénomène à partir d'un jeu de données. En tant que sous-catégorie de l'Inférence Statistique, les méthodes d'Inférence Bayésienne proposent de calibrer un modèle probabiliste paramétrique afin de décrire des données observées, avec la particularité que ces méthodes incorporent des connaissances a priori sur les paramètres du modèle considéré.

Plus précisément, le cadre de l'Inférence Bayésienne peut être explicité comme suit. Soit un espace mesuré $(\mathsf{Y}, \mathcal{Y}, \nu)$, où ν est une mesure σ-finie sur $(\mathsf{Y}, \mathcal{Y})$. Supposons que nous avons accès à un ensemble de données $\mathscr{D}$ généré à partir d'un modèle probabiliste dominé ayant pour densité $p(\mathscr{D}|y)$, où le paramètre $y \in \mathsf{Y}$ est une variable latente elle-même simulée selon une distribution *a priori* de densité p_0 par rapport à ν. La quantité phare de l'Inférence Bayésienne est la *densité a posteriori* de la variable latente y étant donné l'ensemble $\mathscr{D}$:

$$p(y|\mathscr{D}) = \frac{p(y, \mathscr{D})}{p(\mathscr{D})} = \frac{p_0(y)p(\mathscr{D}|y)}{p(\mathscr{D})} \ ,$$

où $p(\mathscr{D}) = \int_{\mathsf{Y}} p_0(y)p(\mathscr{D}|y)\nu(\mathrm{d}y)$ est la *loi marginale*. Cette dernière permet de quantifier l'incertitude du paramètre y suite à l'observation des données $\mathscr{D}$, au travers de quantités telles que la loi marginale $p(\mathscr{D})$ ou encore la *moyenne a posteriori*

$$\int_{\mathsf{Y}} y\, p(y|\mathscr{D})\nu(\mathrm{d}y) \ .$$

Plus généralement, le succès des méthodes d'Inférence Bayésienne repose sur notre capacité à calculer des intégrales de la forme suivante :

$$\int_{\mathsf{Y}} g(y)p(y|\mathscr{D})\nu(\mathrm{d}y) \ , \tag{B.1}$$

où g est une fonction d'intérêt définie sur Y.

Le problème susmentionné est difficile à résoudre : il n'existe pas d'expression analytique générale pour (B.1) et bien qu'une expression analytique soit connue pour certains choix de modèles probabilistes, celle-ci requiert souvent des temps de calculs trop longs en pratique (e.g. le calcul de la loi marginale pour un modèle de mélange Gaussien bayésien, voir Blei, Kucukelbir, and McAuliffe, 2017).

Cette difficulté à calculer (B.1) est particulièrement prégnante dans le contexte des données massives. En effet, la modélisation de grands volumes de données - avec une structure sous-jacente des données potentiellement compliquée - engendre des densités a posteriori très complexes. Dès lors, il devient crucial de trouver des algorithmes d'Inférence Bayésienne applicables au traitement des données massives et ce, avec des temps de calcul raisonnables.

Les méthodes d'Inférence Bayésienne exactes étant souvent impossibles à mettre en œuvre en pratique, une alternative consiste à faire appel à des méthodes d'Inférence Bayésienne *approchées*. Ces dernières appartiennent principalement à deux grandes catégories : (i) les méthodes de Monte Carlo (e.g. les méthodes par échantillonnage préférentiel adaptatif (Oh and Berger, 1992), les méthodes de Monte Carlo par chaînes de Markov (Neal, 1993), les méthodes de Monte Carlo séquentielles (Doucet, Freitas, and Gordon, 2001)), qui sont des méthodes d'*échantillonnage* (ii) les méthodes d'Inférence Variationnelle (e.g. l'algorithme Variational Bayes (Jordan et al., 1999) et l'algorithme Expectation Propagation (Minka, 2001)), qui reposent sur des techniques d'*optimisation*.

En l'état, les méthodes d'Inférence Variationnelle sont souvent plébiscitées du fait de leurs avantages numériques. Elles ont en effet été appliquées avec succès à des tâches d'apprentissage automatique en grande dimension faisant intervenir des modèles probabilistes complexes (Hoffman et al., 2013; Kingma and Welling, 2014; Ranganath, Gerrish, and Blei, 2014). Néanmoins, et contrairement aux méthodes de Monte Carlo, les méthodes d'Inférence Variationnelle utilisent des techniques d'optimisation sur un espace de densités restreint ; cela signifie qu'il y a un potentiel écart entre la densité a posteriori et l'approximation retournée à la fin de la

procédure d'optimisation et donc que les garanties théoriques pour ces méthodes font fréquemment défaut (Yao et al., 2018; Campbell and Li, 2019).

De ce fait, la recherche s'est tournée vers la construction d'algorithmes d'Inférence Variationnelle bénéficiant de solides garanties théoriques (e.g. Alquier, Ridgway, and Chopin, 2016; Domke, 2019; Alquier and Ridgway, 2020). Ce processus s'est également accompagné d'avancées permettant de combiner les méthodes de Monte Carlo et d'Inférence Variationnelle (pour ne citer que quelques exemples : Burda, Grosse, and Salakhutdinov, 2016; Li and Turner, 2016; Mandt, Hoffman, and Blei, 2017; Naesseth et al., 2018; Thin et al., 2020; Naesseth, Lindsten, and Blei, 2020).

Dans cette thèse, nous nous attachons à étudier comment les méthodes de Monte Carlo adaptatives, et plus spécifiquement les méthodes d'échantillonnage préférentiel adaptatif, peuvent être associées aux procédures d'Inférence Variationnelle afin de construire des algorithmes fondés théoriquement et applicables au traitement des données massives. Pour ce faire, nous commençons par rappeler les bases des méthodes de Monte Carlo en allant jusqu'aux méthodes d'échantillonnage préférentiel adaptatif, en gardant en ligne de mire les applications au cadre bayésien.

B.2 Méthodes de Monte Carlo et Inférence Bayésienne

Les méthodes de Monte Carlo dans leur ensemble visent à approximer des intégrales de la forme

$$I(g) := \int_{\mathsf{Y}} g(y)p(y)\nu(\mathrm{d}y) \ ,$$

où g est une fonction intégrable définie sur Y et p est une densité de probabilité par rapport à ν. Notons maintenant $\mathbb{P}$ la mesure de probabilité définie sur $(\mathsf{Y}, \mathcal{Y})$ et de dérivée de Radon-Nikodym par rapport à ν donnée par $\mathrm{d}\mathbb{P}/\mathrm{d}\nu = p$. Le problème ci-dessus peut alors être vu comme le calcul d'une espérance par rapport à la distribution de probabilité $\mathbb{P}$:

$$I(g) = \mathbb{E}_p[g(Y)] \ ,$$

où Y est une variable aléatoire définie sur l'espace probabilisé $(\mathsf{Y}, \mathcal{Y}, \mathbb{P})$.

La première idée des méthodes de Monte Carlo est de remplacer le calcul explicite de l'espérance $\mathbb{E}_p[g(Y)]$ par une approximation faisant intervenir la moyenne empirique de M réalisations indépendantes.

B.2.1 Monte Carlo standard

Soit $Y_1, Y_2, \ldots$ une suite de variables aléatoires indépendantes et identiquement distribuées partageant la même distribution de probabilité $\mathbb{P}$. Pour tout $M \in \mathbb{N}^\star$, l'estimateur $\hat{I}_M(g)$ de $I(g)$ donné par

$$\hat{I}_M(g) = \frac{1}{M} \sum_{m=1}^{M} g(Y_m) \tag{B.2}$$

est sans biais (i.e. $\mathbb{E}_p[\hat{I}_M(g)] = I(g)$) et la loi des grands nombres indique que sous l'hypothèse $I(g) = \mathbb{E}_p[|g(Y_1)|] < \infty$, nous pouvons écrire

$$\lim_{M\to\infty} \hat{I}_M(g) = I(g) \ , \quad \text{presque sûrement.}$$

En supposant de surcroît que $\mathbb{E}_p[|g(Y_1)|^2] < \infty$, le théorème central limite donne la convergence en distribution de $\sqrt{M}(\hat{I}_M(g) - I(g))$ vers $\mathcal{N}(0, \mathbb{V}\mathrm{ar}_p[g(Y_1)])$ lorsque M tend vers l'infini.

Ainsi, il s'agirait de choisir $p(y) = p(y|\mathscr{D})$ pour tout $y \in \mathsf{Y}$ pour pouvoir utiliser les méthodes de Monte Carlo dans un cadre bayésien. Il suffirait dès lors de parvenir à simuler sous la distribution a posteriori pour trouver en (B.2) une approximation non-biaisée de (B.1).

Il existe cependant d'importants modèles bayésiens pour lesquels nous ne savons pas simuler directement sous la distribution a posteriori et pour lesquels même la constante de renormalisation $p(\mathscr{D})$ est inconnue. A titre illustratif, nous fournissons ci-après un exemple dans le cas d'un modèle de Régression Logistique Bayésienne pour de la classification binaire.

Exemple 1 (Régression Logistique Bayésienne)**.** *Ce modèle est tiré de Gershman, Hoffman, and Blei, 2012. Nous observons les données $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$ qui sont constituées de I variables binaires, $c_i \in \{-1, 1\}$, et d'un vecteur de taille L pour chaque observation, $\boldsymbol{x}_i \in \mathbb{R}^L$. Les variables latentes $y = \{\boldsymbol{\omega}, \beta\}$ correspondent aux L coefficients de la régression $\omega_\ell \in \mathbb{R}$ ainsi qu'à un paramètre de précision $\beta \in \mathbb{R}^+$. Le modèle choisi est le suivant :*

$$\begin{aligned} p_0(\beta) &= \mathrm{Gamma}(\beta; a, b) \ , \\ p_0(\omega_\ell|\beta) &= \mathcal{N}(\omega_\ell; 0, \beta^{-1}) \ , \quad 1 \leqslant \ell \leqslant L \ , \\ p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{\omega}) &= \frac{1}{1 + e^{-\boldsymbol{\omega}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I \ , \end{aligned}$$

où a et b sont des hyperparamètres supposés fixés à l'avance. Pour tout $y \in \mathsf{Y}$, nous avons alors $p(y, \mathscr{D}) \propto p_0(y) \prod_{i=1}^I p(c_i|\boldsymbol{x}_i, y)$ avec $p_0(y) = \prod_{\ell=1}^L p_0(\omega_\ell|\beta) p_0(\beta)$. La quantité problématique dans ce modèle est la fonction sigmoid, qui empêche de simuler selon la distribution a posteriori. Elle rend également la constante de renormalisation $p(\mathscr{D})$ inconnue, tout comme la distribution prédictive a posteriori, en charge de prédire le label c_{new} lorsque nous sommes confrontés à une nouvelle observation x_{new}:

$$p(c_{\mathrm{new}}|\boldsymbol{x}_{\mathrm{new}}, \mathscr{D}) = \int_{\mathsf{Y}} p(c_{\mathrm{new}}|\boldsymbol{x}_{\mathrm{new}}, y) p(y|\mathscr{D}) \nu(\mathrm{d}y) \ .$$

Fort heureusement, les méthodes dites d'échantillonnage préférentiel permettent d'outrepasser ces complications.

B.2.2 Échantillonnage préférentiel

L'idée clef de l'échantillonnage préférentiel est d'introduire une certaine densité de probabilité q par rapport à ν, que nous appelons la *proposition*. Nous supposons deux choses sur q : (i) nous savons simuler selon q et (ii) le support de q contient le support de $g \times p$, c'est-à-dire que pour tout $y \in \mathsf{Y}$, $g(y)p(y) \neq 0$ implique que $q(y) > 0$ (une condition suffisante pour obtenir cette propriété étant que le support de q contient le support de p). Dans ce cas, en définissant $w(y) = p(y)/q(y)$ pour tout $y \in \mathsf{Y}$, nous avons que :

$$I(g) = \int_{\mathsf{Y}} g(y)p(y)\nu(\mathrm{d}y) = \int_{\mathsf{Y}} g(y)\frac{p(y)}{q(y)}q(y)\nu(\mathrm{d}y) = \mathbb{E}_q\left[w(Y)g(Y)\right] \ .$$

Soit $Y_1, Y_2, \ldots$ une suite de variables aléatoires indépendantes et identiquement distribuées simulées selon q, nous obtenons cette fois-ci qu'un nouvel estimateur (sans biais) de $I(g)$ est donné pour tout $M \in \mathbb{N}^\star$ par

$$\hat{I}_M^{IS}(g) = \frac{1}{M}\sum_{m=1}^{M} w(Y_m)g(Y_m) \ . \tag{B.3}$$

Si nous revenons maintenant au cadre de l'Inférence Bayésienne, une avancée importante est alors que l'estimateur $\hat{I}_M^{IS}(g)$ ne requiert plus de savoir simuler selon la densité a posteriori afin d'estimer des intégrales de la forme (B.1).

Il nous reste toutefois un obstacle à surmonter, lié au fait que la densité a posteriori est souvent calculable à une constante de normalisation près. Nous considérons pour cela l'estimateur normalisé (SNIS) donné pour tout $M \in \mathbb{N}^\star$ par

$$\hat{I}_M^{SNIS}(g) = \frac{\frac{1}{M}\sum_{m=1}^{M} w(Y_m)g(Y_m)}{\frac{1}{M}\sum_{m=1}^{M} w(Y_m)} \ .$$

Contrairement aux estimateurs introduits précédemment, $\hat{I}_M^{SNIS}(g)$ est biaisé. Néanmoins, lorsque les conditions de support sont satisfaites et que $\mathbb{E}_q[|w(Y_1)g(Y_1)|] < \infty$, la loi des grands nombres fournit la convergence presque sûre vers $I(g)$ pour les deux estimateurs $\hat{I}_M^{IS}(g)$ et $\hat{I}_M^{SNIS}(g)$.

Si nous ajoutons de plus l'hypothèse que $\mathbb{E}_q[|w(Y_1)g(Y_1)|^2] < \infty$ (et dans le cas de l'estimateur SNIS que $\mathbb{E}_q[w(Y_1)^2(1+g(Y_1)^2)] < \infty$), nous obtenons le résultat suivant

$$\begin{aligned}
&\sqrt{M}(\hat{I}_M^{IS}(g) - I(g)) \to_{\mathcal{L}} \mathcal{N}(0, \mathbb{V}\mathrm{ar}_q[w(Y_1)g(Y_1)]) \\
&\sqrt{M}(\hat{I}_M^{SNIS}(g) - I(g)) \to_{\mathcal{L}} \mathcal{N}(0, \mathbb{V}\mathrm{ar}_q[w(Y_1)(g(Y_1) - I(g))]) \ ,
\end{aligned}$$

où la notation $\to_{\mathcal{L}}$ désigne la convergence en distribution.

Ce résultat permet d'observer que $\mathbb{V}\mathrm{ar}_q[w(Y_1)g(Y_1)]$ et $\mathbb{V}\mathrm{ar}_q[w(Y_1)(g(Y_1) - I(g))]$ sont minimales lorsque $q \propto |g| \times p$ et $q \propto |g - I(g)|p$ respectivement, ce qui illustre le fait que les performances des méthodes par échantillonnage préférentiel dépendent fortement du choix de la proposition q (voir Robert and Casella, 2005 et la Figure

B.1). A noter toutefois qu'il devient moins efficace de choisir une proposition q qui dépend de g dès lors que nous cherchons à estimer des intégrales de la forme (B.1) pour un grand nombre de fonctions g ; il convient alors de viser directement la densité cible p (Delyon and Portier, 2021).

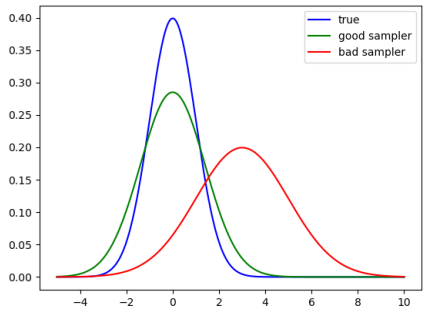


FIGURE B.1: Dans cette figure, "true" représente la distribution cible p. Les deux distributions restantes sont ainsi susceptibles de retourner des estimations très différentes de $I(g)$.

Comme il n'est pas forcément aisé de savoir tout de go comment bien choisir q lorsque nous sommes confrontés à un modèle probabiliste complexe, des procédures adaptatives au cours desquelles la densité q est progressivement améliorée peuvent être envisagées. Ceci nous amène au concept d'échantillonnage préférentiel adaptatif.

B.2.3 Échantillonnage préférentiel adaptatif

L'objectif des méthodes d'échantillonnage préférentiel adaptatif est de partir d'une densité de probabilité initiale q_1 et de construire itérativement une suite de densités de probabilité $(q_n)_{n\geqslant 1}$ permettant d'améliorer nos approximations de $I(g)$ au fur et à mesure que n augmente.

Bien qu'initialement limitées à des procédures en deux étapes (Kloek and Van Dijk, 1978; Geweke, 1989), les méthodes d'échantillonnage préférentiel adaptatif ont depuis évolué vers des procédures multi-étapes (Oh and Berger, 1992) de sorte qu'un algorithme d'échantillonnage préférentiel adaptatif peut typiquement s'écrire à la manière de l'Algorithme 16.

En choisissant $p = p(\cdot, \mathscr{D})$, les paires retournées par l'Algorithme 16 peuvent dès lors être utilisées dans l'estimation d'intégrales de la forme (B.1), en ayant par exemple recours à l'estimateur au temps n suivant :

$$\hat{I}^{SNAIS}_{M_n,n}(g) = \frac{\frac{1}{M_n}\sum_{m=1}^{M_n} w_n(Y_{m,n})g(Y_{m,n})}{\frac{1}{M_n}\sum_{m=1}^{M_n} w_n(Y_{m,n})} .$$

Un état de l'art détaillé des méthodes d'échantillonnage préférentiel adaptatif est disponible dans Bugallo et al., 2017. Parmi les avancées notables dans le domaine de

Algorithme 16: *Échantillonnage préférentiel adaptatif*

Entrée: N: nombre total d'itérations, $(M_n)_{1\leqslant n\leqslant N}$: politique de répartition des ressources, q_1: proposition initiale.
Sortie: Retourne les paires $(Y_{m,n}, w_n(Y_{m,n}))_{1\leqslant m\leqslant M_n, 1\leqslant n\leqslant N}$.
pour $n = 1 \dots N$ **faire**

1. Générer M_n points $(Y_{m,n})_{1\leqslant m\leqslant M_n}$ simulés indépendamment sous q_n.
2. Calculer les poids d'importance $(w_n(Y_{m,n}))_{1\leqslant m\leqslant M_n}$, où pour tout $y \in \mathsf{Y}$, nous définissons $w_n(y) = p(y)/q_n(y)$.
3. Mettre à jour la proposition q_n.

fin

l'échantillonnage préférentiel adaptatif se trouvent alors des méthodes construisant la suite $(q_n)_{n\geqslant 1}$ en minimisant un critère bien choisi à chaque étape (e.g. Douc et al., 2007a; Douc et al., 2007b; Cappé et al., 2008 et Portier and Delyon, 2018) ainsi que des méthodes proposant de nouveaux raffinements au niveau des poids d'importance (Martino et al., 2017).

Nous présentons maintenant les méthodes d'Inférence Variationnelle.

B.3 Méthodes d'Inférence Variationnelle et Inférence Bayésienne

Les méthodes d'Inférence Variationnelle (Jordan et al., 1999) cherchent à approcher la densité a posteriori par une densité variationnelle plus simple q appartenant à une famille de densité $\mathcal{Q}$ et facilitant le calcul d'intégrales de la forme (B.1).

Pour ce faire, ces approches proposent de résoudre un problème d'optimisation faisant intervenir une certaine mesure de dissimilarité D entre la distribution a posteriori $\mathbb{P}_{|\mathscr{D}}$ et la distribution variationnelle $\mathbb{Q}$:

$$\inf_{q\in\mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \ ,$$

où $\mathbb{P}_{|\mathscr{D}}$ sont $\mathbb{Q}$ des densités de probabilités sur $(\mathsf{Y}, \mathcal{Y})$ que l'on suppose absolument continues par rapport à ν (ce que nous indiquons aussi par la notation $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$) et de dérivées de Radon-Nikodym par rapport à ν : $q = \mathrm{d}\mathbb{Q}/\mathrm{d}\nu$ et $p(\cdot|\mathscr{D}) = \mathrm{d}\mathbb{P}_{|\mathscr{D}}/\mathrm{d}\nu$.

Les méthodes d'Inférence Variationnelle s'attachent alors à bien choisir D et à trouver des familles variationnelles $\mathcal{Q}$ afin de mener la procédure d'optimisation efficacement tout en étant capable de capturer une structure compliquée au sein de la densité a posteriori.

Dans cette section, nous rappelons en premier lieu le choix le plus traditionnel pour D et $\mathcal{Q}$ en Inférence Variationnelle. Nous détaillons par la suite les avancées en Inférence Variationnelle importantes à mentionner dans le cadre de cette thèse.

B.3.1 L'Inférence Variationnelle au sens traditionnel

Un choix traditionnel et extrêmement fréquent en Inférence Variationnelle consiste à utiliser la divergence de Kullback-Leibler comme mesure de dissimilarité D. Nous rappelons la définition de cette divergence maintenant.

Définition 1 (Divergence de Kullback-Leibler)**.** *Soient $\mathbb{Q}$ et $\mathbb{P}$ deux mesures de probabilités sur $(\mathsf{Y}, \mathcal{Y})$ absolument continues par rapport à ν i.e. $\mathbb{Q} \preceq \nu$, $\mathbb{P} \preceq \nu$. Nous notons $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$ and $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$ les dérivées de Radon-Nikodym de $\mathbb{Q}$ et $\mathbb{P}$ par rapport à ν. La divergence de Kullback-Leibler (KL) est alors donnée par:*

$$D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y)$$

et est à valeurs dans $[0, +\infty]$.

L'Inférence Variationnelle au sens traditionnel se concentre sur la minimisation de la divergence de Kullback-Leibler exclusive

$$\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{B.4}$$

ainsi que sur la minimisation de la divergence de Kullback-Leibler inclusive

$$\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{P}_{|\mathscr{D}}||\mathbb{Q}) \ . \tag{B.5}$$

Parmi ces deux problèmes d'optimisation, un intérêt plus poussé a été porté sur la résolution de (B.4) en raison de l'Evidence Lower BOund (ELBO) : pour tout $q \in \mathcal{Q}$, nous pouvons effectivement écrire que

$$\begin{aligned} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) &= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D}) \\ &= -\mathrm{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D}) \ , \end{aligned}$$

où la fonction ELBO est définie pour tout $q \in \mathcal{Q}$ par

$$\mathrm{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \ . \tag{B.6}$$

Le résultat ci-dessus signifie que l'ELBO agit comme une fonction objectif alternative ne faisant pas intervenir la constante de renormalisation $p(\mathscr{D})$, c'est-à-dire que le problème d'optimisation

$$\sup_{q\in\mathcal{Q}} \mathrm{ELBO}(q;\mathscr{D})$$

est strictement équivalent à (B.4).

Le nom ELBO tient alors son origine dans le fait que $\mathrm{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ avec égalité si et seulement si $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$. Ce résultat, établi en utilisant l'inégalité de Jensen appliquée à la fonction strictement concave $u \mapsto \log(u)$ nous dit effectivement que l'ELBO est une borne inférieure (lower bound) du log de la loi marginale $p(\mathscr{D})$ (the evidence).

Nous passons maintenant au choix de la famille variationnelle $\mathcal{Q}$: l'idée traditionnelle de l'Inférence Variationnelle est de fonctionner sous l'hypothèse de champs moyen, que nous allons expliciter ci-après. Ceci nous permettra d'expliquer brièvement comment l'ELBO et l'approche par champs moyen peuvent être combinées ensemble dans le cadre de l'Inférence Variationnelle.

L'approche par champs moyen fait l'hypothèse simplificatrice suivante : la variable latent y est constituée de L composantes indépendantes $(y_1, \ldots, y_L) \in \mathsf{Y}_1 \times \ldots \times \mathsf{Y}_L$ de sorte que $\mathcal{Q}$ se décompose de la manière suivante :

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\},$$

et que la dépendance en la variable latente y_ℓ n'apparaît qu'au travers de la densité variationnelle associée q_ℓ (nous écrivons $\nu(\mathrm{d}y) = \bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$).

Pour ce choix de famille variationnelle $\mathcal{Q}$, l'expression de l'ELBO (B.6) se simplifie. En maintenant tous les facteurs variationnels constants sauf celui correspondant à la coordonnée ℓ, nous pouvons alors déduire une formule de mise à jour optimale pour le facteur q_ℓ :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right) \ , \text{ pour } \nu_\ell\text{-presque tout } y_\ell \in \mathsf{Y}_\ell \ , \tag{B.7}$$

où nous utilisons la notation $\mathbb{E}_{-\ell}$ pour désigner l'espérance par rapport à q en omettant le facteur q_ℓ. En effet, sous l'hypothèse de champs moyen

$$\begin{aligned}
\mathrm{ELBO}(q;\mathscr{D}) &= \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y,\mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log q_\ell(y_\ell) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell} \\
&= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y,\mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}
\end{aligned}$$

où $c_{-\ell}$ est une constante qui ne dépend pas de q_ℓ (et par commodité nous faisons un léger abus de notation dans l'écriture de $\mathbb{E}_{-\ell}\left[\log p(y,\mathscr{D})\right]$ en utilisant la même notation pour les variables $(y_k)_{1\leqslant k\leqslant L, k\neq\ell}$ et les variables aléatoires sous $\prod_{k=1,k\neq\ell}^{L} q_k$).

D'après l'inégalité de Jensen, le terme de droite est maximisé lorsque $q_\ell(y_\ell)$ est proportionnel à $\exp\left(\mathbb{E}_{-\ell}[\log p(y,\mathscr{D})]\right)$ pour ν_ℓ-presque tout $y_\ell \in \mathsf{Y}_\ell$; nous retrouvons ainsi la condition d'optimalité annoncée dans (B.7) pour le facteur q_ℓ.

L'idée naturelle est maintenant de s'appuyer sur (B.7) pour effectuer la mise à jour de chaque facteur q_ℓ de manière cyclique, jusqu'à atteindre la convergence vers un optimum (local) : cette procédure porte le nom de l'algorithme Coordinate Ascent Variational Inference (Bishop, 2006) et elle est décrite dans l'Algorithme 17.

Algorithme 17: *Coordinate Ascent Variational Inference (CAVI)*

Entrée: $(q_\ell)_{1\leqslant\ell\leqslant L}$: facteurs variationnels initiaux.
Sortie: Retourne la densité variationnelle optimisée q définie pour tout $y \in \mathsf{Y}$ par $q(y) = \prod_{\ell=1}^{L} q_\ell(y_\ell)$.
tant que *l'ELBO n'a pas convergé* **faire**
  pour $\ell = 1 \dots L$ **faire**
    $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$, pour ν_ℓ-presque tout $y_\ell \in \mathsf{Y}_\ell$
  fin
  Calculer l'ELBO.
fin

L'exemple qui suit illustre la manière dont les formules de mise à jour apparaissant dans l'Algorithme 17 sont obtenues en pratique. Cet exemple jouet, tiré de Hernandez-Lobato et al., 2016 et pour lequel la densité a posteriori est connue, nous permettra également de visualiser les caractéristiques de l'approche par champ moyen dans le cadre de l'Inférence Variationnelle.

Exemple 2 (Régression Linéaire Bayésienne)**.** *Nous observons les données $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$ composées de I variables $(c_i)_{1\leqslant i\leqslant I}$ en dimension 1 et de I vecteurs $(\boldsymbol{x}_i)_{1\leqslant i\leqslant I}$ en dimension 2, où chaque paire $(c_i, \boldsymbol{x}_i)$ appartient à $\mathbb{R}\times\mathbb{R}^2$. Les variables latentes correspondent aux deux coefficients de la régression $y = \{y_1, y_2\} \in \mathbb{R}^2$. Le modèle choisi est le suivant :*

$$
\begin{aligned}
&p_0(y) = \mathcal{N}(y; \mu_0, \Lambda_0^{-1}) \ , \\
&p(c_i|\boldsymbol{x}_i, y) = \mathcal{N}(c_i; y^T\boldsymbol{x}_i, \sigma^2) \ , \quad 1 \leqslant i \leqslant I \ ,
\end{aligned}
$$

où μ_0, Λ_0 et σ sont des hyperparamètres supposés fixés.

Dans ce cas, la densité a posteriori est connue et elle vérifie

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

avec $\Lambda = \Lambda_0 + \sigma^{-2}\sum_{i=1}^{I}\boldsymbol{x}_i\boldsymbol{x}_i^T$ et $\Lambda\mu = \Lambda_0\mu_0 + \sigma^{-2}\sum_{i=1}^{I} c_i\boldsymbol{x}_i$. Sous l'hypothèse de champs moyen, nous cherchons q sous la forme $q(y) = q_1(y_1)q_2(y_2)$ avec pour tout $\ell = \{1, 2\}$ et tout $y_\ell \in \mathsf{Y}_\ell$

$$q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right) \ .$$

En notant $\mu = (\mu_\ell)_{1\leqslant\ell\leqslant 2}$ *et* $\Lambda = (\Lambda_{\ell,k})_{1\leqslant\ell,k\leqslant 2}$ *avec* $\Lambda_{1,2} = \Lambda_{2,1}$*, il vient :*

$$\begin{aligned} q_1(y_1) &\propto \exp\left(\mathbb{E}_{q_2}\left[-\frac{1}{2}\left\{(y_1-\mu_1)^2\Lambda_{1,1} + 2(y_1-\mu_1)(y_2-\mu_2)\Lambda_{1,2}\right\}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{q_2}[y_2]-\mu_2)\Lambda_{1,2}\right]\right\}\right) , \end{aligned}$$

et nous en déduisons que $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{q_2}[y_2]-\mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$ *et par symétrie que* $q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{q_1}[y_1]-\mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$*. Par conséquent, en notant* $m_1 = \mathbb{E}_{q_1}[y_1]$ *and* $m_2 = \mathbb{E}_{q_2}[y_2]$*, l'algorithme CAVI revient à effectuer les mises à jours suivantes*

$$\begin{aligned} m_1 &\leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2-\mu_2)\Lambda_{1,2} \\ m_2 &\leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1-\mu_1)\Lambda_{1,2} . \end{aligned}$$

Le seul point fixe de ce schéma itératif étant donné par $m_1 = \mu_1$ *et* $m_2 = \mu_2$*, nous obtenons finalement que* $q_1(y_1) = \mathcal{N}(y_1; \mu_1, \Lambda_{1,1}^{-1})$ *et* $q_2(y_2) = \mathcal{N}(y_2; \mu_2, \Lambda_{2,2}^{-1})$*. Nous pouvons alors visualiser la distribution a posteriori et la distribution variationnelle optmisée par l'algorithme CAVI sur la Figure B.2, dans laquelle* $\mu = [0, 0]$*,* $\Lambda_{1,1} = \Lambda_{2,2} = 3$ *et* $\Lambda_{1,2} = -2$*.*

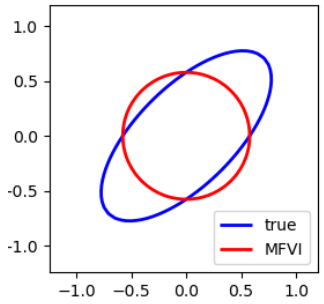


FIGURE B.2: Approche par champs moyen pour la Régression Linéaire Bayésienne de l'Exemple 2 (adaptée de Hernandez-Lobato et al., 2016). Ici, "true" et "MVFI" représentent respectivement la distribution a posteriori et la distribution sous l'hypothèse de champs moyen obtenue par minimisation de la Kullback-Leibler exclusive (avec des contours de taille 1-sigma).

Plus généralement, l'algorithme CAVI permet d'obtenir des formules de mise à jour explicites lorsque l'on travaille avec des modèles faisant intervenir des familles exponentielles conjuguées bien choisies. Parmi ces modèles, nous trouvons notamment les modèles de mélange gaussiens (Bishop, 2006) ainsi que les modèles d'Allocation de Dirichlet latente (Blei, Ng, and Jordan, 2003).

En effet, partant d'un jeu de données $\mathscr{D} = (x_\ell)_{1\leqslant\ell\leqslant L}$, ces derniers introduisent les variables latentes $y = \{\beta, \omega_1, \ldots, \omega_L\}$, où β est une variable latente globale et où pour tout $\ell = 1 \ldots L$ la variable latente locale ω_ℓ est associée à x_ℓ, de sorte que

$$p(y, \mathscr{D}) = p(\beta)\prod_{\ell=1}^{L} p(\omega_\ell, x_\ell|\beta)$$

Ils choisissent ensuite une densité variationnelle vérifiant l'hypothèse de champs moyen suivante :

$$q(y) = q(\beta|\psi) \prod_{\ell=1}^{L} q(\omega_\ell|\phi_\ell) \ ,$$

où $\{\psi, \phi_1, ..., \phi_L\}$ sont les paramètres variationnels à optimiser via l'algorithme CAVI. Ils déduisent finalement des formules de mise à jour pour ces paramètres variationnels grâce à des choix appropriés de $p(\beta), p(\omega_\ell, x_\ell|\beta), q(\beta|\psi)$ et $q(\omega_\ell|\phi_\ell)$ (voir Blei, Kucukelbir, and McAuliffe, 2017 pour les détails concernant les modèles de mélange gaussiens et les modèles d'Allocation de Dirichlet latente).

Nous avons vu comment l'algorithme CAVI permet d'obtenir des formules de mise à jour pour les paramètres variationnels lorsque D correspond à la Kullback-Leibler exclusive, $\mathcal{Q}$ vérifie l'hypothèse de champs moyen et $p(\cdot, \mathscr{D})$ est un modèle bien choisi appartenant aux modèles exponentiels conjugués. Dans le cadre des données massives, il reste une dernière difficulté à surmonter pour obtenir un algorithme utilisable en pratique : l'algorithme CAVI doit d'abord optimiser l'ensemble des paramètres variationnels locaux $\{\phi_1, ..., \phi_L\}$ avant de ré-estimer le paramètre variationnel global ψ, ce qui le rend inefficace lorsqu'il fait face à un grand volume de données.

Pour remédier à cette situation, la littérature en Inférence Variationnelle a mis à contribution les techniques d'optimisation stochastique (Bottou, 2010; Robbins and Monro, 1951). L'algorithme Stochastic Variational Inference (Hoffman et al., 2013) a ainsi permis l'apprentissage en grande dimension pour des modèles complexes comme celui d'Allocation de Dirichlet latente et le succès numérique de cette approche sur des jeux de données contenant des millions d'observations a ravivé l'intérêt pour les méthodes d'Inférence Variationnelle (voir Blei, Kucukelbir, and McAuliffe, 2017 et Zhang et al., 2019 pour des états de l'art sur ce domaine).

Le reste de cette section est dédié aux avancées majeures en Inférence Variationnelle que nous aurons l'occasion de remettre en perspective au cours de cette thèse.

B.3.2 A la rencontre des méthodes de Monte Carlo

Comme nous l'avons souligné précédemment, l'usage de la Kullback-Leibler exclusive sous des hypothèses de champs moyen facilite considérablement l'application des méthodes d'Inférence Variationnelle aux données massives.

Cependant, l'une des principales limitations de cette approche provient du fait que non seulement l'hypothèse de champs moyen restreint le choix des modèles, mais également que les formules de mises à jour obtenues sont spécifiques au modèle considéré et requièrent de ce fait d'être établies à la main (voir Blei, Kucukelbir, and McAuliffe, 2017 et la Figure B.2).

Pour répondre à cette difficulté, les techniques d'Inférence Variationnelle dites "Black-Box" ont été déployées (Ranganath, Gerrish, and Blei, 2014) : cette nouvelle classe d'algorithmes permet de minimiser la Kullback-Leibler exclusive pour un choix de modèles beaucoup plus général. Prenons pour D la Kullback-Leibler exclusive et supposons que nous travaillons avec une famille variationnelle paramétrique de la forme :

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \tag{B.8}$$

(où T est par exemple $\mathbb{R}^d$). L'idée centrale de l'algorithme Black-Box Variational Inference est d'utiliser le gradient de l'ELBO ainsi que des approximations de Monte Carlo au cours de la procédure d'optimisation. En se plaçant sous des hypothèses de dérivabilité classiques et d'après Paisley, Blei, and Jordan, 2012, le gradient de l'ELBO (B.6) est en effet donné par :

$$\begin{aligned}
\nabla \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D}) &= \nabla \left(\int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) \nu(\mathrm{d}y) \right) \\
&= \int_{\mathsf{Y}} \nabla k(\theta, y) \times \left[\log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) - 1 \right] \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} k(\theta, y) \nabla \left[\log k(\theta, y) \right] \log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) \nu(\mathrm{d}y) - \int_{\mathsf{Y}} \nabla k(\theta, y) \nu(\mathrm{d}y)
\end{aligned}$$

où nous avons utilisé que pour tout $y \in \mathsf{Y}$, $\nabla k(\theta, y) = k(\theta, y) \nabla \left[\log k(\theta, y) \right]$, une astuce appelée le REINFORCE trick dans la littérature (Williams, 1992). Comme de plus $\int_{\mathsf{Y}} \nabla k(\theta, y) \nu(\mathrm{d}y) = \nabla \left(\int_{\mathsf{Y}} k(\theta, y) \nu(\mathrm{d}y) \right) = 0$, il vient que

$$\begin{aligned}
\nabla \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D}) &= \int_{\mathsf{Y}} k(\theta, y) \nabla \left[\log k(\theta, y) \right] \log \left(\frac{p(y,\mathscr{D})}{k(\theta,y)} \right) \nu(\mathrm{d}y) \\
&= \mathbb{E}_{k(\theta,\cdot)} \left[\nabla \log \left[k(\theta, Y) \right] \times \log \left(\frac{p(Y,\mathscr{D})}{k(\theta,Y)} \right) \right]
\end{aligned}$$

c'est-à-dire que le gradient de l'ELBO s'écrit comme une espérance par rapport à la densité variationnelle $k(\theta, \cdot)$. Les méthodes de Monte Carlo entrent alors en jeu : étant données M variables aléatoires $Y_1, \ldots, Y_M$ indépendantes et identiquement distribuées selon $k(\theta, \cdot)$, un estimateur non-biaisé de l'espérance ci-dessus est

$$\frac{1}{M} \sum_{m=1}^{M} \nabla \log \left[k(\theta, Y_m) \right] \log \left(\frac{p(Y_m,\mathscr{D})}{k(\theta,Y)} \right) \ .$$

L'algorithme Black-Box Variational Inference se base sur ce résultat pour introduire une suite de vitesses d'apprentissage $(\gamma_n)_{n \geqslant 1}$ et effectuer des pas de Descente de Gradient Stochastique afin de construire la suite $(\theta_n)_{n \geqslant 1}$ suivant l'Algorithme 18 (le signe "+" dans l'étape de descente de gradient provenant du fait que l'algorithme maximise l'ELBO et minimise donc $-$ELBO).

Algorithm 18: *Black-Box Variational Inference*

Entrée: N: nombre total d'itérations, $(M_n)_{n\geqslant 1}$: politique d'allocation des ressources, $(\gamma_n)_{n\geqslant 1}$: politique des vitesses d'apprentissage, θ_1: valeur initiale.

Sortie: Retourne le paramètre optimisé θ_{N+1}.

pour $n = 1 \dots N$ **faire**

1. Générer M_n points $(Y_{m,n})_{1\leqslant m\leqslant M_n}$ simulés indépendamment sous $k(\theta_n, \cdot)$.
2. Définir

$$\theta_{n+1} = \theta_n + \gamma_n \frac{1}{M_n} \sum_{m=1}^{M_n} \nabla \log\left[k(\theta, Y_{m,n})\right]|_{\theta=\theta_n} \log\left(\frac{p(Y_{m,n}, \mathscr{D})}{k(\theta_n, Y_{m,n})}\right) .$$

fin

La particularité de ce schéma itératif se trouve dans le caractère non-biaisé des estimateurs du gradient de l'ELBO apparaissant dans les pas de Descente de Gradient Stochastique. Sous certaines hypothèses sur la politique des vitesses d'apprentissage ainsi que sur la fonction objectif, la suite $(k(\theta_n, \cdot))_{n\geqslant 1}$ converge en effet vers un optimum de l'ELBO : l'algorithme Black-Box Variational Inference permet donc de minimiser (au moins localement) la divergence de Kullback-Leibler exclusive.

Il est important de noter qu'une grande variance des estimateurs du gradient de l'ELBO pourrait constituer un revers potentiel des méthodes d'Inférence Variationnelle de type Black-Box. Le succès de ces méthodes a ainsi en grande partie été attribué à la mise en œuvre de diverses techniques de réduction de variance (e.g. méthode de Rao-Blackwell, variables de contrôle (Ranganath, Gerrish, and Blei, 2014), reparamétrisation (Kingma and Welling, 2014) et méthodes de Quasi-Monte Carlo (Buchholz, Wenzel, and Mandt, 2018)).

Nous nous sommes intéressés jusqu'à présent au cas de la Kullback-Leibler (exclusive) dans notre choix de D. Néanmoins, un second attrait fondamental des méthodes d'Inférence Variationnelle de type Black-Box est qu'elles autorisent une gamme plus large de fonctions objectif par-delà la divergence de Kullback-Leibler exclusive. En particulier, ces procédures permettent de travailler avec la famille des α-divergences.

B.3.3 Méthodes d'Inférence Variationnelle basées sur la α-divergence

Les distributions variationnelles obtenues via la minimisation d'une divergence de Kullback-Leibler présentent des caractéristiques parfois indésirables d'un point de vue pratique (Minka, 2001; Hoffman et al., 2013; Blei, Kucukelbir, and McAuliffe, 2017), e.g. elles ont tendance à sous-estimer/sur-estimer la variance de la distribution a posteriori pour la Kullback-Leibler exclusive/inclusive (un effet qui peut

également être amplifié lorsque ce choix de divergence s'accompagne d'une hypothèse de champs moyen, voir la Figure B.2).

Par conséquent, une autre ramification de la recherche en Inférence Variationnelle s'est penchée sur la question du choix de la mesure de dissimilarité D. Notamment, Minka, 2004 et Minka, 2005 font partie des premiers travaux à faire appel à la α-divergence (Zhu and Rohwer, 1995a; Zhu and Rohwer, 1995b) dans un contexte d'Inférence Variationnelle. Nous mentionnons ici les principales propriétés de cette famille, avant d'expliquer pourquoi elle constitue un outil privilégié en Inférence Variationnelle.

La α-divergence est une famille de divergences qui tire son origine dans la littérature de la théorie de l'information (e.g. Cichocki and Amari, 2010) et qui généralise la divergence de Kullback-Leibler. Nous rappelons ci-dessous sa définition entre deux mesures de probabilité $\mathbb{Q}$ et $\mathbb{P}$.

Définition 2. *Soit $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Soient deux mesures de probabilité $\mathbb{Q}$ et $\mathbb{P}$ définies sur $(\mathsf{Y}, \mathcal{Y})$ et absolument continues par rapport à ν i.e. $\mathbb{Q} \preceq \nu$, $\mathbb{P} \preceq \nu$. Nous notons $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$ et $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$ les dérivées de Radon-Nikodym de $\mathbb{Q}$ et $\mathbb{P}$ par rapport à ν. La α-divergence entre $\mathbb{Q}$ et $\mathbb{P}$ est donnée par :*

$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)} \right)^\alpha - 1 \right] p(y) \nu(\mathrm{d}y) \ ,$$

et est à valeurs dans $[0, +\infty]$.

Sous des hypothèses de dérivabilité classiques, la α-divergence se prolonge par continuité en $\alpha = 0$ et en $\alpha = 1$ de telle sorte que nous retrouvons la Kullback-Leibler inclusive et exclusive : pour tout $y \in \mathsf{Y}$, nous pouvons en effet écrire :

$$\begin{aligned} \lim_{\alpha \to 0} \frac{1}{\alpha} \left[\frac{1}{\alpha - 1} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha - 1} \right] &= \nabla \left[\frac{1}{\alpha - 1} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha - 1} \right] \Bigg|_{\alpha=0} \\ &= -\log \left(\frac{q(y)}{p(y)} \right) \end{aligned}$$

ainsi que :

$$\begin{aligned} \lim_{\alpha \to 1} \frac{1}{\alpha - 1} \left[\frac{1}{\alpha} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha} \right] &= \nabla \left[\frac{1}{\alpha} \left(\frac{q(y)}{p(y)} \right)^\alpha - \frac{1}{\alpha} \right] \Bigg|_{\alpha=1} \\ &= 1 - \frac{q(y)}{p(y)} + \log \left(\frac{q(y)}{p(y)} \right) \frac{q(y)}{p(y)} \end{aligned}$$

d'où nous déduisons d'une part que $\lim_{\alpha \to 0} D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{P}||\mathbb{Q})$ et d'autre part que $\lim_{\alpha \to 1} D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{Q}||\mathbb{P})$ (nous utilisons par ailleurs les notations $D_0(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{P}||\mathbb{Q})$ et $D_1(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{Q}||\mathbb{P})$ dans l'ensemble du manuscrit).

Il convient par ailleurs de noter que la famille des α-divergences inclut également la distance de Hellinger et la divergence du χ^2 correspondant chacune à l'ordre $\alpha = 0.5$ et $\alpha = 2$ respectivement.

Soit maintenant f_α la fonction *convexe* définie sur $(0, +\infty)$ par $f_0(u) = u - 1 - \log(u)$, $f_1(u) = 1 - u + u\log(u)$ et $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$ pour tout $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Dans ce cas, pour tout $\alpha \in \mathbb{R}$,

$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \; . \tag{B.9}$$

Écrit sous cette forme, le terme de droite dans (B.9) peut se voir comme la définition générale d'une α-divergence (Cichocki and Amari, 2010).

Cette formulation inscrit la α-divergence dans la famille des f-divergences (Morimoto, 1963a; Morimoto, 1963b) au travers de la convexité de f_α et la proposition qui suit rappelle les propriétés essentielles de la α-divergence (voir Minka, 2005; Cichocki and Amari, 2010; Cichocki, Cruces, and Amari, 2011; Erven and Harremoes, 2014 et Sason, 2018 pour plus de détails concernant la famille des α-divergences).

Proposition 3. *La α-divergence (étendue par continuité en $\alpha = 0$ et en $\alpha = 1$) est positive et égale à zéro si et seulement si $\mathbb{Q} = \mathbb{P}$. De plus, elle est convexe en $(\mathbb{Q}, \mathbb{P})$ et la définition de la α-divergence est invariante par rapport à la transformation $\tilde{f}_{\alpha,c}(u) = f_\alpha(u) + c(u-1)$ pour toute constante arbitraire c, c'est-à-dire que f_α peut-être remplacée de manière équivalente par $\tilde{f}_{\alpha,c}$ dans* (B.9).

Ainsi, un problème d'optimisation plus général que celui visant à minimiser la divergence de Kullback-Leibler exclusive (B.4) ou inclusive (B.5) revient à considérer

$$\inf_{q \in \mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \; . \tag{B.10}$$

et les caractéristiques de la densité variationnelle solution du problème d'optimisation (B.10) vont varier selon la valeur de α (Minka, 2005).

Plus spécifiquement, il y deux régimes principaux : soit $\alpha \leqslant 0$ et la α-divergence est *inclusive*, c'est-à-dire qu'elle favorise les densités variationnelles q couvrant tous les modes, soit $\alpha \geqslant 1$ et la α-divergence est *exclusive*, ce qui signifie que q va être attirée par le mode de plus grande masse (le cas $\alpha \in (0, 1)$ correspondant à un entre-deux). Ces deux régimes s'expliquent par le fait que $D_\alpha(\mathbb{Q}||\mathbb{P})$ explose dès que le support de q est plus grand que celui de p pour $\alpha \geqslant 1$ et inversement, $D_\alpha(\mathbb{Q}||\mathbb{P})$ explose dès que le support de p est plus grand que celui de q pour $\alpha \leqslant 0$.

Cette propriété inclusive/exclusive de la α-divergence est illustrée dans la Figure B.3 ci-après, dans laquelle la cible est une distrubution multimodale que nous cherchons à approcher par la Gaussienne q optimale en terme de la α-divergence $D_\alpha(\mathbb{Q}||\mathbb{P})$, et ce, pour différentes valeurs de α.

Si nous reprenons maintenant le modèle considéré dans l'Exemple 2, l'effet du paramètre α sur la densité variationnelle optimale vérifiant l'hypothèse de champs

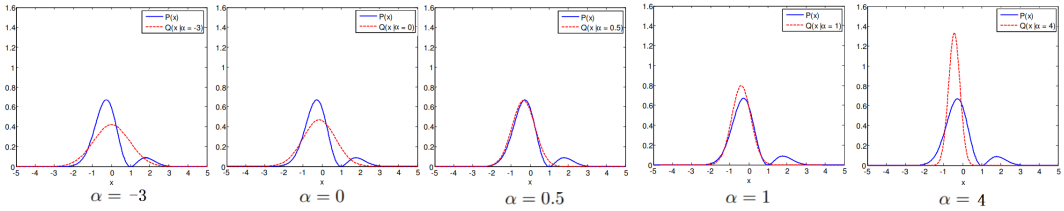


FIGURE B.3: En bleu la cible p multimodale et en rouge la Gaussienne q minimisant la α-divergence $D_\alpha(\mathbb{Q}||\mathbb{P})$ pour différentes valeurs de α. (adapté des notes de cours de Cevher disponibles ici: https://www.ece.rice.edu/~vc3/elec633/AlphaDivergence.pdf)

moyen peut être observé sur la Figure B.4 (voir Appendix A.1 pour les détails des calculs), ce qui met une fois de plus en évidence la propriété inclusive/exclusive de la α-divergence.

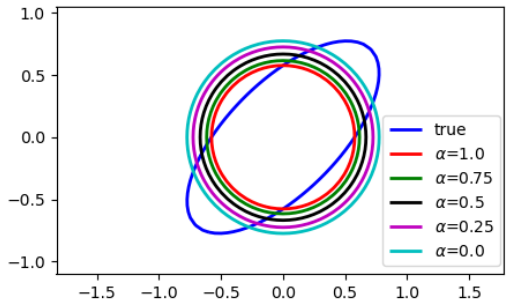


FIGURE B.4: Approximation variationnelle optimale sous l'hypothèse de champs moyen selon la valeur de α pour le modèle de Régression Linéaire Bayésienne de l'Exemple 2 (adapté de Hernandez-Lobato et al., 2016). Ici, "true" est la distribution a posteriori cible et les autres courbes désignent les distributions variationnelles obtenues par α-divergence minimisation (avec des contours de taille 1-sigma).

La propriété inclusive/exclusive de la α-divergence permet d'interpoler entre le comportement de la Kullback-Leibler inclusive et celui de la Kullback-Leibler exclusive, ce qui rend le problème d'optimisation (B.10) attractif pour réguler la variance de la densité variationnelle. Ceci explique l'intérêt qui a été porté à cette famille de divergences dans Minka, 2004 et Minka, 2005. Toutefois, ces travaux restent limités à des distributions au sein de la famille exponentielle.

L'émergence d'algorithmes d'Inférence Variationnelle mettant à profit les méthodes de Monte Carlo a permis de construire de nouveaux algorithmes basés sur la α-divergence aux performances empiriques prometteuses (Hernandez-Lobato et al., 2016; Li and Turner, 2016; Dieng et al., 2017; Kuleshov and Ermon, 2017).

Ces méthodes exploitent le fait que la forme spécifique de f_α permet de travailler sans la constante de renormalisation gênante $p(\mathscr{D})$ et elles se divisent en deux groupes distincts : d'un côté les méthodes *biaisées* (Hernandez-Lobato et al., 2016; Li

and Turner, 2016) et de l'autre les méthodes *non-biaisées* (Dieng et al., 2017; Kuleshov and Ermon, 2017).

Les méthodes biaisées considèrent une version légèrement modifiée de (B.10) qui repose sur la très liée α-divergence de Renyi (Rényi, 1961; Erven and Harremoes, 2014)

$$\begin{aligned} D_\alpha^{(\mathrm{AR})}(\mathbb{Q}||\mathbb{P}) &= \frac{1}{\alpha-1}\log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{\alpha-1}\nu(\mathrm{d}y)\right) \\ &= \frac{1}{\alpha-1}\log\left(1+\alpha(\alpha-1)D_\alpha(\mathbb{Q}||\mathbb{P})\right) \ . \end{aligned}$$

En particulier, Li and Turner, 2016 formalisent le concept de Variational Renyi (VR) bound, une nouvelle fonction objectif généralisant l'ELBO et étant définie pour tout $\alpha \in \mathbb{R}\setminus\{1\}$ et toute densité variationnelle $q \in \mathcal{Q}$ par

$$\mathcal{L}_\alpha(q;\mathscr{D}) := \frac{1}{1-\alpha}\log\left(\int_{\mathsf{Y}}\left(\frac{p(y,\mathscr{D})}{q(y)}\right)^{1-\alpha} q(y)\nu(\mathrm{d}y)\right)$$

de telle sorte qu'ils cherchent à résoudre

$$\sup_{q\in\mathcal{Q}} \mathcal{L}_\alpha(q;\mathscr{D}) \ .$$

Ils montrent que, selon le signe de α, la VR bound agit comme une borne inférieure ou supérieure du log de la loi marginale $\log p(\mathscr{D})$ et retrouvent l'ELBO lorsque $\alpha \to 1$ (Li and Turner, 2016, Theorem 1).

La procédure d'optimisation est alors menée en suivant l'idée de l'algorithme Black-Box Variational Inference, c'est-à-dire par Descente de Gradient Stochastique sur $-\mathcal{L}_\alpha(q;\mathscr{D})$, où q appartient à une famille variationnelle paramétrique de la forme (B.8). Du fait du $\log$, cette procédure fait intervenir un estimateur biaisé du gradient de la VB bound, là où les méthodes non-biaisées considèrent la fonction objectif donnée pour tout $q \in \mathcal{Q}$ par

$$\Psi_\alpha(q;\mathscr{D}) := \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y)$$

et visent à résoudre par Descente de Gradient Stochastique le problème d'optimisation équivalent à (B.10) (voir Appendix A.2) suivant :

$$\inf_{q\in\mathcal{Q}} \Psi_\alpha(q;\mathscr{D}) \ . \tag{B.11}$$

Les avancées en Inférence Variationnelle par minimisation de la α-divergence incluent notamment un calibrage automatique de l'hyperparamètre α (Wang, Liu, and Liu, 2018) ainsi que plusieurs tentatives pour comprendre quelle approche - biaisée ou non-biaisée - choisir d'un point de vue à la fois théorique et pratique (Geffner and Domke, 2020a; Geffner and Domke, 2020b).

Nous avons rappelé les bases des méthodes d'échantillonnage préférentiel adaptatif et abordé un ensemble de méthodes d'Inférence Variationnelle dépassant le cadre traditionnel de la minimisation de la divergence de Kullback-Leibler exclusive sous des hypothèses de champs moyens (par des techniques dite Black-Box et en offrant un choix plus large de fonctions objectif). Toutefois, et comme nous allons le montrer au cours de cette thèse, des améliorations plus poussées peuvent être proposées afin de mieux encore appréhender la complexité de la densité a posteriori.

B.4 Objectif de la thèse et résumé des chapitres à venir

A ce stade, une première remarque intéressante est que les méthodes d'Inférence Variationnelle construisent une suite de densités de probabilités qui est progressivement améliorée afin de minimiser un certain critère. En ce sens, elles peuvent être vues comme un exemple d'étape 3 dans l'Algorithme 16.

Parmi ces techniques, les méthodes d'Inférence Variationnelle de type Black-Box (minimisant la Kullback-Leibler exclusive ou plus généralement la α-divergence) attirent particulièrement l'attention, les échantillons utilisés dans la construction de la suite de propositions $(k(\theta_n, \cdot))_{n\geqslant 1}$ pouvant également servir à approcher des intégrales de la forme (B.1).

Il paraît dès lors pertinent de mettre à contribution les techniques vues en Inférence Variationnelle dans le but d'améliorer les algorithmes de Monte Carlo adaptatifs. Les performances des méthodes d'Inférence Variationnelle étant limitées par le choix de la famille variationnelle $\mathcal{Q}$ ainsi que celui de la mesure de dissimilarité D, nous pourrions alors nous demander : dans quelle mesure est-il possible d'enrichir $\mathcal{Q}$ par-delà le cadre de l'Inférence Variationnelle de type Black-Box pour de la minimisation de la α-divergence tout en maintenant des procédures d'optimisation efficaces ?

Afin de répondre à cette question, cette thèse s'attache à construire de nouveaux algorithmes d'Inférence Variationnelle pour de la minimisation de α-divergences (i) pouvant être utilisés en échantillonnage préférentiel adaptatif et (ii) augmentant le degré d'expressivité de la famille variationnelle $\mathcal{Q}$.

Plus précisément, nos travaux se décomposent en trois chapitres, qui sont basés sur trois articles distincts. Le fil rouge entre ces trois travaux est que nous nous sommes intéressés à la construction d'algorithmes d'Inférence Variationnelle itératifs entraînant une *décroissance systématique* de la α-divergence à chaque étape. Nous fournissons un bref résumé de chacun de ces chapitres ci-après.

- Chapitre 2 Daudel, Douc, and Portier, 2021.
"Infinite-dimensional gradient-based descent for Alpha-divergence minimisation". *A paraître dans Annals of Statistics*.

Nous introduisons la (α, Γ)-descent, un nouvel algorithme itératif agissant sur les mesures et minimisant la α-divergence. Cette procédure basée sur le gradient étend

la famille variationnelle usuelle en ajoutant un a priori sur les paramètres variationnels sous la forme d'une mesure. Nous montrons qu'elle entraîne une décroissance systématique de la α-divergence pour une grande famille de fonctions Γ et nous obtenons des résultats de convergence. Nous recouvrons l'algorithme Entropic Mirror Descent comme cas particulier de cette procédure et nous présentons une alternative à cet algorithme appelée la Power Descent. Un aspect remarquable de ce travail est qu'en faisant appel à des approximations de Monte Carlo, ces deux algorithmes permettent d'optimiser les poids de mélange de n'importe quel modèle de mélange, sans requérir d'information sur la distribution des paramètres variationnels. Nous démontrons empiriquement les bénéfices de l'approche Power Descent par rapport à l'Entropic Mirror Descent lorsque la dimension augmente.

- Chapitre 3 Daudel and Douc, 2021.

"Mixture weights optimisation for Alpha-divergence Variational Inference".
Soumis en tant qu'article de conférence au moment de la rédaction du manuscrit.

Nous établissons la preuve complète de la convergence de la Power Descent vers les poids de mélange optimaux lorsque $\alpha < 1$. Comme cet algorithme pour la minimisation de α-divergences est défini pour tout $\alpha \in \mathbb{R} \setminus \{1\}$ et ne couvre donc pas le cas classique de la minimisation de la Kullback-Leibler exclusive ($\alpha = 1$), nous l'étendons au cas $\alpha = 1$ et nous montrons que nous retrouvons un algorithme d'Entropic Mirror Descent. Ceci nous amène à étudier plus en détail les liens entre Power Descent et Entropic Mirror Descent : des approximations de premier ordre nous permettent alors de dépasser le cadre de la (α, Γ)-descent et d'introduire la Renyi Descent, un nouvel algorithme pour lequel nous établissons une vitesse de convergence en $O(1/N)$. Enfin, nous comparons empiriquement le comportement de la Power Descent (algorithme non-biaisé) à celui de la Renyi Descent (algorithme biaisé) avant de discuter des avantages potentiels de ces algorithmes l'un par rapport à l'autre.

- Chapitre 4 Daudel, Douc, and Roueff, 2021.

"Monotonic Alpha-divergence minimisation".
Soumis en tant qu'article de journal au moment de la rédaction du manuscrit.

Nous proposons une méthodologie complète permettant la minimisation de α- divergences par décroissance systématique de la α-divergence à chaque étape. Dans sa forme la plus générale, notre travail nous permet de mettre à jour simultanément les poids de mélange et les paramètres des composantes d'un modèle de mélange donné. Notre approche nous permet d'améliorer plusieurs algorithmes déjà utilisés dans le cadre de la minimisation de α-divergence tels des algorithmes de Gradient Descent et de Power Descent. De plus, nous revisitons sous un angle neuf et généralisons un algorithme d'Expectation-Maximisation intégré. Enfin, en considérant le cas particulier des modèles de mélange Gaussiens et en faisant appel à des approximations de Monte Carlo, nous démontrons empiriquement que notre méthodologie

apporte des améliorations numériques, tout en illustrant les bénéfices pratiques liés à la flexibilité nouvelle acquise au travers de l'hyperparamètre α de la α-divergence.

Bibliography

Alquier, Pierre and James Ridgway (2020). "Concentration of tempered posteriors and of their variational approximations". In: *The Annals of Statistics* 48.3, pp. 1475 –1497. DOI: 10.1214/19-AOS1855. URL: https://doi.org/10.1214/19-AOS1855.

Alquier, Pierre, James Ridgway, and Nicolas Chopin (2016). "On the properties of variational approximations of Gibbs posteriors". In: *Journal of Machine Learning Research* 17.236, pp. 1–41. URL: http://jmlr.org/papers/v17/15-290.html.

Beck, Amir and Marc Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3, pp. 167 –175. ISSN: 0167-6377. DOI: https://doi.org/10.1016/S0167-6377(02)00231-6. URL: http://www.sciencedirect.com/science/article/pii/S0167637702002316.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518, 859–877. ISSN: 1537-274X. DOI: 10.1080/01621459.2017.1285773. URL: http://dx.doi.org/10.1080/01621459.2017.1285773.

Blei, David M., Andrew Y. Ng, and Michael Jordan (2003). "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research*. Vol. 3, pp. 993–1022.

Bottou, Léon (2010). "Large-Scale Machine Learning with Stochastic Gradient Descent". In: *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*. Ed. by Yves Lechevallier and Gilbert Saporta. Paris, France: Springer, pp. 177–187. URL: http://leon.bottou.org/papers/bottou-2010.

Bubeck, Sébastien (2015). "Convex Optimization: Algorithms and Complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4, pp. 231–357. ISSN: 1935-8237. DOI: 10.1561/2200000050. URL: http://dx.doi.org/10.1561/2200000050.

Buchholz, Alexander, Florian Wenzel, and Stephan Mandt (2018). "Quasi-Monte Carlo Variational Inference". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of

Machine Learning Research. PMLR, pp. 668–677. URL: http://proceedings.mlr.press/v80/buchholz18a.html.

Bugallo, Monica F. et al. (2017). "Adaptive Importance Sampling: The past, the present, and the future". In: *IEEE Signal Processing Magazine* 34.4, pp. 60–79. DOI: 10.1109/MSP.2017.2699226.

Burda, Yuri, Roger B. Grosse, and Ruslan Salakhutdinov (2016). "Importance weighted autoencoders". In: *International Conference on Learning Representations (ICLR)*.

Campbell, Trevor and Xinglong Li (2019). "Universal Boosting Variational Inference". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., pp. 3484–3495. URL: https://proceedings.neurips.cc/paper/2019/file/07a4e20a7bbeeb7a736682b26b16ebe8-Paper.pdf.

Cappé, Olivier et al. (2008). "Adaptive importance sampling in general mixture classes". In: *Statistics and Computing* 18.4, pp. 447–459.

Chopin, Nicolas (2004). "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference". In: *Ann. Statist.* 32.6, pp. 2385–2411. DOI: 10.1214/009053604000000698. URL: https://doi.org/10.1214/009053604000000698.

Chérief-Abdellatif, Badr-Eddine, Pierre Alquier, and Mohammad Emtiyaz Khan (2019). "A Generalization Bound for Online Variational Inference". In: *Proceedings of the 29th International Conference on Machine Learning*. Vol. 101, pp. 662–677.

Cichocki, Andrzej and Shun-ichi Amari (2010). "Families of Alpha- Beta- and Gamma-Divergences: Flexible and Robust Measures of Similarities". In: *Entropy* 12.6, 1532–1568. ISSN: 1099-4300. DOI: 10.3390/e12061532. URL: http://dx.doi.org/10.3390/e12061532.

Cichocki, Andrzej, Sergio Cruces, and Shun-ichi Amari (2011). "Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization". In: *Entropy* 13.1, 134–170. ISSN: 1099-4300. DOI: 10.3390/e13010134. URL: http://dx.doi.org/10.3390/e13010134.

Daudel, Kamélia and Randal Douc (2021). "Mixture weights optimisation for Alpha-Divergence Variational Inference". In: *Submitted*. URL: http://arxiv.org/abs/2106.05114.

Daudel, Kamélia, Randal Douc, and François Portier (2021). "Infinite-dimensional gradient-based descent for alpha-divergence minimisation". In: *To appear in the Annals of Statistics*. arXiv: 2005.10618 [math.ST].

Daudel, Kamélia, Randal Douc, and François Roueff (2021). "Monotonic Alpha- divergence Minimisation". In: *Submitted*. arXiv: 2103.05684 [stat.CO].

Delyon, Bernard and François Portier (2021). "Safe adaptive importance sampling: A mixture approach". In: *The Annals of Statistics* 49.2, pp. 885 –917. DOI: 10.1214/20-AOS1983. URL: https://doi.org/10.1214/20-AOS1983.

Dhaka, Akash Kumar et al. (2021). "Challenges and Opportunities in High-dimensional Variational Inference". In: *arxiv preprint arxiv:2103.01085*. arXiv: 2103.01085 [cs.LG]. URL: https://arxiv.org/abs/2103.01085.

Dieng, Adji Bousso et al. (2017). "Variational Inference via \chi Upper Bound Minimization". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 2732–2741. URL: http://papers.nips.cc/paper/6866-variational-inference-via-chi-upper-bound-minimization.pdf.

Domke, Justin (2019). "Provable Gradient Variance Guarantees for Black-Box Variational Inference". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/bd4c9ab730f5513206b999ec0d90d1fb-Paper.pdf.

— (2020). "Provable Smoothness Guarantees for Black-Box Variational Inference". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 2587–2596. URL: http://proceedings.mlr.press/v119/domke20a.html.

Douc, Randal et al. (2007a). "Convergence of adaptive mixtures of importance sampling schemes". In: *Ann. Statist.* 35.1, pp. 420–448. DOI: 10.1214/009053606000001154. URL: https://doi.org/10.1214/009053606000001154.

Douc, Randal et al. (2007b). "Minimum variance importance sampling via population Monte Carlo". en. In: *ESAIM: Probability and Statistics* 11, pp. 427–447. DOI: 10.1051/ps:2007028. URL: http://www.numdam.org/item/PS_2007__11__427_0/.

Doucet, Arnaud, Nando de Freitas, and Neil J. Gordon, eds. (2001). *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer. ISBN: 978-1-4419-2887-0. DOI: 10.1007/978-1-4757-3437-9. URL: https://doi.org/10.1007/978-1-4757-3437-9.

Erven, Tim van and Peter Harremoes (2014). "Rényi Divergence and Kullback-Leibler Divergence". In: *IEEE Transactions on Information Theory* 60.7, 3797–3820. ISSN: 1557-9654. DOI: 10.1109/tit.2014.2320500. URL: http://dx.doi.org/10.1109/TIT.2014.2320500.

Geffner, Tomas and Justin Domke (2020a). "Empirical Evaluation of Biased Methods for Alpha Divergence Minimization". In: *3rd Symposium on Advances in Approximate Bayesian Inference*, pp. 1–12. URL: https://openreview.net/pdf?id=ihUcld16Mpu.

— (2020b). *On the Difficulty of Unbiased Alpha Divergence Minimization*. arXiv: 2010.09541 [stat.ML].

Gershman, Samuel, Matthew D. Hoffman, and David M. Blei (2012). "Nonparametric variational inference". In: *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh, Scotland, UK.

Geweke, John (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration". In: *Econometrica* 57.6, pp. 1317–1339. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913710.

Hernandez-Lobato, Jose et al. (2016). "Black-Box Alpha Divergence Minimization". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1511–1520. URL: http://proceedings.mlr.press/v48/hernandez-lobatob16.html.

Hoffman, Matthew D. et al. (2013). "Stochastic Variational Inference". In: *Journal of Machine Learning Research* 14.4, pp. 1303–1347. URL: http://jmlr.org/papers/v14/hoffman13a.html.

Hsieh, Ya-Ping, Chen Liu, and Volkan Cevher (2019). "Finding Mixed Nash Equilibria of Generative Adversarial Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 2810–2819. URL: http://proceedings.mlr.press/v97/hsieh19b.html.

Jaakkola, Tommi S. and Michael I. Jordan (1998). "Improving the Mean Field Approximation via the Use of Mixture Distributions". In: *Jordan M.I. (eds) Learning in Graphical Models. NATO ASI Series (Series D: Behavioural and Social Sciences), Springer* 89.

Jordan, Michael I. et al. (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2, pp. 183–233. ISSN: 1573-0565. DOI: 10.1023/A:1007665907178. URL: https://doi.org/10.1023/A:1007665907178.

Kingma, Diederik P and Max Welling (2014). "Auto-Encoding Variational Bayes". In: arXiv: 1312.6114 [stat.ML].

Kloek, Teun and Herman K Van Dijk (1978). "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo". In: *Econometrica* 46.1, pp. 1–19. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913641.

Kuleshov, Volodymyr and Stefano Ermon (2017). "Neural Variational Inference and Learning in Undirected Graphical Models". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/14e422f05b68cc0139988e128ee880df-Paper.pdf.

Kullback, S. and R. A. Leibler (1951). "On Information and Sufficiency". In: *Ann. Math. Statist.* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694. URL: https://doi.org/10.1214/aoms/1177729694.

Li, Yingzhen and Richard E Turner (2016). "Rényi Divergence Variational Inference". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al.

Curran Associates, Inc., pp. 1073–1081. URL: http://papers.nips.cc/paper/6208-renyi-divergence-variational-inference.pdf.

Mandt, Stephan, Matthew D. Hoffman, and David M. Blei (2017). "Stochastic Gradient Descent as Approximate Bayesian Inference". In: *Journal of Machine Learning Research* 18.134, pp. 1–35. URL: http://jmlr.org/papers/v18/17-214.html.

Martino, Luca et al. (May 2017). "Layered Adaptive Importance Sampling". In: *Statistics and Computing* 27.3, 599–623. ISSN: 0960-3174. DOI: 10.1007/s11222-016-9642-5. URL: https://doi.org/10.1007/s11222-016-9642-5.

Minka, Thomas P. (2001). "Expectation Propagation for Approximate Bayesian Inference". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. Seattle, Washington: Morgan Kaufmann Publishers Inc., pp. 362–369. ISBN: 1-55860-800-1. URL: http://dl.acm.org/citation.cfm?id=2074022.2074067.

Minka, Tom (2004). *Power EP*. Tech. rep. MSR-TR-2004-149, p. 6. URL: https://www.microsoft.com/en-us/research/publication/power-ep/.

— (2005). *Divergence Measures and Message Passing*. Tech. rep. MSR-TR-2005-173, p. 17. URL: https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/.

Morimoto, Tetsuzo (1963a). "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten". In: *Magyar Tud. Akad. Mat. Kutat Int.*, 85–108.

— (1963b). "Markov Processes and the H-Theorem". In: *Journal of the Physical Society of Japan* 18.3, pp. 328–331.

Naesseth, Christian, Fredrik Lindsten, and David Blei (2020). "Markovian Score Climbing: Variational Inference with KL(p|| q)". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 15499–15510. URL: https://proceedings.neurips.cc/paper/2020/file/b20706935de35bbe643733f856d9e5d6-Paper.pdf.

Naesseth, Christian et al. (2018). "Variational Sequential Monte Carlo". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, pp. 968–977. URL: http://proceedings.mlr.press/v84/naesseth18a.html.

Neal, Radford M. (1993). *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Tech. rep.

Nemirovski, Arkadi (2004). "Prox-method with rate of convergence O(1/T) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems". In: *SIAM Journal on Optimization* 15, pp. 229–251.

Nemirovski, Arkadi. et al. (2009). "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4, pp. 1574–1609.

DOI: 10.1137/070704277. eprint: https://doi.org/10.1137/070704277. URL: https://doi.org/10.1137/070704277.

Oh, Man-Suk and James O. Berger (1992). "Adaptive importance sampling in monte carlo integration". In: *Journal of Statistical Computation and Simulation* 41.3-4, pp. 143–168. DOI: 10.1080/00949659208810398. eprint: https://doi.org/10.1080/00949659208810398. URL: https://doi.org/10.1080/00949659208810398.

Paisley, John, David Blei, and Michael Jordan (2012). "Variational Bayesian Inference with Stochastic Search". In: *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh, Scotland, UK, 1363–1370.

Portier, François and Bernard Delyon (2018). "Asymptotic optimality of adaptive importance sampling". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/1bc0249a6412ef49b07fe6f62e6dc8de-Paper.pdf.

Ranganath, Rajesh, Sean Gerrish, and David Blei (2014). "Black Box Variational Inference". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Ed. by Samuel Kaski and Jukka Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, pp. 814–822. URL: http://proceedings.mlr.press/v33/ranganath14.html.

Ranganath, Rajesh, Dustin Tran, and David Blei (2016). "Hierarchical Variational Models". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 324–333. URL: http://proceedings.mlr.press/v48/ranganath16.html.

Rényi, Alfréd (1961). "On Measures of Entropy and Information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 547–561. URL: https://projecteuclid.org/euclid.bsmsp/1200512181.

Robbins, Herbert and Sutton Monro (1951). "A Stochastic Approximation Method". In: *Ann. Math. Statist.* 22.3, pp. 400–407. DOI: 10.1214/aoms/1177729586. URL: https://doi.org/10.1214/aoms/1177729586.

Robert, Christian P. and George Casella (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387212396.

Royden, H.L. and P. Fitzpatrick (2010). *Real Analysis (4th Edtion)*. Prentice Hall. ISBN: 9780131437470. URL: https://books.google.fr/books?id=0Y5fAAAACAAJ.

Sason, Igal (2018). "On f-Divergences: Integral Representations, Local Behavior, and Inequalities". In: *Entropy* 20.5, p. 383. ISSN: 1099-4300. DOI: 10.3390/e20050383. URL: http://dx.doi.org/10.3390/e20050383.

Stone, Charles J. (Dec. 1982). "Optimal Global Rates of Convergence for Nonparametric Regression". In: *Ann. Statist.* 10.4, pp. 1040–1053. DOI: 10.1214/aos/1176345969. URL: https://doi.org/10.1214/aos/1176345969.

Thin, Achille et al. (2020). *MetFlow: A New Efficient Method for Bridging the Gap between Markov Chain Monte Carlo and Variational Inference*. arXiv: 2002.12253 [stat.ML].

Titsias, Michalis K. and Francisco Ruiz (2019). "Unbiased Implicit Variational Inference". In: *Proceedings of Machine Learning Research*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 167–176. URL: http://proceedings.mlr.press/v89/titsias19a.html.

Wang, Dilin, Hao Liu, and Qiang Liu (2018). "Variational Inference with Tail-adaptive f-Divergence". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 5737–5747. URL: http://papers.nips.cc/paper/7816-variational-inference-with-tail-adaptive-f-divergence.pdf.

Williams, Ronald J. (1992). "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". In: *Mach. Learn.* 8, pp. 229–256. DOI: 10.1007/BF00992696. URL: https://doi.org/10.1007/BF00992696.

Yao, Yuling et al. (2018). "Yes, but Did It Work?: Evaluating Variational Inference". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, pp. 5581–5590. URL: http://proceedings.mlr.press/v80/yao18a.html.

Yin, Mingzhang and Mingyuan Zhou (2018). "Semi-Implicit Variational Inference". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, pp. 5660–5669. URL: http://proceedings.mlr.press/v80/yin18b.html.

Zhang, Cheng et al. (2019). "Advances in Variational Inference". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8, pp. 2008–2026. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2889774.

Zhu, Huaiyu and Richard Rohwer (1995a). "Bayesian invariant measurements of generalization". In: *Neural Processing Letters* 2, pp. 28–31. DOI: 10.1007/BF02309013.

— (1995b). *Information Geometric Measurements of Generalisation*. Tech. rep. NCRG/4350.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Méthodes de Monte Carlo adaptatives pour les modèles complexes

Mots clés : Méthodes de Monte Carlo, Inférence Variationnelle, Alpha-divergence

Résumé : Cette thèse s'inscrit dans le domaine de l'Inférence Statistique et plus précisément dans le cadre de l'Inférence Bayésienne, dont le but est de modéliser un phénomène à partir d'un jeu de données tout en incorporant des connaissances a priori sur les paramètres de ce modèle.
L'émergence des données massives nécessite le recours à des modèles bayésiens complexes à même de décrire la structure de ces données. De tels modèles requièrent à leur tour la construction et l'étude d'algorithmes adaptatifs capables de traiter de larges volumes de données lorsque les paramètres du modèle choisi évoluent dans un espace en grande dimension.
Deux catégories principales de méthodes tentent de répondre à cette problématique : les méthodes de Monte Carlo, s'appuyant sur de l'échantillonnage, et les méthodes d'Inférence Variationnelle, reposant sur de l'optimisation. En faisant appel à la littérature de l'optimisation et plus récemment aux méthodes de Monte Carlo, des avancées majeures en Inférence Variationnelle ont permis de lever une partie des obstacles computationnels rencontrés en Inférence Bayésienne.
Toutefois, les résultats théoriques et empiriques des méthodes d'Inférence Variationnelle sont souvent affectés par : (i) un choix inapproprié de la fonction objectif apparaissant dans le problème d'optimisation et (ii) un espace de recherche ne contenant pas la cible car trop restreint.
Dans cette thèse, nous cherchons à remédier aux deux difficultés susmentionnées en construisant des algorithmes adaptatifs applicables aux modèles complexes et se situant à l'intersection des méthodes de Monte Carlo et d'Inférence Variationnelle.
Nos travaux suggèrent d'utiliser la α-divergence comme fonction objectif plus générale et proposent d'enrichir l'espace de recherche par-delà les schémas traditionnels utilisés en Inférence Variationnelle. La spécificité de notre approche réside dans l'obtention de nouveaux algorithmes avantageux d'un point de vue numérique et bénéficiant également de solides fondements théoriques, qui se manifestent au travers d'une décroissance systématique de la α-divergence à chaque étape de nos algorithmes. En outre, nos travaux mettent en lumière d'importants liens entre les méthodes de Monte Carlo et celles d'Inférence Variationnelle.

Title : Adaptive Monte Carlo methods for complex models

Keywords : Monte Carlo methods, Variational Inference, Alpha-divergence

Abstract : This thesis lies in the field of Statistical Inference and more precisely in Bayesian Inference, where the goal is to model a phenomenon given some data while taking into account prior knowledge on the model parameters.
The availability of large datasets sparked the interest in using complex models for Bayesian Inference tasks that are able to capture potentially complicated structures inside the data. Such a context requires the development and study of adaptive algorithms that can efficiently process large volumes of data when the dimension of the model parameters is high.
Two main classes of methods attempt to fulfil this role : sampling-based Monte Carlo methods and optimisation-based Variational Inference methods. By relying on the optimisation literature and more recently on Monte Carlo methods, the latter have made it possible to construct fast algorithms that overcome some of the computational hurdles encountered in Bayesian Inference.
Yet, the theoretical results and empirical performances of Variational Inference methods are often impacted by two factors : one, an inappropriate choice of the objective function appearing in the optimisation problem and two, a search space that is too restrictive to match the target at the end of the optimisation procedure.
This thesis explores how we can remedy the two issues mentioned above in order to build improved adaptive algorithms for complex models at the intersection of Monte Carlo and Variational Inference methods.
In our work, we suggest selecting the α-divergence as a more general class of objective functions and we propose several ways to enlarge the search space beyond the traditional framework used in Variational Inference. The specificity of our approach in this thesis is then that it derives numerically advantageous adaptive algorithms with strong theoretical foundations, in the sense that they provably ensure a systematic decrease in the α-divergence at each step. In addition, we unravel important connections between the sampling-based and the optimisation-based methodologies.

Institut Polytechnique de Paris
91120 Palaiseau, France