Mixture weights optimisation for Alpha-Divergence Variational Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper focuses on α-divergence minimisation methods for Variational Inference. More precisely, we are interested in algorithms optimising the mixture weights of any given mixture model, without any information on the underlying distribution of its mixture components parameters. The Power Descent, defined for all $\alpha \neq 1$, is one such algorithm and we establish in our work the full proof of its convergence towards the optimal mixture weights when $\alpha < 1$. Since the α-divergence recovers the widely-used forward Kullback-Leibler when $\alpha \to 1$, we then extend the Power Descent to the case $\alpha = 1$ and show that we obtain an Entropic Mirror Descent. This leads us to investigate the link between Power Descent and Entropic Mirror Descent: first-order approximations allow us to introduce the Renyi Descent, a novel 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.

1 Introduction

Bayesian Inference involves being able to compute or sample from the posterior density. For many useful models, the posterior density can only be evaluated up to a normalisation constant and we must resort to approximation methods.

One major category of approximation methods is Variational Inference, a wide class of optimisation methods which introduce a simpler density family $\mathcal{Q}$ and use it to approximate the posterior density (see for example Variational Bayes [1, 2] and Stochastic Variational Inference [3]). The crux of these methods consists in being able to find the best approximation of the posterior density among the family $\mathcal{Q}$ in the sense of a certain divergence, most typically the Kullback-Leibler divergence. However, The Kullback-Leibler divergence is known to have some undesirable properties (e.g posterior overestimation/underestimation [4]) and as a consequence, the α-divergence [5, 6] and Renyi's α-divergence [7, 8] have gained a lot of attention recently as a more general alternative [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

Noticeably, [17] introduced the (α, Γ)-descent, a general family of gradient-based algorithms that are able to optimise the *mixture weights* of mixture models 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, in an Sequential Monte Carlo fashion [20], to select the mixture components according to their overall importance in the set of component parameters. From there, one is able to optimise the weights and the components parameters alternatively [17]. 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 $\alpha \in \mathbb{R} \setminus \{1\}$ 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 [17] showed that the Power Descent outperforms the Entropic Mirror Descent when $\alpha < 1$ as the dimension increases.

Nonetheless, the global convergence of the Power Descent algorithm when $\alpha < 1$, as stated in [17], 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 $\alpha \geqslant 1$ ([4] and [17]) and which often may lead to posterior variance underestimation. We are thus interested in studying Variational Inference methods for optimising the mixture weights of mixture models when $\alpha < 1$. After recalling the basics of the Power Descent algorithm in Section 2, we make the following contributions in the paper:

- In Section 3, we derive the full convergence proof of the Power Descent algorithm towards the optimal mixture weights when $\alpha < 1$ (Theorem 2).
- Since the α-divergence becomes the traditional forward Kullback-Leibler when $\alpha \to 1$, we first bridge in Section 4 the gap between the cases $\alpha < 1$ and $\alpha > 1$ of the Power Descent: we obtain that the Power Descent recovers an Entropic Mirror Descent performing forward Kullback-Leibler minimisation (Proposition 1). 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 that we call the *Renyi Descent* and that is proved in Theorem 3 to converge at an $O(1/N)$ rate towards its optimum for all $\alpha \in \mathbb{R}$.
- Finally, we run some numerical experiments in Section 5 to compare the behavior of the Power Descent and the Renyi Descent altogether, before discussing the potential benefits of one approach over the other.

2 Background

We start by introducing some notation. 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 probabilistic model $p(\mathscr{D}|y)$ parameterised by a hidden random variable $y \in \mathsf{Y}$ that is drawn from a certain prior $p_0(y)$. The posterior density of the latent variable y given the data $\mathscr{D}$ is then given by:

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

where the normalisation constant $p(\mathscr{D}) = \int_{\mathsf{Y}} p_0(y)p(\mathscr{D}|y)\nu(\mathrm{d}y)$ is called the *marginal likelihood* or *model evidence* and is oftentimes unknown.

To approximate the posterior density, the Power Descent considers a variational family $\mathcal{Q}$ that is large enough to contain mixture models and that we redefine now: letting $(\mathsf{T}, \mathcal{T})$ be a measurable space, $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ be a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density k defined on $\mathsf{T} \times \mathsf{Y}$, the Power Descent considers the following approximating family

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

where M is a convenient subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on $(\mathsf{T}, \mathcal{T})$. This choice of approximating family extends the typical parametric family commonly-used in Variational Inference since it amounts to putting a prior over the parameter θ (in the form of a measure) and does describe the class of mixture models when μ is a weighted sum of Dirac measures.

Problem statement Denote by $\mathbb{P}$ the probability measure on $(\mathsf{Y}, \mathcal{Y})$ with corresponding density $p(\cdot|\mathscr{D})$ with respect to ν and for all $\mu \in \mathrm{M}_1(\mathsf{T})$, for all $y \in \mathsf{Y}$, denote $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y)$. Furthermore, given $\alpha \in \mathbb{R}$, let 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 $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Then, the α-divergence between μK and $\mathbb{P}$ (extended by continuity to the cases $\alpha = 0$ and $\alpha = 1$ as for example done in [21]) is given by

$$D_\alpha(\mu K||\mathbb{P}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \;,$$

and the goal of the Power Descent is to find

$$\mathrm{arginf}_{\mu\in\mathsf{M}} D_\alpha(\mu K||\mathbb{P}) \;. \tag{1}$$

More generally, letting p be any measurable positive function on $(\mathsf{Y},\mathcal{Y})$, the Power Descent aims at solving

$$\mathrm{arginf}_{\mu\in\mathsf{M}} \Psi_\alpha(\mu;p) \;, \tag{2}$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $\Psi_\alpha(\mu;p) = \int_{\mathsf{Y}} f_\alpha\left(\mu k(y)/p(y)\right) p(y)\nu(\mathrm{d}y)$. The Variational Inference optimisation problem (1) can then be seen as an instance of (2) that is equivalent to optimising $\Psi_\alpha(\mu;p)$ with $p(y) = p(y,\mathscr{D})$ (see Appendix A.1). In the following, the dependency on p in Ψ_α may be dropped throughout the paper for notational ease when no ambiguity occurs and we now present the Power Descent algorithm.

The Power Descent algorithm. The optimisation problem (2) can be solved for all $\alpha \in \mathbb{R}\setminus\{1\}$ by using the Power Descent algorithm introduced in [17] : given an initial measure $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi_\alpha(\mu_1) < \infty$, $\alpha \in \mathbb{R}\setminus\{1\}$, $\eta > 0$ and κ such that $(\alpha-1)\kappa \geqslant 0$, the Power descent algorithm is an iterative scheme which builds the sequence of probability measures $(\mu_n)_{n\in\mathbb{N}^\star}$

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \;, \qquad n \in \mathbb{N}^\star \;, \tag{3}$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$, the one-step transition $\mu \mapsto \mathcal{I}_\alpha(\mu)$ is given by Algorithm 1 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 1: *Power descent one-step transition* $(\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)})$

---

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

In this algorithm, $b_{\mu,\alpha}$ can be understood as the gradient of Ψ_α. Algorithm 1 then consists in applying the transform function Γ to the translated gradient $b_{\mu,\alpha}+\kappa$ and projecting back onto the space of probability measures.

A remarkable property of the Power Descent algorithm, which has been proven in [17] (it is a special case of [17, Theorem 1] with $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$), is that under (A1) as defined below

(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 $\eta \in (0,1]$ (this result is recalled in Theorem 4 of Appendix A.2 for the sake of completeness). Under the additional assumptions that $\kappa > 0$ and

$$\sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}| < \infty \quad \text{and} \quad \Psi_\alpha(\mu_1) < \infty \;, \tag{4}$$

the Power Descent is also known to converge towards its optimal value at an $O(1/N)$ rate when $\alpha > 1$ [17, Theorem 3]. On the other hand, when $\alpha < 1$, the convergence towards the optimum as written in [17] holds under different assumptions including

(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 [17, Theorem 4], that is recalled below under the form of Theorem 1, states the convergence of the Power Descent algorithm towards the global optimum.

Theorem 1 ([17, Theorem 4])**.** *Assume* (A1) *and* (A2)*. Let $\alpha < 1$ and let $\kappa \leqslant 0$. Then, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $\Psi_\alpha(\mu) < \infty$ 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) *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) $(\Psi_\alpha(\mu_n))_{n \in \mathbb{N}^\star}$ *is nonincreasing,*

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

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

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) 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 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 1 that shall be retained in our upcoming convergence result.

A first comment is that (A1) is mild since the assumption that $p(y) > 0$ for all $y \in \mathsf{Y}$ can be discarded and is kept for convenience [17, Remark 4]. A second comment is that (A2) is also mild and covers (4) as it amounts to assuming that $b_{\mu,\alpha}(\theta)$ and $\Psi_\alpha(\mu)$ are uniformly bounded with respect to μ and θ. To see this, we give below an example for which (A2) is satisfied.

Example 1. *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 $p(y) = c \times [0.5\mathcal{N}(y; \theta_1^\star, \boldsymbol{I_d}) + 0.5\mathcal{N}(y; \theta_2^\star, \boldsymbol{I_d})]$ for all $y \in \mathsf{Y}$ where $\theta_1^\star, \theta_2^\star \in \mathsf{T}$ and $\boldsymbol{I_d}$ is the identity matrix. Then,* (A2) *holds (see Appendix B.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 $\mathbb{R}^J$:

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

and we also define $\mathcal{S}_J^+ = \{\boldsymbol{\lambda} \in \mathcal{S}_J \ : \ \forall j \in \{1, \ldots, J\} \, , \ \lambda_j > 0\}$. We let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$ be fixed and for all $\boldsymbol{\lambda} \in \mathcal{S}_J$, we define $\mu_{\boldsymbol{\lambda},\Theta} \in \mathrm{M}_1(\mathsf{T})$ by $\mu_{\boldsymbol{\lambda},\Theta} = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$.

Consequently, $\mu_{\boldsymbol{\lambda},\Theta} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y)$ corresponds to a mixture model and if we let $(\mu_n)_{n \in \mathbb{N}^\star}$ be defined by $\mu_1 = \mu_{\boldsymbol{\lambda},\Theta}$ and (3), 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 $\boldsymbol{\lambda}_n = (\lambda_{1,n}, \ldots, \lambda_{J,n}) \in \mathcal{S}_J$ satisfies the initialisation $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}$ and the update formula:

$$\boldsymbol{\lambda}_{n+1} = \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_n) \, , \ n \in \mathbb{N}^\star \, , \tag{5}$$

where for all $\boldsymbol{\lambda} \in \mathcal{S}_J$,

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

with $\Gamma(v) = [(\alpha-1)v+1]^{\frac{\eta}{1-\alpha}}$ for all $v \in \mathrm{Dom}_\alpha$. Finally, let us rewrite (A2) in the simplified case where μ is a sum of Dirac measures, which gives (A3) below.

(A3) (i) For all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;

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

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

We then have the following theorem, which establishes the full proof of the global convergence towards the optimum for the mixture weights when $\alpha < 1$.

Theorem 2. *Assume* (A1) *and* (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 $\boldsymbol{\lambda} \in \mathcal{S}_J$, $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}) < \infty$ 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 (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) $(\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n,\Theta}))_{n \in \mathbb{N}^\star}$ *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) $\Psi_\alpha(\mu_{\boldsymbol{\lambda}_\star,\Theta}) = \inf_{\boldsymbol{\lambda}' \in \mathcal{S}_J} \Psi_\alpha(\mu_{\boldsymbol{\lambda}',\Theta})$.

The proof of this result builds on Theorem 1 and Theorem 4 and is deferred to Appendix B.2. Notice that since Ψ_α depends on $\boldsymbol{\lambda}$ through $\mu_{\boldsymbol{\lambda},\Theta}K$ in Theorem 2, 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 1, 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 $(b_{\mu_{\boldsymbol{\lambda}_n,\Theta},\alpha}(\theta_j))_{1 \leqslant j \leqslant J}$ in the update formula (5) (see Algorithm 3 of Appendix B.3).

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 $\alpha \to 1$, 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 $\mu = \mu_{\boldsymbol{\lambda},\Theta}$ and $\alpha < 1$ (and more generally when $\mu \in \mathrm{M}_1(\mathsf{T})$ and $\alpha \in \mathbb{R}$).

4 Power Descent and Entropic Mirror Descent

Recall from Section 2 that the Power Descent is defined for all $\alpha \in \mathbb{R} \setminus \{1\}$. In this section, we first establish in Proposition 1 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 runs 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) and which are deferred to Appendix C.1, alongside with the proof of Proposition 1.

Proposition 1 (Limiting case $\alpha \to 1$)**.** *Assume* (A1) *and* (D1). *Then, for 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

$$\mathcal{I}_1(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta)e^{-\eta b_{\mu,1}(\theta)}}{\mu\left(e^{-\eta b_{\mu,1}}\right)} \quad \textit{and} \quad b_{\mu,1}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \log\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y) \ . \tag{6}$$

Here, we recognise the one-step transition associated to the Entropic Mirror Descent applied to Ψ_1. This algorithm is a special case of [17] with $\Gamma(v) = e^{-\eta v}$ and $\alpha = 1$ and as such, it is known to lead to a systematic decrease in the forward Kullback-Leibler divergence and to enjoy an $O(1/N)$ convergence rate under the assumptions that (4) holds and $\eta \in (0,1)$ [17, 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

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta)\left[(\alpha-1)(b_{\mu_n,\alpha}(\theta)+\kappa)+1\right]^{\frac{\eta}{1-\alpha}}}{\mu_n(\left[(\alpha-1)(b_{\mu_n,\alpha}+\kappa)+1\right]^{\frac{\eta}{1-\alpha}})} \;, \quad n \in \mathbb{N}^\star \;.$$

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

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta)\exp\left[-\eta'(b_{\mu_n,\alpha}(\theta)+\kappa)\right]}{\mu_n(\exp\left[-\eta'(b_{\mu_n,\alpha}+\kappa)\right])} \;, \quad n \in \mathbb{N}^\star \;, \tag{7}$$

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 (7), we see that this leads us to consider the update formula given by

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta)\exp\left[-\eta\frac{b_{\mu_n,\alpha}(\theta)}{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1}\right]}{\mu_n\left(\exp\left[-\eta\frac{b_{\mu_n,\alpha}}{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1}\right]\right)} \;, \quad n \in \mathbb{N}^\star \;. \tag{8}$$

Observe then that (8) can again be seen as an Entropic Mirror Descent, but applied this time to the objective function defined for all $\alpha \in \mathbb{R} \setminus \{0,1\}$ by

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

meaning 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 Appendix C.2 for the derivation of (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 [9, 10]), we can motivate the iterative scheme (8) by observing that we recover the Variational Renyi bound introduced in [10] up to a constant $-\alpha^{-1}$ when we let $p = p(\cdot, \mathscr{D})$, $\kappa = 0$ and $\alpha > 0$ in Ψ_α^{AR}. For this reason we call the algorithm given by (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 [17] 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

Table 1: Summary of the theoretical results obtained in this paper compared to [17]

	Power Descent	Renyi Descent
[17]	$\alpha < 1$: convergence under restrictive assumptions; $\alpha > 1$: $O(1/N)$ convergence rate	not covered
This paper	$\alpha < 1$: full proof of convergence for mixture weights; extension to $\alpha = 1$ with $O(1/N)$ convergence rate	$O(1/N)$ convergence rate

(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 3. *Assume* (A1) *and* (A4). *Let $\alpha \in \mathbb{R} \setminus \{1\}$ 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 $\Psi_\alpha(\mu_1) < \infty$. Then, the following assertions hold.*

(i) The sequence $(\mu_n)_{n \in \mathbb{N}^\star}$ defined by (8) *is well-defined and the sequence $(\Psi_\alpha(\mu_n))_{n \in \mathbb{N}^\star}$ is non-increasing.*

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

$$\Psi_\alpha(\mu_N) - \Psi_\alpha(\mu^\star) \leqslant \frac{L_{\alpha,2}}{N} \left[KL(\mu^\star || \mu_1) + L \frac{L_{\alpha,3}}{L_{\alpha,1}(\alpha - 1)\kappa} \Delta_1 \right] , \tag{9}$$

where $\mu^\star$ is such that $\Psi_\alpha(\mu^\star) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta)$, $\mathrm{M}_{1,\mu_1}(\mathsf{T})$ denotes the set of probability measures dominated by μ_1, $KL(\mu^\star || \mu_1) = \int_\mathsf{T} \log(\mathrm{d}\mu^\star / \mathrm{d}\mu_1)\,\mathrm{d}\mu^\star$, $\Delta_1 = \Psi_\alpha(\mu_1) - \Psi_\alpha(\mu^\star)$ and $L_{\alpha,2}$, L, $L_{\alpha,3}$, $L_{\alpha,1}$ are finite constants defined in (20).

The proof of this result is deferred to Appendix C.3 and we present in the next example an application of this theorem to the particular case of mixture models.

Example 2. *Let $\alpha \in \mathbb{R} \setminus \{1\}$, 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*

$$\Psi_\alpha(\mu_N) - \Psi_\alpha(\mu^\star) \leqslant \frac{|\alpha - 1|(|B|_{\infty,\alpha} + |\kappa|)}{N} \left[\frac{\log J}{\eta} + \frac{\sqrt{2\log(J)}|B|_{\infty,\alpha}}{(\alpha - 1)\kappa(1 - \eta|\kappa|^{-1}|B|_{\infty,\alpha})} \right] ,$$

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 (20) *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 the Renyi Descent enjoys an $O(1/\sqrt{N})$ convergence rate as a Entropic Mirror Descent algorithm for the sequence $(\Psi_\alpha(N^{-1} \sum_{n=1}^N \mu_n))_{N \in \mathbb{N}^\star}$ under our assumptions when η is proportional to $1/\sqrt{N}$, N being fixed (see [22] or [23, 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 [23, Theorem 6.2] that we do not assume in Theorem 3. 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 $N^{-1} \sum_{n=1}^N \mu_n$.

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 $\mu = \mu_{\boldsymbol{\lambda},\Theta}$ and by resorting to an unbiased estimate of $(b_{\mu,\alpha}(\theta_j))_{1 \leqslant j \leqslant J}$ (see Algorithm 4 of Appendix C.4).

The theoretical results we have obtained are summarised in Table 1 and we next move on to numerical experiments.

5 Simulation study

Let the target p be a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant c such that $p(y) = c \times [0.5\mathcal{N}(y; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(y; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, 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, a fully adaptive algorithm can be 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 2.

Algorithm 2: *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 $\theta_{1,t+1}, \ldots, \theta_{J,t+1}$.

Many choices of Exploration step can be envisioned in Algorithm 2 since there is no constraint on $\{\theta_1, \ldots, \theta_J\}$. Here, we consider the same Exploration step as the one they used in [17]: h is set to be proportional to $J^{-1/(4+d)}$ and the particles are updated by i.i.d sampling according to $\mu_{\boldsymbol{\lambda},\Theta} k_h$ (and we refer to Appendix C.5 for some details about alternative possible choices of 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 3 and 4, 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 $(b_{\mu_{\boldsymbol{\lambda},\Theta},\alpha}(\theta_{j,t}))_{1 \leqslant J}$ at each iteration $n = 1 \ldots N$. We take $J = 100$, $M \in \{100, 1000, 2000\}$, $\alpha = 0.5$, $\kappa = 0$, $\eta_0 = 0.3$ and the initial particles $\theta_1, \ldots, \theta_J$ 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 Variational Renyi bound introduced in [10] (which requires next to none additional computations).

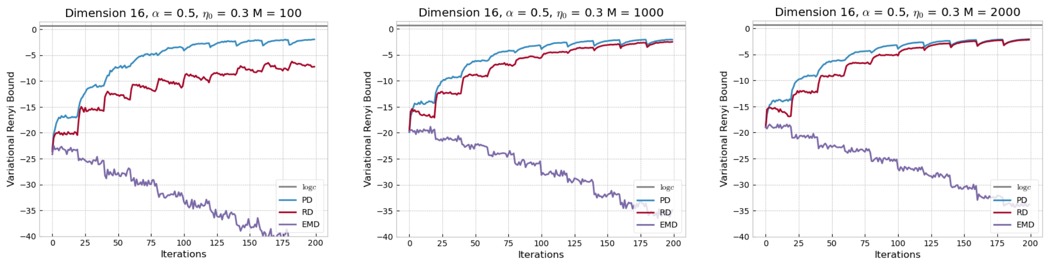
The results for the Power Descent and the Renyi Descent are displayed on Figure 1 below and we add the Entropic Mirror Descent applied to Ψ_α as a reference.

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

Figure 1: Plotted is the average Variational Renyi 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.



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 most suitable between biased and unbiased α-divergence minimisation is still an open issue in the literature, both theoretically and empirically [15, 16, 19]. 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 [11, 10] and low Signal-to-Noise ratio when $\alpha \neq 0$ [16] 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.

6 Conclusion

We investigated algorithms that can be used to perform mixture weights optimisation for α-divergence minimisation regardless of how the mixture parameters are obtained. 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. Since our contributions are mainly theoretical, we believe these will not result in any negative societal impacts.

References

[1] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[2] Matthew James. Beal. Variational algorithms for approximate bayesian inference. *PhD thesis*, 01 2003.

[3] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

[4] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

[5] Huaiyu Zhu and Richard Rohwer. Information geometric measurements of generalisation. Technical Report NCRG/4350, Aug 1995.

[6] Huaiyu Zhu and Richard Rohwer. Bayesian invariant measurements of generalization. *Neural Processing Letters*, 2:28–31, December 1995.

[7] Alfréd Rényi. 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*, pages 547–561, Berkeley, Calif., 1961. University of California Press.

[8] Tim van Erven and Peter Harremoes. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, Jul 2014.

[9] Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520, New York, New York, USA, 20–22 Jun 2016. PMLR.

[10] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1073–1081. Curran Associates, Inc., 2016.

[11] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via \chi upper bound minimization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2732–2741. Curran Associates, Inc., 2017.

[12] Volodymyr Kuleshov and Stefano Ermon. Neural variational inference and learning in undirected graphical models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] Robert Bamler, Cheng Zhang, Manfred Opper, and Stephan Mandt. Perturbative black box variational inference. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5079–5088. Curran Associates, Inc., 2017.

[14] Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5737–5747. Curran Associates, Inc., 2018.

[15] Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. In *3rd Symposium on Advances in Approximate Bayesian Inference*, pages 1–12, 2020.

[16] Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. *arXiv preprint arXiv:2010.09541*, 2020.

[17] Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *To appear in the Annals of Statistics*, 2021.

[18] Kamélia Daudel, Randal Douc, and François Roueff. Monotonic alpha-divergence minimisation. *arXiv preprint arxiv:2103.05684*, 2021.

[19] Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael Riis Andersen, Jonathan Huggins, and Aki Vehtari. Challenges and opportunities in high-dimensional variational inference. *arxiv preprint arxiv:2103.01085*, 2021.

[20] Arnaud Doucet, Nando Freitas, Kevin Murphy, and Stuart Russell. Sequential monte carlo methods in practice. 01 2013.

[21] Andrzej Cichocki and Shun-ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, Jun 2010.

[22] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167 – 175, 2003.

[23] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 01 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] End of Section 5 and Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Section 3 and Section 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Section 5 and supplementary.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Supplementary.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A

A.1 Equivalence between (1) and (2) with $p(y) = p(y, \mathscr{D})$

- Case $\alpha = 1$ with $f_1(u) = 1 - u + u\log(u)$ for all $u > 0$. Then,

$$\begin{aligned}
D_1(\mu K||\mathbb{P}) &= \int_{\mathsf{Y}} f_1\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} \mu k(y) \log\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right) \nu(\mathrm{d}y) + 0 \\
&= \int_{\mathsf{Y}} \mu k(y) \log\left(\frac{\mu k(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D}) \\
&= \int_{\mathsf{Y}} f_1\left(\frac{\mu k(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) + 1 - p(\mathscr{D}) + \log p(\mathscr{D})
\end{aligned}$$

  Thus,

$$\mathrm{arginf}_{\mu\in\mathsf{M}} D_1(\mu K||\mathbb{P}) = \mathrm{arginf}_{\mu\in\mathsf{M}} \Psi_1(\mu; p) \quad \text{with} \quad p(y) = p(y,\mathscr{D})$$

- Case $\alpha = 0$ with $f_0(u) = u - 1 - \log(u)$ for all $u > 0$.

$$\begin{aligned}
D_0(\mu K||\mathbb{P}) &= \int_{\mathsf{Y}} f_0\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} -\log\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} -\log\left(\frac{\mu k(y)}{p(y,\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) - \log p(\mathscr{D}) \\
&= \frac{1}{p(\mathscr{D})}\left[\int_{\mathsf{Y}} f_1\left(\frac{\mu k(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) + p(\mathscr{D}) - 1 - p(\mathscr{D})\log p(\mathscr{D})\right]
\end{aligned}$$

  Thus

$$\mathrm{arginf}_{\mu\in\mathsf{M}} D_0(\mu K||\mathbb{P}) = \mathrm{arginf}_{\mu\in\mathsf{M}} \Psi_0(\mu; p) \quad \text{with} \quad p(y) = p(y,\mathscr{D})$$

- 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 $u > 0$.

$$\begin{aligned}
&D_\alpha(\mu K||\mathbb{P}) \\
&= \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{\mu k(y)}{p(y|\mathscr{D})}\right)^\alpha - 1\right] p(y|\mathscr{D})\nu(\mathrm{d}y) \\
&= p(\mathscr{D})^{\alpha-1}\int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{\mu k(y)}{p(y,\mathscr{D})}\right)^\alpha - 1\right] p(y,\mathscr{D})\nu(\mathrm{d}y) + \frac{p(\mathscr{D})^\alpha - 1}{\alpha(\alpha-1)} \\
&= p(\mathscr{D})^{\alpha-1}\int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) + \frac{\alpha p(\mathscr{D})^{\alpha-1} + (1-\alpha)p(\mathscr{D})^\alpha - 1}{\alpha(\alpha-1)} \qquad (10)
\end{aligned}$$

  Thus,

$$\mathrm{arginf}_{\mu\in\mathsf{M}} D_\alpha(\mu K||\mathbb{P}) = \mathrm{arginf}_{\mu\in\mathsf{M}} \Psi_\alpha(\mu; p) \quad \text{with} \quad p(y) = p(y,\mathscr{D})$$

A.2 [17, Theorem 1] with $\Gamma(v) = [(\alpha-1)v + 1]^{\eta/(1-\alpha)}$

Theorem 4 ([17, Theorem 1] with $\Gamma(v) = [(\alpha-1)v + 1]^{\eta/(1-\alpha)}$)**.** *Assume that p and k are as in* (A1). *Let $\alpha \in \mathbb{R}\setminus\{1\}$, 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*

$$0 < \mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty \qquad (11)$$

holds and $\Psi_\alpha(\mu) < \infty$. Then, the two following assertions hold.

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

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

B

B.1 Proof that (A2) is satisfied in Example 1

Proof that (A2) *is satisfied in Example 1.*

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

$$\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} k(\theta, y) \times \sup_{\theta' \in \mathsf{T}} \left(\frac{k(\theta', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) < \infty \tag{12}$$

and

$$\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} \left|\log\left(\frac{k_h(\theta, y)}{p(y)}\right)\right| p(y)\nu(\mathrm{d}y) < \infty \; . \tag{13}$$

(i) We start by proving (12). 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 $y \in \mathsf{Y}$,

$$\frac{k_h(\theta, y)}{k_h(\theta', y)} \leqslant e^{\frac{\|y\|2r + r^2}{h^2}} \tag{14}$$

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 (14), 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 (13). For all $y \in \mathsf{Y}$ and all $\theta \in \mathsf{T}$, we have

$$\begin{aligned} e^{-\sup_{\theta \in \mathsf{T}} \frac{\|y-\theta\|^2}{2h^2}} &\leqslant (2\pi h^2)^{d/2} k_h(\theta, y) \leqslant 1 \\ e^{-\max_{i \in \{1,2\}} \frac{\|y-\theta_i^\star\|^2}{2}} &\leqslant c^{-1}(2\pi)^{d/2} p(y) \leqslant 1\end{aligned}$$

and we can deduce for all $y \in \mathsf{Y}$ and all $\theta \in \mathsf{T}$

$$\begin{aligned}\left|\log\left(\frac{k_h(\theta, y)}{p(y)}\right)\right| &\leqslant \sup_{\theta \in \mathsf{T}} \frac{\|y-\theta\|^2}{2h^2} + \max_{i \in \{1,2\}} \frac{\|y-\theta_i^\star\|^2}{2} + d|\log h| + |\log c| \\ &\leqslant \frac{(\|y\| + r)^2}{2}\left[\frac{1}{h^2} + 1\right] + d|\log h| + |\log c| \; .\end{aligned} \tag{15}$$

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 (13) holds.

□

B.2 Proof of Theorem 2

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 2. *Assume* (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*

$$\Psi_\alpha(\zeta_1) = \Psi_\alpha(\zeta_2) = \inf_{\zeta \in \mathsf{M}} \Psi_\alpha(\zeta).$$

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 $t \in (0,1)$, $f_\alpha(t u_y + (1-t) v_y) \leqslant t f_\alpha(u_y) + (1-t) f_\alpha(v_y)$ by convexity of f_α and we obtain

$$\Psi_\alpha(t\zeta_1 + (1-t)\zeta_2) \leqslant t \Psi_\alpha(\zeta_1) + (1-t)\Psi_\alpha(\zeta_2) = \inf_{\zeta \in \mathsf{M}} \Psi_\alpha(\zeta) \ . \tag{16}$$

Furthermore, $t\zeta_1 + (1-t)\zeta_2 \in \mathsf{M}$ which implies that we have equality in (16).

Consequently, for all $t \in (0,1)$:

$$\int_{\mathsf{Y}} \underbrace{[t f_\alpha(u_y) + (1-t) f_\alpha(v_y) - f_\alpha(t u_y + (1-t) v_y)]}_{\geqslant 0} p(y) \nu(\mathrm{d}y) = 0 \ .$$

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 $\zeta_1 \mathcal{R} \zeta$. □

Lemma 3. *Assume* (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,*

$$\Psi_\alpha(\mu^\star) = \inf_{\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})} \Psi_\alpha(\zeta) \ . \tag{17}$$

Furthermore, for all $\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})$, $\Psi_\alpha(\mu^\star) = \Psi_\alpha(\zeta)$ *implies that* $\mu^\star \mathcal{R} \zeta$.

Proof. Let $\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})$ be such that $\Psi_\alpha(\zeta) \leqslant \Psi_\alpha(\mu^\star)$. We have that

$$\zeta\left(b_{\mu^\star,\alpha} - \mu^\star(b_{\mu^\star,\alpha})\right) \leqslant \Psi_\alpha(\zeta) - \Psi_\alpha(\mu^\star) \leqslant 0 \ . \tag{18}$$

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 (18) yields $\Psi_\alpha(\zeta) = \Psi_\alpha(\mu^\star)$ and we recover (17).

Finally, assume there exists $\zeta \in \mathrm{M}_{1,\mu^\star}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star) = \Psi_\alpha(\zeta)$. Then, since $\mathrm{M}_{1,\mu^\star}(\mathsf{T})$ is a convex set, we have by Lemma 2 that $\mu^\star \mathcal{R} \zeta$. □

We now move on to the proof of Theorem 2.

Proof of Theorem 2. For convenience, we define the notation $\Psi_{\alpha,\Theta}(\boldsymbol{\lambda}) := \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$ 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 (A3), we have that $|\Psi_{\alpha,\Theta}(\boldsymbol{\lambda})| < \infty$ and that (11) is satisfied for all $\mu_{\boldsymbol{\lambda},\Theta}$ such that $\boldsymbol{\lambda} \in \mathcal{S}_J$. This means that the sequence $(\boldsymbol{\lambda}_n)_{n \in \mathbb{N}^\star}$ defined by (5) is well-defined, that the sequence $(\Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_n))_{n \in \mathbb{N}^\star}$ is lower-bounded and that $\Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_n)$ is finite for all $n \in \mathbb{N}^\star$. As $(\Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_n))_{n \in \mathbb{N}^\star}$ is nonincreasing by Theorem 4-(i), it converges in $\mathbb{R}$ and in particular we have

$$\lim_{n \to \infty} \Psi_{\alpha,\Theta} \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_n) - \Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_n) = 0 \ .$$

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 $\boldsymbol{\lambda} \mapsto \Psi_{\alpha,\Theta} \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}) - \Psi_{\alpha,\Theta}(\boldsymbol{\lambda})$ is continuous we obtain that $\Psi_{\alpha,\Theta} \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\bar{\boldsymbol{\lambda}}) = \Psi_{\alpha,\Theta}(\bar{\boldsymbol{\lambda}})$ and hence by Theorem 4-(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 $\mathrm{M}_{1,\mu_{\boldsymbol{\lambda},\Theta}}(\mathsf{T}) = \mathrm{M}_{1,\mu_{\boldsymbol{\lambda}',\Theta}}(\mathsf{T}) = \left\{\mu_{\boldsymbol{\lambda}'',\Theta} \;:\; \boldsymbol{\lambda}'' \in \tilde{\mathcal{S}}_{J,R}\right\}$, Lemma 3 implies that

$$\Psi_{\alpha,\Theta}(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\lambda}'' \in \tilde{\mathcal{S}}_{J,R}} \Psi_{\alpha,\Theta}\left(\boldsymbol{\lambda}''\right) = \Psi_{\alpha,\Theta}(\boldsymbol{\lambda}') \;.$$

Applying again Lemma 3, we get $\mu_{\boldsymbol{\lambda},\Theta} \mathcal{R} \mu_{\boldsymbol{\lambda}',\Theta}$, that is, $\mu_{\boldsymbol{\lambda},\Theta} K = \mu_{\boldsymbol{\lambda}',\Theta} K$. 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 2 applied to the convex subset of measures $\mathsf{M} = \mathcal{S}_J$, the function $\Psi_{\alpha,\Theta}$ 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, $\mu_{\boldsymbol{\lambda},\Theta} \mathcal{R} \mu_{\boldsymbol{\lambda}',\Theta}$ if and only if $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$. Then, by Theorem 4-(i) and by definition of $\boldsymbol{\lambda}_\star$

$$\Psi_{\alpha,\Theta} \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_\star) \leqslant \Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_\star) = \inf_{\boldsymbol{\lambda}' \in \mathcal{S}_J} \Psi_{\alpha,\Theta}(\boldsymbol{\lambda}') \leqslant \Psi_{\alpha,\Theta} \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_\star) \;,$$

and hence, $\Psi_{\alpha,\Theta} \circ \mathcal{I}_\alpha^{\mathrm{mixt}}(\boldsymbol{\lambda}_\star) = \Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_\star)$, showing that $\boldsymbol{\lambda}_\star \in F$ by Theorem 4-(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 $\ell \in \{1, \dots, L\}$,

$$\mathcal{I}_\alpha^{\mathrm{mixt}}(W_\ell) \cap \left(\bigcup_{j \neq \ell} W_j\right) = \emptyset \tag{19}$$

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 (19), 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 $(\mu_{\boldsymbol{\lambda}_n,\Theta})_{n \in \mathbb{N}^\star}$ weakly converges to $\mu_{\boldsymbol{\lambda}^\ell,\Theta}$ as $n \to \infty$ and Theorem 1 can be applied. Since $\boldsymbol{\lambda}_1 \in \mathcal{S}_J^+$, we have $\mathrm{M}_{1,\mu_{\boldsymbol{\lambda}_1,\Theta}}(\mathsf{T}) = \left\{\mu_{\boldsymbol{\lambda}',\Theta} \;:\; \boldsymbol{\lambda}' \in \mathcal{S}_J\right\}$ and Theorem 1-(iii) then shows that $\mu_{\boldsymbol{\lambda}^\ell,\Theta}$ is the global arginf of Ψ_α over all $\left\{\mu_{\boldsymbol{\lambda}',\Theta} \;:\; \boldsymbol{\lambda}' \in \mathcal{S}_J\right\}$. Therefore, $\ell = 1$, i.e., $\boldsymbol{\lambda}^\ell = \boldsymbol{\lambda}^1 = \boldsymbol{\lambda}_\star$ and

$$\Psi_{\alpha,\Theta}(\boldsymbol{\lambda}_\star) = \inf_{\boldsymbol{\lambda}' \in \mathcal{S}_J} \Psi_{\alpha,\Theta}(\boldsymbol{\lambda}') \;.$$

□

B.3 The Power Descent for mixture models: practical version

The algorithm below provides one possible approximated version of the Power Descent algorithm, where we have set $\Gamma(v) = [(\alpha - 1)v + 1]^{\frac{\eta}{1-\alpha}}$ with $\eta \in (0, 1]$.

Algorithm 3: *Practical version of the Power 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, $\Gamma(v) = [(\alpha-1)v+1]^{\frac{\eta}{1-\alpha}}$ with $\eta \in (0,1]$, 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 $\mu_{\boldsymbol{\lambda},\Theta} k$.

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

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

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

Iteration step : Set

$$\boldsymbol{\lambda} \leftarrow \frac{1}{w_{\boldsymbol{\lambda}}} \boldsymbol{W_\lambda}$$

C

C.1 Proof of Proposition 1

We first state (D1), which summarises the necessary convergence and differentiability assumptions needed in the proof of proposition 1.

(D1) (i) we have $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} k(\theta, y) \times \sup_{\theta' \in \mathsf{T}} \left(\frac{k(\theta', y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) < \infty$;

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

(iii) 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. 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 $(\alpha - 1)\kappa > 0$.

Proof of proposition 1. 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 $\theta \in \mathsf{T}$,

$$\begin{aligned}
\lim_{\alpha \to 1} [(\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1]^{\frac{\eta}{1-\alpha}} &= \exp\left(\lim_{\alpha \to 1} -\eta \frac{\log[(\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1]}{\alpha-1} \right) \\
&= \exp\left(\lim_{\alpha \to 1} -\eta \frac{\int_{\mathsf{Y}} k(\theta,y) \left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} \log\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y) + \kappa}{\int_{\mathsf{Y}} k(\theta,y) \left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) + (\alpha-1)\kappa} \right) \\
&= \exp\left[-\eta \int_{\mathsf{Y}} k(\theta, y) \log\left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y) \right] \exp(-\eta\kappa)
\end{aligned}$$

In addition, by the Dominated Convergence Theorem (and (D1)-(iii) when $\alpha > 1$), we have

$$\lim_{\alpha \to 1} \mu \left([(\alpha - 1)(b_{\mu,\alpha} + \kappa) + 1]^{\frac{\eta}{1-\alpha}} \right) = \mu \left(\exp \left[-\eta \int_{\mathsf{Y}} k(\cdot, y) \log \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y) \right] \right) \exp(-\eta\kappa) \ .$$

Thus,

$$\lim_{\alpha \to 1} [\mathcal{I}_\alpha(\mu)](h) = \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta) h(\theta) e^{-\eta \int_{\mathsf{Y}} k(\theta,y) \log\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)}}{\mu\left(e^{-\eta \int_{\mathsf{Y}} k(\cdot,y) \log\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y)}\right)} = [\mathcal{I}_1(\mu)](h) \ .$$

□

C.2 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

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

Lemma 4. *Assume* (A1). *The gradient of* $\Psi_\alpha^{AR}(\mu)$ *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,

$$\begin{aligned}
\Psi_\alpha^{AR}(\mu + \varepsilon\mu') &= \frac{1}{\alpha(\alpha-1)} \log \left(\int_{\mathsf{Y}} [(\mu + \varepsilon\mu')k(y)]^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right) \\
&= \frac{1}{\alpha(\alpha-1)} \log \left(\int_{\mathsf{Y}} \mu k(y)^\alpha \left[1 + \alpha\varepsilon \frac{\mu' k(y)}{\mu k(y)} \right] p(y)^{1-\alpha} \nu(\mathrm{d}y) + (\alpha - 1)\kappa + o(\varepsilon) \right)
\end{aligned}$$

where we used that $(1 + u)^\alpha = 1 + \alpha u + o(u)$ as $u \to 0$. Thus,

$$\begin{aligned}
\Psi_\alpha^{AR}(\mu + \varepsilon\mu') &= \Psi_\alpha^{AR}(\mu) + \frac{1}{\alpha(\alpha-1)} \log \left(1 + \alpha\varepsilon \frac{\int_{\mathsf{Y}} \mu' k(y) \left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) + (\alpha-1)\kappa} + o(\varepsilon) \right) \\
&= \Psi_\alpha^{AR}(\mu) + \varepsilon \frac{1}{\alpha - 1} \frac{\int_{\mathsf{Y}} \mu' k(y) \left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) + (\alpha-1)\kappa} + o(\varepsilon) \\
&= \Psi_\alpha^{AR}(\mu) + \varepsilon \int_{\mathsf{T}} \mu'(\mathrm{d}\theta) \frac{1}{\alpha-1} \frac{b_{\mu,\alpha}(\theta) + 1/(\alpha-1)}{\mu(b_{\mu,\alpha}) + \kappa + 1/(\alpha-1)} + o(\varepsilon)
\end{aligned}$$

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

$$\mu_{n+1}(\mathrm{d}\theta) = \mu_n(\mathrm{d}\theta) \frac{e^{-\frac{\eta}{\alpha-1} \frac{b_{\mu_n,\alpha}(\theta)}{\mu_n(b_{\mu_n,\alpha}) + \kappa + 1/(\alpha-1)}}}{\mu_n(e^{-\frac{\eta}{\alpha-1} \frac{b_{\mu_n,\alpha}}{\mu_n(b_{\mu_n,\alpha}) + \kappa + 1/(\alpha-1)}})} \ , \quad n \in \mathbb{N}^\star \ .$$

C.3 Proof of Theorem 3

As we shall see, the proof can be adapted from the proof of [17, 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

$$\begin{aligned}
L_{\alpha,2} &= \eta^{-1} \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} [(\alpha-1)(b_{\mu,\alpha}(\theta)+\kappa)+1] \\
L &= \eta^2 \sup_{v\in\mathrm{Dom}_\alpha^{AR}} e^{-\eta v} \\
L_{\alpha,3} &= \sup_{v\in\mathrm{Dom}_\alpha^{AR}} e^{\eta v} \\
L_{\alpha,1} &= \inf_{v\in\mathrm{Dom}_\alpha^{AR}} \{1-\eta(\alpha-1)(v-\kappa')\} \times \eta \inf_{v\in\mathrm{Dom}_\alpha^{AR}} e^{-\eta v} .
\end{aligned} \tag{20}$$

C.3.1 Recalling [17, Lemma 5]

Let (ζ, μ) be a couple of probability measures where ζ is dominated by μ which we denote by $\zeta \preceq \mu$ and define

$$A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1-g(\theta)] , \tag{21}$$

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 [17, Lemma 5] in Lemma 5 below.

Lemma 5. *[17, Lemma 5] Assume* (A1). *Then, for all $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ such that $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$, we have*

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta) . \tag{22}$$

Moreover, equality holds in (22) *if and only if $\zeta = \mu$.*

C.3.2 Adaptation of [17, Theorem 1]

Lemma 6. *Assume* (A1) *and* (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*

$$0 < \mu\left\{\exp\left(-\eta\frac{b_{\mu,\alpha}+1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right)\right\} < \infty \tag{23}$$

holds and $\Psi_\alpha(\mu) < \infty$. Then, the two following assertions hold.

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

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

Proof. The proof builds on the proof of [17, 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 (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

$$\begin{aligned}A_\alpha&=[(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1]\times\mathbb{E}[(V-\kappa')\tilde{\Gamma}^{\alpha-1}(V)(1-\tilde{\Gamma}(V))]\\&=[(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1]\times\mathbb{C}\mathrm{ov}((V-\kappa')\tilde{\Gamma}^{\alpha-1}(V),1-\tilde{\Gamma}(V))\ .\end{aligned}\tag{24}$$

Under (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 $(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1>0$. □

C.3.3 Adaptation of [17, Lemma 6]

Consider the probability space $(\mathsf{T},\mathcal{T},\mu)$ and denote by $\mathbb{V}\mathrm{ar}_\mu$ the associated variance operator.

Lemma 7. *Assume* (A1) *and* (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* (23) *holds and $\Psi_\alpha(\mu)<\infty$. Then,*

$$\frac{(\alpha-1)\kappa L_{\alpha,1}}{2}\mathbb{V}\mathrm{ar}_\mu\left(\frac{b_{\mu,\alpha}+1/(\alpha-1)}{(\alpha-1)(\mu(b_{\mu,\alpha})+\kappa)+1}\right)\leqslant\Psi_\alpha(\mu)-\Psi_\alpha\circ\mathcal{I}_\alpha^{AR}(\mu)\ ,\tag{25}$$

where

$$L_{\alpha,1}:=\inf_{v\in\mathrm{Dom}_\alpha^{AR}}\{1-\eta(\alpha-1)(v-\kappa')\}\times\inf_{v\in\mathrm{Dom}_\alpha^{AR}}\eta e^{-\eta v}\ .$$

Proof. The proof of Lemma 7 builds on the proof of [17, Lemma 6], which can be found in the supplementary material of [17]. Using (24) combined with the fact that under (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,

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

and we thus obtain (25). □

C.3.4 Adaptation of the proof of [17, Theorem 2] to obtain Theorem 3

Proof of Theorem 3. The proof of Theorem 3 builds on the proof of [17, Theorem 2], which can be found in the supplementary material of [17]. We prove the assertions successively.

(i) The proof of (i) simply consists in verifying that we can apply Lemma 6. For all $\mu \in \mathrm{M}_1(\mathsf{T})$, (23) 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 6 combined with $\Psi_\alpha(\mu_n) < \infty$ implies that $\Psi_\alpha(\mu_{n+1}) \leqslant \Psi_\alpha(\mu_n) < \infty$, we obtain by induction that $(\Psi_\alpha(\mu_n))_{n\in\mathbb{N}^\star}$ is non-increasing.

(ii) Let $n \in \mathbb{N}^\star$, set $\Delta_n = \Psi_\alpha(\mu_n) - \Psi_\alpha(\mu^\star)$ 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

$$\Delta_n \leqslant L_{\alpha,2}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\mathrm{d}\mu^\star + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}(V_n)L_{\alpha,3}\right] . \tag{26}$$

The convexity of f_α implies that

$$\Delta_n \leqslant \int_{\mathsf{T}} b_{\mu_n,\alpha}(\mathrm{d}\mu_n - \mathrm{d}\mu^\star) \tag{27}$$

$$= \int_{\mathsf{T}} \left(b_{\mu_n,\alpha} + \frac{1}{\alpha-1}\right)(\mathrm{d}\mu_n - \mathrm{d}\mu^\star)$$

$$= \frac{(\alpha-1)(\mu_n(b_{\mu_n,\alpha})+\kappa)+1}{\eta}\int_{\mathsf{T}}(\mu_n(\eta V_n) - \eta V_n)\mathrm{d}\mu^\star . \tag{28}$$

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

$$\Delta_n \leqslant L_{\alpha,2}\int_{\mathsf{T}}\left[\mu_n(\eta V_n) + \log\mu_n\left(e^{-\eta V_n}\right) + \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right)\right]\mathrm{d}\mu^\star . \tag{29}$$

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 (29) implies (26). To conclude, we apply Lemma 7 to $g = \frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}$ and combining with (26), we obtain

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

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

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

and we recover (9) using (i), that $KL(\mu^\star||\mu_{N+1}) \geqslant 0$ and that $\Delta_{N+1} \geqslant 0$.

□

C.4 The Renyi Descent for mixture models: practical version

The algorithm below provides one possible approximated version of the Renyi Descent algorithm, where we have set $\Gamma(v) = e^{-\eta v}$ with $\eta > 0$.

Algorithm 4: *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 $\eta > 0$, 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 $\mu_{\boldsymbol{\lambda},\Theta} k$.

Expectation step : Compute $\boldsymbol{B}_{\boldsymbol{\lambda}} = (b'_j)_{1 \leqslant j \leqslant J}$ where for all $j = 1 \dots J$

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

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

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

C.5 Alternative Exploration step in Algorithm 2

We present here several possible alternative choices of Exploration step in Algorithm 2, beyond the one we have made in Section 5 and that is based on [18]. Our goal here is not to discriminate between all of them, but to illustrate the generality of our approach.

Gradient Descent. One could use a Gradient Descent approach to optimise the mixture components parameters $\{\theta_{1,t+1}, \dots, \theta_{J,t+1}\}$ in the spirit of Renyi's α-divergence gradient-based methods (e.g [9, 10]) or α-divergence gradient-based methods (e.g [11, 12]).

The particular case $\alpha \in [0, 1)$. Following [18], if we consider the specific case $\alpha \in [0, 1)$ another possibility would be to set at time t: for all $j = 1 \dots J$

$$\theta_{j,t+1} = \mathrm{argmax}_{\theta_j \in \mathsf{T}} \int_{\mathsf{Y}} \gamma^t_{j,\alpha}(y) \log(k(\theta_j, y)) \nu(\mathrm{d}y) \tag{30}$$

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

$$\gamma^t_{j,\alpha}(y) = k(\theta_{j,t}, y) \left(\frac{\mu_{\boldsymbol{\lambda},\Theta} k(y)}{p(y)} \right)^{\alpha - 1} .$$

Indeed, [18] showed that the above update formulas for $\{\theta_{1,t+1}, \dots, \theta_{J,t+1}\}$ ensure a systematic decrease in the α-divergence and they notably explained how these update formulas could even outperform typical Renyi's α / α-divergence gradient-based approaches (we refer to [18] for details).

Furthermore, in the particular case of d-dimensional Gaussian kernels with $k(\theta_{j,t}, y) = \mathcal{N}(y; m_{j,t}, \Sigma_{j,t})$ and where $\theta_{j,t} = (m_{j,t}, \Sigma_{j,t}) \in \mathsf{T}$ denotes the mean and covariance matrix of the j-th Gaussian component density, they obtained that the maximisation procedure (30) amounts to setting

$$\forall j = 1 \dots J, \quad m_{j,t+1} = \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^t(y) y \, \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^t(y) \nu(\mathrm{d}y)}$$

$$\Sigma_{j,t+1} = \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^t(y)(y - m_{j,t})(y - m_{j,t})^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^t(y) \nu(\mathrm{d}y)} .$$

These update formulas can then always be made feasible by resorting to Monte Carlo approximations and can be used as a valid Exploration step. If we were to focus on solely updating the means $(m_{j,t+1})_{1 \leqslant j \leqslant J}$, we could for example consider the Exploration step given by:

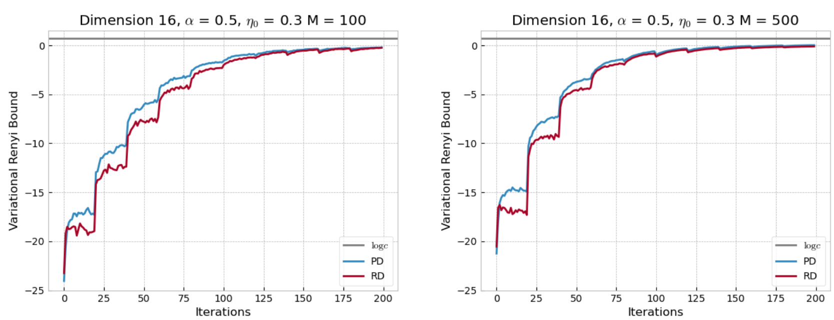
$$\forall j = 1 \dots J, \quad \theta_{j,t+1} = m_{j,t+1} = \frac{\sum_{m=1}^M \hat{\gamma}_j^{(t)}(Y_m'; \boldsymbol{\lambda}) \cdot Y_m'}{\sum_{m=1}^M \hat{\gamma}_j^{(t)}(Y_m'; \boldsymbol{\lambda})}$$

where the M samples $(Y_m')_{1 \leqslant m \leqslant M}$ have been drawn independently from the proposal $\mu_{\boldsymbol{\lambda},\Theta}$ and where we have set

$$\hat{\gamma}_j^{(t)}(y; \boldsymbol{\lambda}) = \frac{k(\theta_{j,t}, y)}{\mu_{\boldsymbol{\lambda},\Theta} k(y)} \left(\frac{\mu_{\boldsymbol{\lambda},\Theta} k(y)}{p(y)} \right)^{\alpha - 1} .$$

We ran Algorithm 2 over 100 replicates for this choice of Exploration step with $M \in \{100, 500\}$ (and keeping the same target p, initial sampler q_0, and hyperparameters $N = 20$, $T = 10$, $\eta = \eta_0 / \sqrt{N}$ with $\eta_0 = 0.3$, $\alpha = 0.5$, $J = 100$, $\kappa = 0.$ and $d = 16$ as those chosen in Section 5). The results when using the Power and the Renyi Descent as Exploitation steps can be visualised in the figure below.

Figure 2: Plotted is the average Variational Renyi bound for the Power Descent (PD) and the Renyi Descent (RD) 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 a similar behavior for the Power and the Renyi Descent, which illustrates the closeness between both algorithms, irrespective of the choice of the Exploration step.