Monotonic alpha-divergence minimisation

Kamélia Daudel, Randal Douc, François Roueff

MONOTONIC ALPHA-DIVERGENCE MINIMISATION

Kamélia Daudel
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
kamelia.daudel@gmail.com

Randal Douc
SAMOVAR, Télécom SudParis
Institut Polytechnique de Paris, France
randal.douc@telecom-sudparis.eu

François Roueff
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
francois.roueff@telecom-paris.fr

ABSTRACT

In this paper, we introduce a novel iterative algorithm which carries 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. Notably, our approach permits to build on various methods previously proposed for α-divergence minimisation such as gradient or power descent schemes. Furthermore, we shed a new light on an integrated Expectation Maximization algorithm. We 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.

1 Introduction

Bayesian inference tasks often induce intractable and hard-to-compute posterior densities which need to be approximated. Among the class of approximating methods, Variational inference methods (e.g Variational Bayes [1, 2]) have attracted a lot of attention as they have empirically been shown to be widely applicable to many high-dimensional machine-learning problems ([3, 4, 5]).

These optimisation-based methods introduce a simpler density family $\mathcal{Q}$ and find the best approximation to the unknown posterior density among this family in terms of a certain divergence, the most common choice of divergence being the forward Kullback-Leibler divergence ([6, 7]).

However, the forward Kullback-Leibler is known to have some drawbacks: its zero-forcing behavior typically results in light tails and covariance underestimation ([8]), which could be especially inconvenient for multimodal posterior densities in high-dimensional settings when the approximating family $\mathcal{Q}$ does not exactly match the posterior density ([9, 10]).

To avoid this hurdle, advances in Variational Inference seeked to employ more general classes of divergences such as the α-divergence ([11, 12]) and Renyi's α-divergence ([13, 14]), which have been used in [15, 16, 17, 18] and [19, 20]. Indeed, thanks to the hyperparameter α these families of divergences interpolate between the forward ($\alpha \to 1$) and the reverse ($\alpha \to 0$) Kullback-Leibler, that is, they provide a more flexible framework between the zero-forcing property of the case $\alpha \to 1$ and the mass-covering behavior of the case $\alpha \to 0$ ([8]).

In the spirit of α-divergence-based methods, we propose in this paper to build a framework for α-divergence minimisation. The particularity of our work will be that is it amenable to mixture models optimisation and that it ensures a monotonic decrease in the α-divergence at each step. The paper is then organised as follows:

- In Section 2, we introduce some notation and we state the optimisation problem we aim at solving in terms of the targeted density, the approximating density $q \in \mathcal{Q}$ and the α-divergence.

- In Section 3, we consider the typical Variational Inference case where q belongs to a parametric family. In this particular case, we state in Theorem 1 conditions which ensure a systematic decrease in the α-divergence at each step for all $\alpha \in [0, 1)$. We then show in Corollary 2 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 3 additional iterative schemes satisfying the conditions of Theorem 1, 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, we further extend the results from Section 3 to the more general case of mixture models. We derive in Theorem 2 and 3 conditions to simultaneously optimise both the weights and the component parameters of a given mixture model, all the while maintaining the systematic decrease in the α-divergence initially enjoyed in Theorem 1. These conditions are then met in Corollary 5 and 6, 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 [20] and provide in Proposition 7 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 [21] as a special case.

- Lastly, we show in Section 5 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.

2 Notation and optimisation problem

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. 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 or too costly to compute.

We denote by $\mathbb{P}$ the probability measure on $(\mathsf{Y}, \mathcal{Y})$ with corresponding density $p(\cdot|\mathscr{D})$ with respect to ν. As for the approximating family, we denote by $\mathbb{Q}$ the probability measure on $(\mathsf{Y}, \mathcal{Y})$ with associated density $q \in \mathcal{Q}$ with respect to ν.

We now specify the optimisation problem we consider in this paper in terms of the α-divergence. We let f_α be the convex function on $(0, +\infty)$ defined by $f_0(u) = -\log(u)$, $f_1(u) = u \log u$ and $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)} \left[u^\alpha - 1\right]$ for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Then, the α-divergence between $\mathbb{Q}$ and $\mathbb{P}$ (extended by continuity to the cases $\alpha = 0$ and $\alpha = 1$ as for example done in [22]) is given by

$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} f_\alpha \left(\frac{q(y)}{p(y|\mathscr{D})} \right) p(y|\mathscr{D})\nu(\mathrm{d}y) \ , \tag{1}$$

and the Variational Inference optimisation problem we aim at solving is

$$\inf_{q \in \mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}) \ .$$

Notably, it can easily be proven that the optimisation problem above is equivalent to solving

$$\inf_{q \in \mathcal{Q}} \Psi_\alpha(q; p) \quad \text{with} \quad p(y) = p(y, \mathscr{D}) \quad \text{for all } y \in \mathsf{Y} \ , \tag{2}$$

where, for all measurable positive function p on $(\mathsf{Y}, \mathcal{Y})$ and for all $q \in \mathcal{Q}$, we have set

$$\Psi_\alpha(q; p) = \int_{\mathsf{Y}} f_\alpha \left(\frac{q(y)}{p(y)} \right) p(y) \nu(\mathrm{d}y) \ . \tag{3}$$

As the normalisation constant does not appear anymore in the optimisation problem (2), this formulation is often preferred in practice. Therefore, we consider the general optimisation problem

$$\inf_{q \in \mathcal{Q}} \Psi_\alpha(q; p) \ , \tag{4}$$

where p is any measurable positive function on $(\mathsf{Y}, \mathcal{Y})$. Note that we may drop the dependency on p in Ψ_α for notational ease and when no ambiguity occurs.

At this stage, we are left with the choice of the approximating family $\mathcal{Q}$ appearing in the optimisation problem (4). The natural idea in Variational Inference and the starting point of our approach is then to work within a parametric family : 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 $\mathsf{T} \times \mathsf{Y}$, we consider a parametric family of the form

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

3 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 paper.

(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 the sequence $(\Psi_\alpha(k(\theta_n, \cdot))_{n \geqslant 1}$ is decreasing. The core idea of our approach will rely on the following proposition.

Proposition 1. *Assume (A1). For all $\alpha \in [0, 1)$ and all $\theta, \theta' \in \mathsf{T}$, it holds that*

$$\Psi_\alpha(k(\theta, \cdot)) \leqslant \int_{\mathsf{Y}} \frac{k(\theta', y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta', y)} \right) \nu(\mathrm{d}y) + \Psi_\alpha(k(\theta', \cdot)) \ . \tag{5}$$

Proof. We treat the two cases $\alpha = 0$ and $\alpha \in (0, 1)$ separately.

(a) Case $\alpha = 0$, with $f_0(u) = -\log(u)$ 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 $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1\right]$ for all $u > 0$. We have that

$$\begin{aligned}\Psi_\alpha(k(\theta,\cdot)) &= \int_{\mathsf{Y}} \frac{\left[\left(\frac{k(\theta,y)}{p(y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) \\ &= \int_{\mathsf{Y}} \left(\frac{k(\theta',y)}{p(y)}\right)^\alpha \frac{\left[\left(\frac{k(\theta,y)}{k(\theta',y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) + \Psi_\alpha(k(\theta',\cdot))\end{aligned}$$

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

$$\frac{1}{\alpha-1}\log(u) = \frac{1}{\alpha(\alpha-1)}\log(u^\alpha) \geqslant f_\alpha(u)\ .$$

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 (5).

□

This result then allows us to deduce Theorem 1 below.

Theorem 1. *Assume (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 $n \geqslant 1$,*

$$\int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \log\left(\frac{k(\theta_{n+1},y)}{k(\theta_n,y)}\right)\nu(\mathrm{d}y) \leqslant 0\ . \tag{6}$$

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 results follows by setting $\theta = \theta_{n+1}$ and $\theta' = \theta_n$ in (5) combined with (6). □

At this point, we seek to find iterative schemes satisfying (6). This leads us to our first corollary.

Corollary 2. *Assume (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*

$$\theta_{n+1} = \operatorname*{argmax}_{\theta\in\mathsf{T}} \int_{\mathsf{Y}} k(\theta_n,y)^\alpha p(y)^{1-\alpha}\log(k(\theta,y))\nu(\mathrm{d}y)\ , \quad n \geqslant 1\ . \tag{7}$$

Then (6) holds and we can apply Theorem 1.

Proof. We have that (6) holds by definition of θ_{n+1} combined with the fact that $\alpha \in [0,1)$ and we can thus apply Theorem 1. □

Let us comment on Corollary 2. A remarkable aspect is that (7) is written as a maximisation problem involving the logarithm of the kernel k. This means that we can use (7) 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 1 (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 (7) yields the following update formulas:*

$$\begin{aligned}\forall n \geqslant 1\ , \quad m_{n+1} &= \frac{\int_{\mathsf{Y}} k(\theta_n,y)^\alpha p(y)^{1-\alpha} y\ \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n,y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)} \\ \Sigma_{n+1} &= \frac{\int_{\mathsf{Y}} k(\theta_n,y)^\alpha p(y)^{1-\alpha}(y-m_n)(y-m_n)^T\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n,y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)}\ .\end{aligned}$$

Example 2 (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 (7) yields the following update formulas:*

$$\forall n \geqslant 1 \,, \quad m_{n+1} = \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) y \, \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) \nu(\mathrm{d}y)}$$
$$\Sigma_{n+1} = \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) (y - m_n)(y - m_n)^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} g^n(y) \nu(\mathrm{d}y)} \,,$$

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 3 (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 (7) yields the following update formulas: for all $n \geqslant 1$,*

$$\theta_{n+1}^{(\ell)} = \operatorname*{argmax}_{\theta^{(\ell)}} \int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \log(k^{(\ell)}(\theta^{(\ell)}, y^{(\ell)})) \nu(\mathrm{d}y) \,, \quad 1 \leqslant \ell \leqslant L \,.$$

Interestingly, while Corollary 2 has a convenient formulation and corresponds to the intuitive choice so that (6) holds, it is also possible to derive alternative schemes satisfying (6) under additional smoothness conditions (see Appendix A.1 for the definition of β-smoothness), as written in Corollary 3.

Corollary 3. *Assume (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*

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \nabla g_n(\theta_n) \,, \quad n \geqslant 1 \,, \tag{8}$$

where $(g_n)_{n \geqslant 1}$ is the sequence of functions defined by: for all $n \geqslant 1$ and all $\theta \in \mathsf{T}$

$$g_n(\theta) = c_n \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \log\left(\frac{k(\theta, y)}{k(\theta_n, y)}\right) \nu(\mathrm{d}y) \,, \tag{9}$$

and g_n is assumed to be β_n-smooth. Then (6) holds and we can apply Theorem 1.

Proof. Since $\gamma_n \in (0, 1]$ and g_n is a β_n-smooth function by assumption, we can apply Lemma 12 and we obtain that for all $n \geqslant 1$,

$$g_n(\theta_n) - g_n\left(\theta_n - \frac{\gamma_n}{\beta_n} \nabla g_n(\theta_n)\right) \geqslant \frac{\gamma_n}{2\beta_n} \|\nabla g_n(\theta_n)\|^2 \,.$$

Thus, by definition of θ_{n+1} in (8), we have

$$0 = g_n(\theta_n) \geqslant g_n(\theta_{n+1}) \,,$$

which in turn implies (6) and the proof is concluded. □

Let us now reflect on the implications of Corollary 3. 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, (8) becomes respectively

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \nabla(\log k(\theta_n, y)) \nu(\mathrm{d}y) \,, \quad n \geqslant 1 \tag{10}$$

$$\theta_{n+1} = \theta_n - \frac{\gamma_n}{\beta_n} \left(\frac{1}{\alpha - 1} \frac{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nabla(\log k(\theta_n, y)) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} k(\theta_n, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)}\right) \,, \quad n \geqslant 1 \,. \tag{11}$$

Here, letting $p(y) = p(y, \mathscr{D})$ for all $y \in \mathsf{Y}$, the iterative schemes (10) and (11) can both be seen as usual gradient descent iterations used for α-divergence

$$D_\alpha(K(\theta,\cdot)||\mathbb{P}) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{k(\theta,y)}{p(y|\mathscr{D})} \right)^\alpha - 1 \right] p(y|\mathscr{D}) \nu(\mathrm{d}y)$$

and Renyi's α-divergence

$$D_\alpha^{(\mathrm{AR})}(K(\theta,\cdot)||\mathbb{P}) = \frac{1}{\alpha(\alpha-1)} \log \left(\int_{\mathsf{Y}} k(\theta,y)^\alpha p(y|\mathscr{D})^{1-\alpha} \nu(\mathrm{d}y) \right)$$

minimisation with a learning policy proportional to $(\gamma_n \beta_n^{-1})_{n \geqslant 1}$. Notice that Renyi's α-divergence is defined following the convention from [22], alternative definitions may use a different scaling factor.

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 3 are satisfied.

Example 4. *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 (9) 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 3 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 1. 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 Extension to mixture models

In order to generalise the approach of Section 3 to mixture models, let us first define the class of mixture models we are going to be working with. 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 $\mathcal{S}_J^+ = \{\boldsymbol{\lambda} \in \mathcal{S}_J \ : \ \forall j \in \{1, \ldots, J\}, \ \lambda_j > 0\}$. The Dirac measure on $(\mathsf{T}, \mathcal{T})$ is denoted by δ_θ where $\theta \in \mathsf{T}$. Now using 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 $\boldsymbol{\lambda} \in \mathcal{S}_J$ and for all $\Theta \in \mathsf{T}^J$, we are interested in the mixture model approximating family given by

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

that is, we consider the optimisation problem

$$\inf_{\boldsymbol{\lambda}\in\mathcal{S}_J,\Theta\in\mathsf{T}^J} \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k;p) \ ,$$

where p is any measurable positive function on $(\mathsf{Y},\mathcal{Y})$. Notice in particular that the framework from Section 3 corresponds to having taken $J=1$ in the optimisation problem above. Let us next 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 $n\geqslant 1$. For convenience, we also introduce the shorthand notation $\mu_n = \sum_{j=1}^J \lambda_{j,n}\delta_{\theta_{j,n}}$ and

$$\gamma_{j,\alpha}^n(y) = k(\theta_{j,n},y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1} \tag{12}$$

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 3 to the case of mixture models is to generalise Proposition 1, which brings us to Proposition 4 below.

Proposition 4. *Assume (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*

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant \int_{\mathsf{Y}}\sum_{j=1}^J \frac{\lambda_j' k(\theta_j',y)}{\alpha-1}\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^{\alpha-1}\log\left(\frac{\lambda_j}{\lambda_j'}\frac{k(\theta_j,y)}{k(\theta_j',y)}\right)\nu(\mathrm{d}y) + \Psi_\alpha(\mu_{\boldsymbol{\lambda}',\Theta'}k) \ . \tag{13}$$

Furthermore, equality holds in (13) 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 f_α, Jensen's inequality implies

$$\begin{aligned}
\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) &= \int_{\mathsf{Y}} f_\alpha\left(\frac{\sum_{j=1}^J \lambda_j k(\theta_j,y)}{p(y)}\right)p(y)\nu(\mathrm{d}y)\\
&\leqslant \int_{\mathsf{Y}}\sum_{j=1}^J \frac{\lambda_j' k(\theta_j',y)}{\sum_{\ell=1}^J \lambda_\ell' k(\theta_\ell',y)} f_\alpha\left(\frac{\lambda_j k(\theta_j,y)}{p(y)\frac{\lambda_j' k(\theta_j',y)}{\sum_{\ell=1}^J \lambda_\ell' k(\theta_\ell',y)}}\right)p(y)\nu(\mathrm{d}y)\\
&= \int_{\mathsf{Y}}\sum_{j=1}^J \frac{\lambda_j' k(\theta_j',y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} f_\alpha\left(\frac{\lambda_j k(\theta_j,y)}{\lambda_j' k(\theta_j',y)}\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)p(y)\nu(\mathrm{d}y) \ .
\end{aligned} \tag{14}$$

We now treat the two cases $\alpha=0$ and $\alpha\in(0,1)$ separately.

(a) Case $\alpha=0$, with $f_0(u)=-\log(u)$ for all $u>0$. In this case, (14) yields

$$\begin{aligned}
\Psi_0(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant &\int_{\mathsf{Y}}\sum_{j=1}^J \lambda_j' \times \frac{-k(\theta_j',y)p(y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}\log\left(\frac{\lambda_j}{\lambda_j'}\frac{k(\theta_j,y)}{k(\theta_j',y)}\right)\nu(\mathrm{d}y)\\
&+\int_{\mathsf{Y}}\sum_{j=1}^J \frac{\lambda_j' k(\theta_j',y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}\times\left[-\log\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)\right]p(y)\nu(\mathrm{d}y)
\end{aligned}$$

which is exactly (13) 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 $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1\right]$ for all $u > 0$. In this setting, (14) gives

$$\begin{aligned}
\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k) \leqslant & \int_{\mathsf{Y}} \sum_{j=1}^{J} \frac{\lambda'_j k(\theta'_j, y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \frac{\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^\alpha \left[\left(\frac{\lambda_j}{\lambda'_j}\frac{k(\theta_j,y)}{k(\theta'_j,y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) \\
& + \int_{\mathsf{Y}} \sum_{j=1}^{J} \frac{\lambda'_j k(\theta'_j, y)}{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)} \frac{\left[\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^\alpha - 1\right]}{\alpha(\alpha-1)} p(y)\nu(\mathrm{d}y) \\
= & \int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda'_j k(\theta'_j, y) \left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right)^{\alpha-1} f_\alpha\left(\frac{\lambda_j}{\lambda'_j}\frac{k(\theta_j,y)}{k(\theta'_j,y)}\right) \nu(\mathrm{d}y) \\
& + \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu_{\boldsymbol{\lambda}',\Theta'}k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \,, \qquad (15)
\end{aligned}$$

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 4 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

$$\frac{1}{\alpha-1}\log(u) = \frac{1}{\alpha(\alpha-1)}\log(u^\alpha) \geqslant f_\alpha(u) \,.$$

Thus, combining with (15) 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 (13) for $\alpha \in (0,1)$.

As for the case of equality, equality in (13) implies equality in (14) which in turn by strict convexity of f_α implies the desired result and concludes the proof of Proposition 4. □

We can then state our second main theorem.

Theorem 2. *Assume (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 $n \geqslant 1$,*

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1} \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \leqslant 0 \qquad (16)$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1} \log\left(\frac{k(\theta_{j,n+1},y)}{k(\theta_{j,n},y)}\right) \nu(\mathrm{d}y) \leqslant 0 \,. \qquad (17)$$

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 (13) combined with (16) and (17). □

We now plan on finding iterative schemes which satisfy (16) and (17). Strikingly, (16) does not depend on Θ_{n+1} nor does (17) depend on $\boldsymbol{\lambda}_{n+1}$. This means that we can treat these two inequalities separately and thus that the weights and component parameters of the mixture can be optimised simultaneously.

Observe also that the dependency in $\lambda_{j,n+1}$ appearing in (16) is simpler than the dependency in $\theta_{j,n+1}$ appearing in (17) and that is expressed through the kernel k . For this reason, we will first study (16). As we shall see, while the natural idea is to perform direct optimisation of the left-hand side of (16), a more general expression for the mixture weights can be derived, which will lead to numerical advantages later illustrated in Section 5.

4.1 Choice of $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$

In the following theorem, we identify an update formula which satisfies (16), regardless of the choice of the kernel k.

Theorem 3. *Assume (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 $n \geqslant 1$*

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \gamma_{\ell,\alpha}^n(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}\,, \quad j = 1\dots J \tag{18}$$

and (17) is satisfied. Then (16) 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 (17) is assumed, it remains to show (16) so that we can apply Theorem 2, before characterising the case of equality. To prove (16), 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 (16) 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

$$\sum_{j=1}^J \lambda_{j,n}\left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right]\log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \leqslant 0\,. \tag{19}$$

Notice then that by definition of $(\lambda_{j,n+1})_{1\leqslant j\leqslant J}$ when $\eta_n = 1$, we can write

$$\boldsymbol{\lambda}_{n+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathcal{S}_J^+} \sum_{j=1}^J \lambda_{j,n}\left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right]\log\left(\frac{\lambda_j}{\lambda_{j,n}}\right)\,.$$

[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 (19) 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 $\epsilon > 0$,

$$\sum_{j=1}^J \lambda_{j,n} \left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right] \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \\ \leqslant \frac{1}{\epsilon}\sum_{j=1}^J \lambda_{j,n}\left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right]\left[1 - \left(\frac{\lambda_{j,n}}{\lambda_{j,n+1}}\right)^{\epsilon}\right] . \quad (20)$$

Now notice that by definition of $(\lambda_{j,n+1})_{1\leqslant j\leqslant J}$ we can write

$$\boldsymbol{\lambda}_{n+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathcal{S}_J^+} \frac{1}{\epsilon}\sum_{j=1}^J \lambda_{j,n}\left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right]\left[1 - \left(\frac{\lambda_{j,n}}{\lambda_j}\right)^{\epsilon}\right]$$

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 $\sum_{j=1}^J \left(\bar{\beta}_j/\lambda_j\right)^{1+\epsilon}\lambda_j \geqslant (\sum_{j=1}^J \bar{\beta}_j)^{1+\epsilon}$ 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 (16) [since combined with (20) it implies (19) which itself implies (16) as seen in the case $\eta_n = 1$]. This settles the case $\eta_n \in (0,1)$.

We can thus apply Theorem 2 and we obtain (i). As for the case of equality, Theorem 2 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 (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 (19) 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 (20) 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 3, the mixture weights update given by (18) 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 $\boldsymbol{\lambda} \in \mathcal{S}_J^+$,

$$h_n(\boldsymbol{\lambda}) = \begin{cases} \sum_{j=1}^J \lambda_{j,n}\left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right]\log\left(\frac{\lambda_j}{\lambda_{j,n}}\right), & \text{if } \eta_n = 1 , \\ \frac{1}{\epsilon}\sum_{j=1}^J \lambda_{j,n}\left[\int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1}\nu(\mathrm{d}y) + \kappa\right]\left[1 - \left(\frac{\lambda_{j,n}}{\lambda_j}\right)^{\epsilon}\right], & \text{if } \eta_n \in (0,1) . \end{cases}$$

More specifically, $h_n(\boldsymbol{\lambda})$ acts as an upper bound of the left-hand side of (18) and we recover exactly the left-hand side of (18) in the particular case $\eta_n = 1$ and $\kappa = 0$.

Now that we have established Theorem 3, we are interested in deriving update formulas for the sequence $(\Theta_n)_{n\geqslant 1}$ satisfying (17).

4.2 Choice of $(\Theta_n)_{n\geqslant 1}$

We investigate three different approaches for choosing $(\Theta_n)_{n\geqslant 1}$.

4.2.1 A minimisation approach

The first idea is to consider the update for $(\Theta_n)_{n\geqslant 1}$ given by: for all $n \geqslant 1$,

$$\Theta_{n+1} = \underset{\Theta\in\mathsf{T}^J}{\operatorname{argmin}}\, g_n(\Theta)$$

where for all $\Theta \in \mathcal{S}_J^+ \times \mathsf{T}^J$, $g_n(\Theta) = \int_{\mathsf{Y}} \sum_{j=1}^J \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha-1} \log\left(\frac{k(\theta_j,y)}{k(\theta_{j,n},y)}\right) \nu(\mathrm{d}y)$. In this case, the full update $(\boldsymbol{\lambda}_{n+1}, \Theta_{n+1})$ can be written as the following optimisation problem

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

and we obtain Corollary 5.

Corollary 5. *Assume (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 (18) and*

$$\theta_{j,n+1} = \underset{\theta_j\in\mathsf{T}}{\operatorname{argmax}} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \log(k(\theta_j, y))\nu(\mathrm{d}y)\ , \quad j = 1\dots J\ . \tag{21}$$

Then (17) holds and we can apply Theorem 3.

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 (17) holds and we can apply Theorem 3. □

Consequently, under the assumptions of Corollary 5 we can define Algorithm 1, which leads to a systematic decrease in Ψ_α at each step and effectively generalises the monotonicity property from Corollary 2 to the case of mixture models. In line with Corollary 3, we next present another possible update formula for $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$.

Algorithm 1: α-divergence minimisation for Mixture Models based on (21)

At iteration n**,**
For all $j = 1\dots J$, set

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

$$\theta_{j,n+1} = \underset{\theta_j\in\mathsf{T}}{\operatorname{argmax}} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \log(k(\theta_j, y))\nu(\mathrm{d}y)\ .$$

4.2.2 A Gradient Descent approach

We shall now resort to gradient descent steps to satisfy (16).

Corollary 6. *Assume (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* (18) *and*

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta_{j,n}) \ , \quad j = 1 \dots J \ , \tag{22}$$

where for all $j = 1 \dots J$*,* $(g_{j,n})_{n\geqslant 1}$ *is defined by: for all* $n \geqslant 1$ *and all* $\theta \in \mathsf{T}$*,*

$$g_{j,n}(\theta) = c_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \ . \tag{23}$$

and $g_{j,n}$ *is assumed to be* $\beta_{j,n}$*-smooth. Then* (17) *holds and we can apply Theorem 3.*

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 12 and we obtain that for all $n \geqslant 1$ and all $j = 1 \dots J$,

$$g_{j,n}(\theta_{j,n}) - g_{j,n}\left(\theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta_{j,n})\right) \geqslant \frac{\gamma_{j,n}}{2\beta_{j,n}} \|\nabla g_{j,n}(\theta_{j,n})\|^2 \ .$$

Thus, by definition of $\theta_{j,n+1}$ in (22), we have

$$0 = g_{j,n}(\theta_{j,n}) \geqslant g_{j,n}(\theta_{j,n+1}).$$

which in turn implies (17) so that we can apply Theorem 3. □

This gives us the monotonicity property for Algorithm 2 by Corollary 6 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 $\theta \in \mathsf{T}$

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

Algorithm 2: α-divergence minimisation for Mixture Models based on (23)

At iteration n**,**
For all $j = 1 \dots J$, set

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

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta_n) \ .$$

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:

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_n}{\beta_n} \lambda_{j,n} \int_{\mathsf{Y}} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \nabla \log \left(k(\theta_{j,n}, y) \right) \nu(\mathrm{d}y) \ , \quad j = 1 \dots J \ , \tag{24}$$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_n}{\beta_n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nabla \log \left(k(\theta_{j,n}, y) \right) \nu(\mathrm{d}y)}{(\alpha - 1) \int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \ , \quad j = 1 \dots J \ . \tag{25}$$

Letting $p(y) = p(y, \mathscr{D})$ for all $y \in \mathsf{Y}$, we recognise usual gradient descent steps on Θ for α-divergence

$$D_\alpha(\mu_{\boldsymbol{\lambda},\Theta} K || \mathbb{P}) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{\mu_{\boldsymbol{\lambda},\Theta} k(y)}{p(y|\mathscr{D})} \right)^\alpha - 1 \right] p(y|\mathscr{D}) \nu(\mathrm{d}y)$$

and Renyi's α-divergence

$$D_\alpha^{(\mathrm{AR})}(\mu_{\boldsymbol{\lambda},\Theta} K || \mathbb{P}) = \frac{1}{\alpha(\alpha-1)} \log \left(\int_{\mathsf{Y}} \mu_{\boldsymbol{\lambda},\Theta} k(y)^\alpha p(y|\mathscr{D})^{1-\alpha} \nu(\mathrm{d}y) \right)$$

minimisation, with 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 the α-divergence (resp. Renyi's α-divergence), $\lambda_{j,n}$ 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

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_n}{\beta_n} \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nabla \log\left(k(\theta_{j,n}, y)\right) \nu(\mathrm{d}y)}{(\alpha-1) \int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y)} , \quad j = 1 \dots J . \tag{26}$$

In this case, we are still in the framework of Corollary 6 and $\lambda_{j,n}$ only appears through $\mu_n k$, a property also shared with the update we introduced in Corollary 5. This further underlines the importance of having worked under the general conditions on $(\boldsymbol{\lambda}_n, \Theta_n)_{n \geqslant 1}$ stated in Theorem 2.

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 update formula for mixture models from [20].

4.2.3 A Power Descent approach

The Power Descent algorithm introduced in [20] 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 $\mu \in \mathrm{M}_1(\mathsf{T})$, they seek 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 for all for all $y \in \mathsf{Y}$, we use the notation $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y)$. The optimisation is then done by applying several one-step transitions of the Power Descent algorithm: given $\mu_1 \in \mathrm{M}_1(\mathsf{T})$, they consider

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \quad n \geqslant 1 , \tag{27}$$

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 $\gamma_{j,\alpha}^n$ in (12) and for μ_n of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_{j,n}}$ with $\Theta_n = \Theta$ and $\eta_n = \eta/(1-\alpha)$ at time n, (18) and (27) coincide.

Interestingly, a monotonicity property has already been proved for the Power Descent algorithm in [20], which uses a different proof technique compared to the one used in the proof of Theorem 3. Indeed, as a particular case of [20, Theorem 1] with $\Gamma = [(\alpha-1)v + 1]^{\eta/(1-\alpha)}$, they are able 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 3, 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 7 below, where the results from [20, Theorem 1] are further extended beyond the case $\eta > 1$ when $\alpha < 0$.

Proposition 7. *Assume that p and k are as in (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 Appendix A.2 and we now make two comments. Firstly, while the results from [20] and Proposition 7 allow for a wider range of values for α and η to be used, a strong improvement compared to [20] is that by Theorem 3 we do not need to keep Θ_n constant anymore at each step of the algorithm. From there, extending Theorem 3 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 (18), we now have a better understanding of the role of the parameter η_n appearing in (18). Indeed, as underlined in [20], 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 5.

We have derived several examples where the conditions of Theorem 3 are met and connected this theorem to the Power Descent algorithm. We will conclude this section by presenting relevant particular cases of Algorithm 1. We start by investigating the case where the kernel k belongs to the Gaussian family.

4.3 Algorithm 1 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 (21), that is

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

yields the following update formulas at time n for the means $(m_{j,n+1})_{1 \leqslant j \leqslant J}$ and covariances matrices $(\Sigma_{j,n+1})_{1 \leqslant j \leqslant J}$:

$$\forall j = 1 \dots J, \quad m_{j,n+1} = \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) y \, \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y)} \tag{28}$$

$$\Sigma_{j,n+1} = \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)(y - m_{j,n})(y - m_{j,n})^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\nu(\mathrm{d}y)} \;. \tag{29}$$

Due to the intractable integrals appearing in (18), (28), and (29), 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 (18), (28), and (29). This leads to Algorithm 3 below, where based on (12) 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 3: α-divergence minimisation for Gaussian 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

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n} \left[\sum_{m=1}^M \hat{\gamma}_{\ell,\alpha}^n(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = \frac{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n})}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) \cdot (Y_{m,n} - m_{j,n})(Y_{m,n} - m_{j,n})^T}{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n})} .$$

We have thus obtained a tractable version of Algorithm 1 which allows us to iteratively update both the weights and component 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 8. *A practical version of Algorithm 1 can be derived in the particular case of Student's distributions, which could be useful for robustification purposes (see Algorithm 4 in Appendix B).*

Remark 9. *We can obtain practical versions of Algorithm 2 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 (25) 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

$$\theta_{j,n+1} = \theta_{j,n} + \gamma \frac{\int_{\mathsf{Y}} \lambda_{j,n} \gamma_{j,\alpha}^n(y)(y - \theta_{j,n}) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} , \quad j = 1 \dots J .$$

In addition, following (26) 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

$$\theta_{n+1} = (1 - \gamma)\, \theta_n + \gamma \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y)\; y\; \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) \nu(\mathrm{d}y)} , \quad j = 1 \dots J ,$$

which coincides with (28) when $\gamma = 1$. Approximated versions of the two above iterative formulas are then given respectively by

$$\forall j = 1 \dots J \ , \quad \theta_{j,n+1} = \theta_{j,n} + \gamma \frac{\lambda_{j,n} \sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^{J} \sum_{m=1}^{M} \lambda_{j,n} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \tag{30}$$

$$\theta_{n+1} = (1-\gamma)\,\theta_n + \gamma \frac{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \tag{31}$$

and tractable versions of Algorithm 2 for Gaussian mixture models can be deduced (see Algorithm 5 and 6 in Appendix C).

Lastly, we focus on the particular case $\alpha = 0$ in Algorithm 1 (and its application to the particular case of Gaussian Mixture Models as seen in Algorithm 3). As we shall see, this case can be linked to the M-PMC algorithm and it will be used to drive our numerical experiments.

4.4 The M-PMC algorithm as a particular case of Algorithm 1

We are interested in interpreting the results we have obtained thus far in the light of the M-PMC algorithm [21]. 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

$$\sup_{(\boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J)} \int_{\mathsf{Y}} \log \left(\sum_{j=1}^{J} \lambda_j k(\theta_j, y) \right) p(y) \nu(\mathrm{d}y) \ , \tag{32}$$

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 [21, Section 2] by introducing the following iterative update formulas for all $j = 1 \dots J$ and for all $n \geqslant 1$

$$\lambda_{j,n+1} = \int_{\mathsf{Y}} \frac{\lambda_{j,n} k(\theta_{j,n}, y)}{\sum_{\ell=1}^{J} \lambda_{\ell,n} k(\theta_{\ell,n}, y)} \frac{p(y)}{\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)} \nu(\mathrm{d}y) \tag{33}$$

$$\theta_{j,n+1} = \operatorname*{argmax}_{\theta_j \in \mathsf{T}} \int_{\mathsf{Y}} \frac{\lambda_{j,n} k(\theta_{j,n}, y)}{\sum_{\ell=1}^{J} \lambda_{\ell,n} k(\theta_{\ell,n}, y)} \log(k(\theta_j, y)) p(y) \nu(\mathrm{d}y) \ . \tag{34}$$

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 1, it follows that the M-PMC algorithm can be seen as a particular example of our framework.

Remark 10. *Equations (33) and (34) are presented in [21] 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 1 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 [21, Section 3] for the particular case of the Gaussian family, in which they use the sampler

$$q_n(y) = \mu_n k(y) = \sum_{j=1}^{J} \lambda_{j,n} k(\theta_{j,n}, y) \ . \tag{35}$$

Thus, comparing Algorithm 3 to the original M-PMC algorithm for Gaussian Mixture Models from [21, 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.

5 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 3. We use a similar setting to the one considered in [21]. 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 3 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 $[1/J, \dots, 1/J]$ (i.e $\boldsymbol{\lambda}_1 = [1/J, \dots, 1/J]$ 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 $n = 1 \dots N$.

As for the choice of sampler at time n, we are first interested in setting q_n as in (35), 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 [21]. We denote the resulting algorithm M-PMC(η, κ), the case $(\eta, \kappa) = (1, 0)$ corresponding to the initial M-PMC algorithm of [21].

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 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 3. Recall from Section 4.2.3 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 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

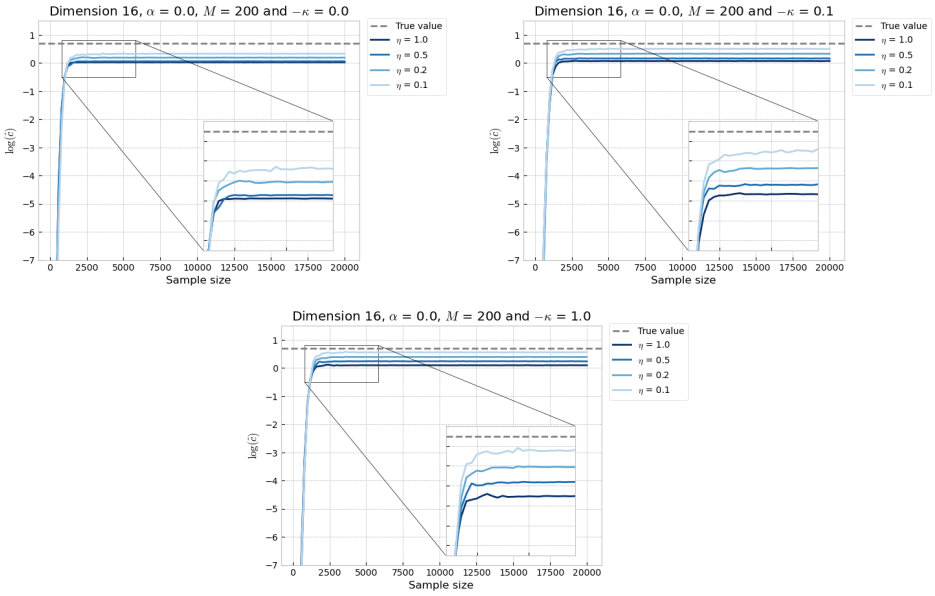

Figure 1: Normalisation constant estimation by the M-PMC(η, κ) algorithm in dimension $d = 16$ for $\eta \in \{1, 0.5, 0.2, 0.1\}$ and $-\kappa \in \{0, 0.1, 1\}$.

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 [21] 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 (18), (28), and (29), 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 2 below.

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 Appendix D for additional plots when $M = \{500, 1000\}$).

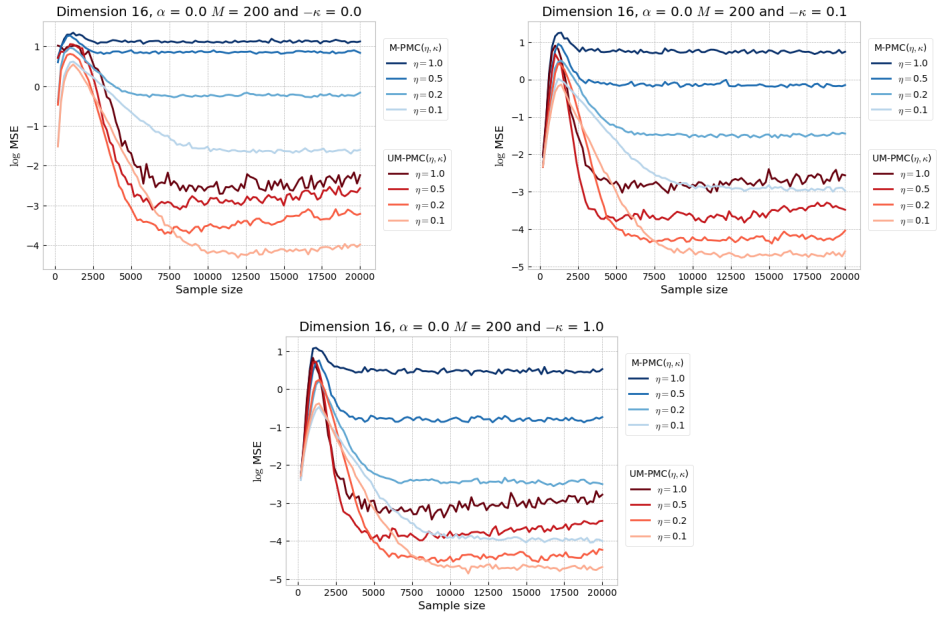

Figure 2: LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms in dimension $d = 16$ for $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$.

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 3. 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 [21].

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 9 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 (30) (see Algorithm 5 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 [23] 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 $\sigma^2_{\mathrm{MH}} \boldsymbol{I_d}$ with $\sigma^2_{\mathrm{MH}} \in \{1, 25\}$ 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 $\hat{m}^{\mathrm{PIMAIS}}_{\mathrm{approx},n}$ (we refer to [23] 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

$$\hat{m}_{\mathrm{approx},n} = \sum_{n'=1}^{n} \sum_{m=1}^{M} w_{m,n'} Y'_{m,n'}$$

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 $\hat{\mathrm{MSE}}$, 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$\hat{\mathrm{MSE}}$ (logarithm of the $\hat{\mathrm{MSE}}$) can then be visualised on Figure 3 below.

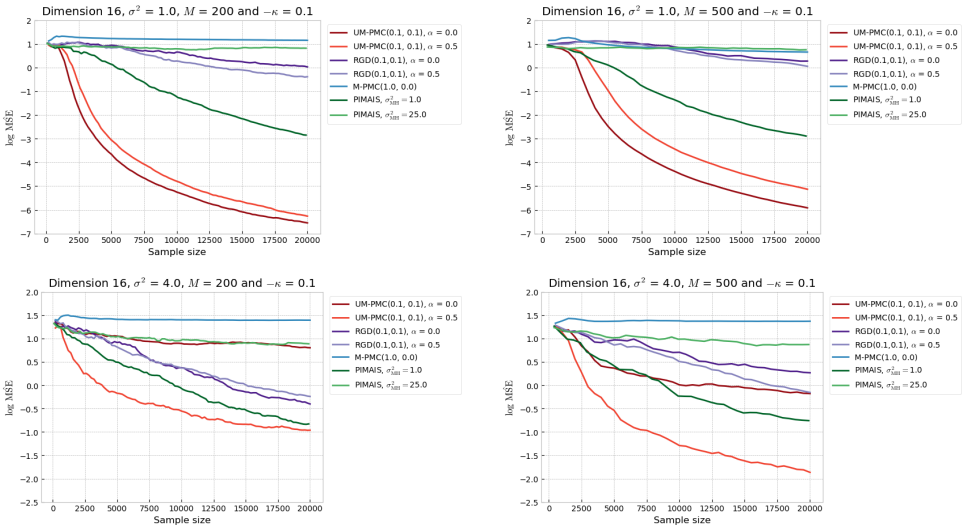

Figure 3: Log$\hat{\mathrm{MSE}}$ for the UM-PMC(η, κ) in dimension $d = 16$ for $\alpha \in \{0., 0.5\}$, $\sigma^2 \in \{1, 4\}$, $\eta = 0.1$ and $-\kappa = 0.1$ compared with the PIMAIS algorithm and the M-PMC(1., 0.) algorithm.

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$\hat{\mathrm{MSE}}$ 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 4.2.2, 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 3, which further illustrates how we were able to successfully improve on this algorithm introduced in [21] by including it into a wider framework.

6 Conclusion

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 component parameters of a given mixture model for all $\alpha \in [0, 1)$.

We then underlined the links between our approach and gradient descent schemes 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 so that it achieves 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 on both theoretical and practical levels. 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 7, another interesting direction consists in generalising the monotonicity property from Theorem 3 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 (18), (28), and (29) will result in further improved numerical results.

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] Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2014.

[6] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Feb 2017.

[7] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.

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

[9] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5581–5590, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[10] Trevor Campbell and Xinglong Li. Universal boosting variational inference. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 3484–3495. Curran Associates, Inc., 2019.

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

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

[13] 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.

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

[15] Tom Minka. Power ep. Technical Report MSR-TR-2004-149, January 2004.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

[22] 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.

[23] L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623, May 2017.

A Deferred results

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 11. *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 12. *Let $\gamma \in (0, 1]$, let g be a β-smooth function defined on T. 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

$$g(\theta) - g\left(\theta - \frac{\gamma}{\beta} \nabla g(\theta)\right) \geqslant \frac{\gamma}{2\beta} \|\nabla g(\theta)\|^2 \, .$$

□

A.2 Monotonicity property for the Power Descent

Preliminary remarks First note that for all $\eta > 0$, the iteration $\mu \mapsto \mathcal{I}_\alpha(\mu)$ is well-defined if we have

$$0 < \mu(|b_{\mu,\alpha} + \kappa|^{\frac{\eta}{1-\alpha}}) < \infty \, . \tag{36}$$

Furthermore, [20] 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 7, 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 [20] 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 [20] beyond the case $\eta \in (0, 1]$ when $\alpha < 0$. This result relies on the existence of an exponent ϱ satisfying condition (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 $\mu \mapsto \mathcal{I}_\alpha(\mu)$

(A2) We have $\varrho \in \mathbb{R} \setminus [0,1]$ and the function $f_{\alpha,\varrho} : u \mapsto f_\alpha(u^{1/\varrho})$ is non-decreasing and concave on $\mathbb{R}_{>0}$.

Proposition 13. *Assume (A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, assume that ϱ satisfies (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 $\zeta \preceq \mu$,*

$$|\varrho|^{-1} \left\{ \mu(|b_{\mu,\alpha} + \kappa|) - \mu\left(|b_{\mu,\alpha} + \kappa| g^\varrho\right) \right\} \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k) \,, \tag{37}$$

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 (37) if and only if $\zeta = \mu$.

Proof. First note that for all $\alpha \in \mathbb{R} \setminus \{1\}$, we have by (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 $\varrho^{-1} f'_\alpha(u) = |\varrho|^{-1} |f'_\alpha(u)|$, $\varrho^{-1}\kappa = |\varrho^{-1}\kappa|$ 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 (A2) and ζ,

$$\begin{aligned}
\Psi_\alpha(\zeta k) &= \int_{\mathsf{Y}} f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} f_{\alpha,\varrho}\left(\left[\frac{\zeta k(y)}{p(y)}\right]^\varrho\right) p(y)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} f_{\alpha,\varrho}\left(\left[\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} \left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\right]^\varrho\right) p(y)\nu(\mathrm{d}y) \\
&\leqslant \int_{\mathsf{Y}} f_{\alpha,\varrho}\left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} \left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^\varrho\right) p(y)\nu(\mathrm{d}y)
\end{aligned} \tag{38}$$

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

$$\begin{aligned}
u_y &= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} \left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^\varrho \\
v_y &= \left(\frac{\mu k(y)}{p(y)}\right)^\varrho
\end{aligned}$$

and note that

$$u_y - v_y = \left(\frac{\mu k(y)}{p(y)}\right)^\varrho \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta,y)}{\mu k(y)} g^\varrho(\theta) - 1\right) \tag{39}$$

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 (38), we get

$$\begin{aligned}
\Psi_\alpha(\zeta k) &\leqslant \int_{\mathsf{Y}} f_{\alpha,\varrho}(u_y) p(y)\nu(\mathrm{d}y) \\
&\leqslant \int_{\mathsf{Y}} f_{\alpha,\varrho}(v_y) p(y)\nu(\mathrm{d}y) + \int_{\mathsf{Y}} f'_{\alpha,\varrho}(v_y)(u_y - v_y) p(y)\nu(\mathrm{d}y)
\end{aligned} \tag{40}$$

Note that the first term of the rhs can be written as

$$\int_{\mathsf{Y}} f_{\alpha,\varrho}(v_y) p(y)\nu(\mathrm{d}y) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) = \Psi_\alpha(\mu k) \tag{41}$$

Using now $f'_{\alpha,\varrho}(v_y) = \varrho^{-1} v_y^{1/\varrho - 1} f'_\alpha(v_y^{1/\varrho})$ and (39), the second term of the rhs of (40) may be expressed as

$$\begin{aligned}
&\int_{\mathsf{Y}} f'_{\alpha,\varrho}(v_y)(u_y - v_y)p(y)\nu(\mathrm{d}y) \\
&= \varrho^{-1} \int_{\mathsf{Y}} \left(\frac{\mu k(y)}{p(y)}\right)^{1-\varrho} f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \\
&\qquad\qquad \left(\frac{\mu k(y)}{p(y)}\right)^{\varrho} \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \frac{k(\theta, y)}{\mu k(y)} g^\varrho(\theta) - 1\right) p(y)\nu(\mathrm{d}y) \\
&= \varrho^{-1} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)\right) g^\varrho(\theta) \\
&\qquad\qquad - \varrho^{-1} \int_{\mathsf{Y}} \mu k(y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y) \\
&= \varrho^{-1} \left\{\mu\left(b_{\mu,\alpha} \cdot g^\varrho\right) - \mu(b_{\mu,\alpha})\right\} \\
&= |\varrho|^{-1} \left\{\mu\left(|b_{\mu,\alpha} + \kappa| g^\varrho\right) - \mu(|b_{\mu,\alpha} + \kappa|)\right\} + |\varrho^{-1}\kappa|(1 - \mu(g^\varrho)) \,,
\end{aligned}$$

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

$$\int_{\mathsf{Y}} f'_{\alpha,\varrho}(v_y)(u_y - v_y)p(y)\nu(\mathrm{d}y) \leqslant |\varrho|^{-1} \left\{\mu\left(|b_{\mu,\alpha} + \kappa| g^\varrho\right) - \mu(|b_{\mu,\alpha} + \kappa|)\right\} \,. \tag{42}$$

Combining this inequality with (40) and (41) finishes the proof of the inequality. Furthermore, if the equality holds in (37), then the equality in Jensen's inequality (42) shows that g is constant μ-a.e. so that $\zeta = \mu$, and the proof is completed. □

Remark 14. *The proof of Proposition 13 relies on f'_α being of constant sign. Notice however that the definition of the α-divergence in (1) is invariant with respect to the transformation $\tilde{f}_\alpha(u) = f_\alpha(u) + \kappa(u-1)$ for any arbitrary constant κ, that is f_α can be equivalently replaced by $\tilde{f}_\alpha$ in (1). This aspect is in fact expressed through the constant κ appearing in the update formula.*

We now plan on setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Proposition 13 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 13, we strengthen the condition (36) as follows to take into account the exponent ϱ

$$0 < \mu(|b_{\mu,\alpha} + \kappa|^{\frac{\eta}{1-\alpha}}) < \infty \text{ and } \mu(|b_{\mu,\alpha} + \kappa| g^\varrho) \leqslant \mu(|b_{\mu,\alpha} + \kappa|) \quad \text{with } g = \frac{|b_{\mu,\alpha}+\kappa|^{\frac{\eta}{1-\alpha}}}{\mu(|b_{\mu,\alpha}+\kappa|^{\frac{\eta}{1-\alpha}})} \,. \tag{43}$$

This leads to the following result.

Proposition 15. *Assume (A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, assume that ϱ satisfies (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 (43). 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 13 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,

$$\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k) + |\varrho|^{-1} \left\{\mu\left(|b_{\mu,\alpha}| g^\varrho\right) - \mu(|b_{\mu,\alpha}|)\right\} \leqslant \Psi_\alpha(\mu k) \tag{44}$$

where the last inequality follows from condition (43).

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 (44) 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 13. Therefore, $\mathcal{I}_\alpha(\mu) = \mu$. □

While Proposition 15 resembles [20, Theorem 1] 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 7. The proof of this theorem requires intermediate results, which are derived in Appendix A.2 alongside with the proof of Proposition 7.

Proof of Proposition 7

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 13, 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 13 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 16 below.

Lemma 16. *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,

$$\mu\left(|b_{\mu,\alpha}|g^\varrho\right) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta)\left(\frac{|b_{\mu,\alpha}(\theta)|^{1/(1-\varrho)}}{g(\theta)}\right)^{1-\varrho} g(\theta) \geqslant \left[\mu\left(|b_{\mu,\alpha}|^{1/(1-\varrho)}\right)\right]^{1-\varrho} \tag{45}$$

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 17. *Assume (A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$ and assume that ϱ satisfies (A2). Then $\eta = (1-\alpha)/(1-\varrho)$ satisfies (43) for any $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_{\mu,\alpha}|) < \infty$.*

Proof. We apply (45) with $g = 1$ and get that

$$\left[\mu\left(|b_{\mu,\alpha}|^{1/(1-\varrho)}\right)\right]^{1-\varrho} \leqslant \mu(|b_{\mu,\alpha}|) < \infty \ . \tag{46}$$

Then (43) 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 (A1) for $\phi > 0$, we obtain $\mu(|b_{\mu,\alpha}|^\phi) > 0$, which concludes the proof. □

While Lemma 17 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 (43) for any $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\mu(|b_{\mu,\alpha}|) < \infty$.

Lemma 18. *Assume (A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$ and assume that ϱ satisfies (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 (43).*

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 (43). Set $\phi = \eta/(1-\alpha)$. Using that $\mu(|b_{\mu,\alpha}|) < \infty$ when $\phi < 0$ and (A1) for $\phi > 0$, we obtain $\mu(|b_{\mu,\alpha}|^{\phi}) > 0$, which concludes the proof. □

Lemma 17 and Lemma 18 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 (A2). Now, in order to prove Proposition 7 and given $\alpha \in \mathbb{R} \setminus \{1\}$, we need to check which values of ϱ satisfy the conditions expressed in (A2).

Proof of Proposition 7. The proof consists in verifying that we can apply Proposition 15, that is, given $\alpha \in \mathbb{R} \setminus \{1\}$, we must find a range of constants ϱ which satisfy (A2). We then use Lemma 17 or Lemma 18 to deduce that, for the provided constants η, (43) holds.

(i) Assumption (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 17 and Lemma 18, (43) 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 (A2) holds for $\varrho \leqslant \alpha$ if $\alpha < 0$ Lemmas 17 and 18 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 17 and 18 can indeed be applied.

Since $u f'_\alpha(u) = \alpha f_\alpha(u) + 1/(\alpha - 1)$, we have, for all $n \geqslant 1$,

$$\begin{aligned} \mu(|b_{\mu,\alpha}|) &= \int_{\mathsf{Y}} \left| \left(\frac{\mu k(y)}{p(y)} \right) f'_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \right| p(y) \nu(\mathrm{d}y) \\ &\leqslant |\alpha| \int_{\mathsf{Y}} \left| f_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \right| p(y) \nu(\mathrm{d}y) + \frac{1}{|\alpha - 1|} \end{aligned} \tag{47}$$

Using that $\Psi_\alpha(\mu k) > -\infty$ (which is a consequence of (A1) and of Jensen's inequality applied to the convex function $u \mapsto u f_\alpha(1/u)$), 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.

□

B Algorithm 1 within 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 2, solving

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

yields the following update formulas

$$\begin{aligned} \forall j = 1 \dots J, \quad m_{j,n+1} &= \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) g_j^n(y) y \, \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) g_j^n(y) \nu(\mathrm{d}y)} \\ \Sigma_{j,n+1} &= \frac{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) g_j^n(y) (y - m_{j,n})(y - m_{j,n})^T \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \gamma_{j,\alpha}^n(y) g_j^n(y) \nu(\mathrm{d}y)} \ , \end{aligned}$$

where for all $y \in \mathsf{Y}$ and for all $j = 1 \dots J$, we have set

$$g_j^n(y) = \frac{\nu_j + d}{\nu_j + (y - m_{j,n})^T (\Sigma_{j,n})^{-1} (y - m_{j,n})} \ .$$

Based on Algorithm 1 and given a sequence of samplers $(q_n)_{n \geqslant 1}$, one may consider in practice Algorithm 4 below.

Algorithm 4: 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

$$\begin{aligned} \lambda_{j,n+1} &= \frac{\lambda_{j,n} \left[\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{l,n} \left[\sum_{m=1}^M \hat{\gamma}_{\ell,\alpha}^n(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}} \\ m_{j,n+1} &= \frac{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) g_j^n(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) g_j^n(Y_{m,n})} \\ \Sigma_{j,n+1} &= \frac{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) g_j^n(Y_{m,n}) \cdot (Y_{m,n} - m_{j,n})(Y_{m,n} - m_{j,n})^T}{\sum_{m=1}^M \hat{\gamma}_{j,\alpha}^n(Y_{m,n}) g_j^n(Y_{m,n})} \ . \end{aligned}$$

C Practical versions of Algorithm 2 within the Gaussian family

Based on the updates (30) and (31) from Remark 9, we obtain below the two practical algorithms for Gaussian Mixture Models (GMMs) optimisation.

Algorithm 5: α-divergence minimisation for GMMs based on (30)

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

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\sum_{m=1}^{M} \hat{\gamma}_{\ell,\alpha}^{n}(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$\theta_{j,n+1} = \theta_{j,n} + \gamma \frac{\lambda_{j,n} \sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^{J} \sum_{m=1}^{M} \lambda_{j,n} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \, .$$

Algorithm 6: α-divergence minimisation for GMMs based on (31)

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

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\sum_{m=1}^{M} \hat{\gamma}_{\ell,\alpha}^{n}(Y_{m,n}) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$\theta_{j,n+1} = \theta_{j,n} + \gamma \frac{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^{M} \hat{\gamma}_{j,\alpha}^{n}(Y_{m,n})} \, .$$

D Additional numerical experiments

In this section we provide further plots based on the numerical experiments from Section 5.

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 and Figure 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.

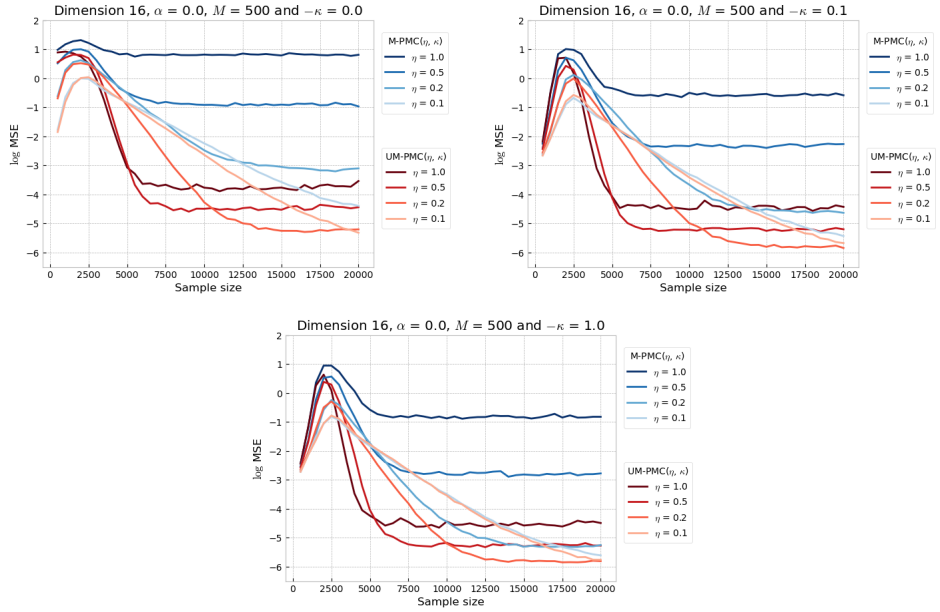

Figure 4: $\underline{M = 500.}$ LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms with $d = 16$, $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$.

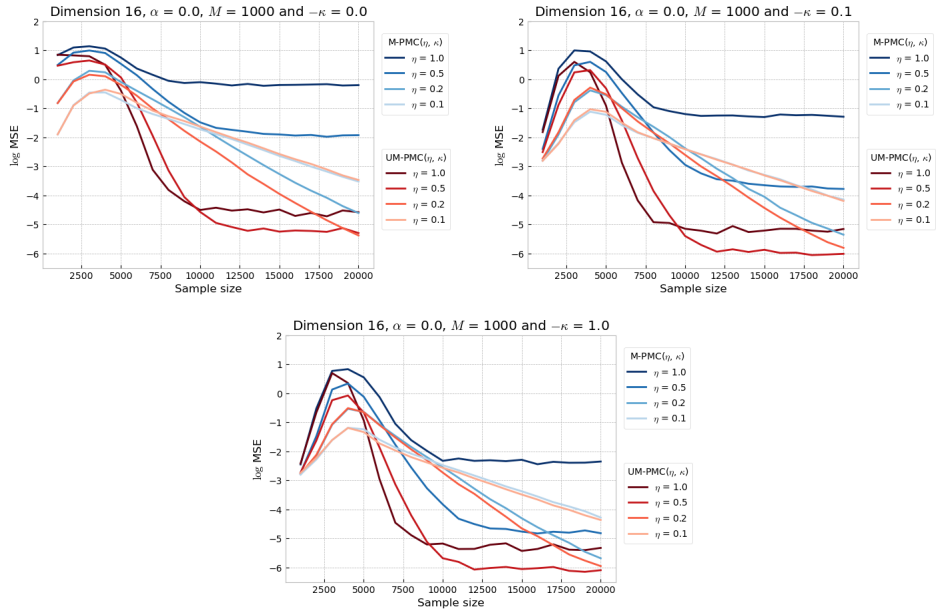

Figure 5: $\underline{M = 1000.}$ LogMSE comparison for the M-PMC(η, κ) and the UM-PMC(η, κ) algorithms with $d = 16$, $\eta \in \{1.0, 0.5, 0.2, 0.1\}$ and $-\kappa = \{0, 0.1, 1\}$.