Infinite-dimensional gradient-based descent for alpha-divergence minimisation

Kamélia Daudel, Randal Douc, François Portier

INFINITE-DIMENSIONAL GRADIENT-BASED DESCENT FOR ALPHA-DIVERGENCE MINIMISATION

By Kamélia Daudel*, Randal Douc† and François Portier*.

*Télécom Paris** *and Télécom SudParis*†

This paper introduces the (α, Γ)-descent, an iterative algorithm which operates on measures and performs α-divergence minimisation in a Bayesian framework. This gradient-based procedure extends the commonly-used variational approximation by adding a prior on the variational parameters in the form of a measure. We prove that for a rich family of functions Γ, this algorithm leads at each step to a systematic decrease in the α-divergence. Our framework recovers the Entropic Mirror Descent (MD) algorithm with improved $O(1/N)$ convergence results and provides an alternative to the Entropic MD that we call the Power descent and for which we prove convergence to an optimum. Moreover, the (α, Γ)-descent allows to optimise the mixture weights of any given mixture model without any information on the underlying distribution of the variational parameters. This renders our method compatible with many choices of parameters updates and applicable to a wide range of Machine Learning tasks. We demonstrate empirically on both toy and real-world examples the benefit of using the Power descent and going beyond the Entropic MD framework, which fails as the dimension grows.

1. Introduction. Bayesian statistics for complex models often induce intractable and hard-to-compute posterior densities which need to be approximated. Variational methods such as Variational Inference (VI) [1, 2] and Expectation Propagation (EP) [3, 4] consider this objective purely as an optimisation problem (which is often non-convex). These approaches seek to approximate the posterior density by a simpler variational density k_θ, characterized by a set of variational parameters $\theta \in \mathsf{T}$, where T is the parameter space. In these methods θ is optimised such that it minimizes a certain objective function, typically the Kullback-Leibler divergence [5] between the posterior and the variational density.

Modern Variational methods improved in three major directions [6, 7] (i) Black-Box inference techniques [8, 9] and Hierarchical Variational Inference methods [10, 11] have been deployed, expanding the variational family

MSC 2010 subject classifications: Primary 62F15; secondary 62F30, 62F35, 62G07, 62L99

Keywords and phrases: Alpha-divergence, Kullback-Leibler divergence, Mirror Descent, Variational Inference

and rendering Variational methods applicable to a wide range of models (ii) Algorithms based on alternative families of divergences such as the α-divergence [12, 13] and Renyi's α-divergence [14, 15] have been introduced [16, 17, 18, 19, 20, 21, 22] to bypass practical issues linked to the Kullback-Leibler divergence [4, 6, 23] (iii) Scalable methods relying on stochastic optimisation techniques [24, 25] have been developed to enable large-scale learning and have been applied to complex probabilistic models [23, 26, 27, 28].

In the spirit of Hierarchical Variational Inference, we offer in this paper to enlarge the variational family by adding a prior on the variational density k_θ and consider

$$q(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \ .$$

This is a more general form than the one found in [11] where μ is parametrised by another parametric model. As for the objective function, we work within the α-divergence family, which admits the Kullback-Leibler and the reverse Kullback-Leibler as limiting cases. These divergences belong to the f-divergence family [29, 30] and as such, they have convexity properties so that the minimisation of the α-divergence between the targeted posterior density and the variational density q with respect to μ can be seen as a convex optimisation problem. The paper is then organised as follows:

- In Section 2, we briefly review basic concepts around the α-divergence family before recalling the basics of Variational methods and formulating formally the optimisation problem we consider.
- In Section 3, we describe the Exact (α, Γ)-descent, an iterative algorithm that performs α-divergence minimisation by updating the measure μ. We establish in Theorem 1 sufficient conditions on Γ for this algorithm to lead at each step to a systematic decrease in the α-divergence. We then investigate the convergence of the algorithm in Theorem 2, 3 and 4. Strikingly, the Infinite-dimensional Entropic Mirror Descent [31, Appendix A] is included in our framework and we obtain an improved $O(1/N)$ convergence rate, which illustrates the generality of our approach. We also introduce a novel algorithm called the Power descent, for which we prove convergence to an optimum and obtain an $O(1/N)$ convergence rate when $\alpha > 1$.
- In Section 4, we define the Stochastic version of the Exact (α, Γ)-descent and apply it to the important case of mixture models [32, 33]. The resulting general-purpose algorithm is Black-Box and does not require any information on the underlying distribution of the variational parameters.
- Finally, Section 5 is devoted to numerical experiments. We demonstrate the benefit of using the Power descent and thus of going beyond the Entropic Mirror Descent framework. We also compare our method to a computa-

tionally equivalent Adaptive Importance Sampling algorithm for Bayesian Logistic Regression on a large dataset.

2. Formulation of the optimisation problem.

2.1. *The α-divergence.* Let $(\mathsf{Y}, \mathcal{Y}, \nu)$ be a measured space, where ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$. Let $\mathbb{Q}$ and $\mathbb{P}$ be two probability measures on $(\mathsf{Y}, \mathcal{Y})$ that are absolutely continuous with respect to ν i.e. $\mathbb{Q} \preceq \nu$, $\mathbb{P} \preceq \nu$. Let us denote by $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$ and $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$ the Radon-Nikodym derivatives of $\mathbb{Q}$ and $\mathbb{P}$ with respect to ν.

Definition 1. *Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. The α-divergence and the Kullback-Leibler (KL) divergence between $\mathbb{Q}$ and $\mathbb{P}$ are respectively defined by :*

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

As $\lim_{\alpha\to 0} D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{P}||\mathbb{Q})$ and $\lim_{\alpha\to 1} D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{Q}||\mathbb{P})$ (see for example [15]), the definition of the α-divergence can be extended to 0 and 1 by continuity and we will use the notation $D_0(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{P}||\mathbb{Q})$ and $D_1(\mathbb{Q}||\mathbb{P}) = D_{KL}(\mathbb{Q}||\mathbb{P})$ throughout the paper. Letting f_α be the convex function on $(0, +\infty)$ defined by $f_0(u) = u - 1 - \log(u)$, $f_1(u) = 1 - u + u\log(u)$ and $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$ for all $\alpha \in \mathbb{R} \setminus \{0, 1\}$, we have that for all $\alpha \in \mathbb{R}$,

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

Written under that form, the r.h.s of (1) corresponds to the general definition of the α-divergence, that is q and p do not need to be normalised in (1) in order to define a divergence. We next remind the reader of a few more results about the α-divergence and we refer to [15, 34, 35, 36] for more details on the α-divergence family.

Proposition 2. *The α-divergence is always non-negative and it is equal to zero if and only if $\mathbb{Q} = \mathbb{P}$. Furthermore, it is jointly convex in $\mathbb{Q}$ and $\mathbb{P}$ and for all $\alpha \in \mathbb{R}$, $D_\alpha(\mathbb{Q}||\mathbb{P}) = D_{1-\alpha}(\mathbb{P}||\mathbb{Q})$.*

Special cases of the α-divergence family include the Hellinger distance [37, 38] and the χ^2-divergence [20] which correspond respectively to order $\alpha = 0.5$ and $\alpha = 2$.

2.2. *Variational Inference within the α-divergence family.* Assume that we have access to some observed variables $\mathscr{D}$ generated from a probabilistic model $p(\mathscr{D}|y)$ parameterised by a hidden random variable $y \in \mathsf{Y}$ that is drawn from a certain prior $p_0(y)$. Bayesian inference involves being able to compute or sample from the posterior density of the latent variable y given the data $\mathscr{D}$:

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

where $p(\mathscr{D}) = \int_{\mathsf{Y}} p_0(y)p(\mathscr{D}|y)\nu(\mathrm{d}y)$ is called the *marginal likelihood* or *model evidence*. For many useful models the posterior density is intractable due to the normalisation constant $p(\mathscr{D})$. One way to bypass this problem is to introduce a variational density q in some tractable density family $\mathcal{Q}$ and to find $q^\star$ such that

$$q^\star = \mathrm{arginf}_{q\in\mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}) \ ,$$

where $\mathbb{P}$ and $\mathbb{Q}$ denote the probability measures on $(\mathsf{Y}, \mathcal{Y})$ with corresponding associated density $p(\cdot|\mathscr{D})$ and q. This optimisation problem still involves the unknown normalisation constant $p(\mathscr{D})$ however it can easily be transformed into the following equivalent optimisation problem

$$q^\star = \mathrm{arginf}_{q\in\mathcal{Q}} \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y,\mathscr{D})}\right) p(y,\mathscr{D})\nu(\mathrm{d}y) \ ,$$

which does not involve the unknown marginal likelihood $p(\mathscr{D})$ anymore (see for example [6] and [19, 20]). The core of Variational Inference methods then consists in designing approximating families $\mathcal{Q}$ which allow efficient optimisation and which are able to capture complicated structure inside the posterior density. Typically, q belongs to a parametric family $q = k_\theta$ where θ is in a certain parametric space T, that is the minimisation occurs over the set of densities

$$\{y \mapsto k_\theta(y) \ : \ \theta \in \mathsf{T}\} \ .$$

In this paper, we offer to perform instead a minimization over

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

where M is a convenient subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on T (and in this case, we equip T with a σ-field denoted by $\mathcal{T}$). In doing so, we extend the minimizing set to a larger space since a parameter θ can be identified with its associated Dirac measure δ_θ. Similarly, a mixture model composed of $\{\theta_1, ..., \theta_J\} \in \mathsf{T}^J$ will correspond to taking μ as a weighted sum of Dirac measures.

More formally, let us consider a measurable space $(\mathsf{T}, \mathcal{T})$. Let p be a measurable positive function on $(\mathsf{Y}, \mathcal{Y})$ and $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ be a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density k defined on $\mathsf{T} \times \mathsf{Y}$. Moreover, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, for all $y \in \mathsf{Y}$, we denote $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y)$ and we define

$$\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ . \tag{2}$$

Note that p, k and ν appear as well in $\Psi_\alpha(\mu)$ i.e $\Psi_\alpha(\mu) = \Psi_\alpha(\mu; p, q, \nu)$, but we drop them for notational ease and when no ambiguity occurs. Notice also that we replaced $k_\theta(y)$ by $k(\theta, y)$ to comply with usual kernel notation. We consider in what follows the general optimisation problem

$$\mathrm{arginf}_{\mu \in \mathsf{M}} \Psi_\alpha(\mu) \ , \tag{3}$$

and in practice, we will choose $p(y) = p(y, \mathscr{D})$.

At this stage, a first remark is that the convexity of Ψ_α is straightforward from the convexity of f_α. Therefore, a simple yet powerful consequence of enlarging the variational family is that the optimisation problem now involves the *convex* mapping

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

whereas the initial optimisation problem was associated to the mapping $\theta \mapsto \int_{\mathsf{Y}} f_\alpha\left(\frac{k_\theta(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$, which is not necessarily convex.

We now move on to Section 3, where we describe the (α, Γ)-descent and state our main theoretical results.

3. The (α, Γ)-descent.

3.1. *An iterative algorithm for optimising Ψ_α.* Throughout the paper we will assume the following conditions on k, p and ν.

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

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

LEMMA 3. *Suppose that (A1) holds. Then, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, we have*

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

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

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

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

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

REMARK 4. *Assumption (A1) can be extended by discarding the assumption that $p(y)$ is positive for all $y \in \mathsf{Y}$. As it complicates the expression of the constant appearing in the bound without increasing dramatically the degree of generality of the results, we chose to maintain this assumption for the sake of simplicity.*

Thus, if there exists a sequence of probability measures $\{\mu_n \, : \, n \in \mathbb{N}^\star\}$ on $(\mathsf{T}, \mathcal{T})$ such that $\Psi_\alpha(\mu_1) < \infty$ and $\Psi_\alpha(\mu_n)$ is non-increasing with n, Lemma 3 guarantees that this sequence converges to a limit in $\mathbb{R}$. We now focus on constructing such a sequence $\{\mu_n \, : \, n \in \mathbb{N}^\star\}$.

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

Algorithm 1: *Exact (α, Γ)-descent one-step transition*

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

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

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \, , \qquad n \in \mathbb{N}^\star \, . \tag{4}$$

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

$$\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty \, . \tag{5}$$

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

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

We refer to [31, Appendix A] for some theoretical background on the Infinite-Dimensional Entropic Mirror Descent. In this light, $b_{\mu,\alpha}$ can be understood as the gradient of Ψ_α. Algorithm 1 then consists in applying a transform function Γ to the gradient $b_{\mu,\alpha}$ and projecting back onto the space of measures.

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

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

$$A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha \left(\frac{g(\theta)\mu k(y)}{p(y)} \right) [1 - g(\theta)] \ , \tag{6}$$

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

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

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta) \ , \tag{7}$$

Moreover, equality holds in (7) if and only if $\zeta = \mu$.

Proof. To prove (7), we introduce the intermediate function

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

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

$$h_\alpha(\zeta, \mu) \geqslant \int_{\mathsf{Y}} \nu(\mathrm{d}y) p(y) f_\alpha \left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) g(\theta)}{p(y)} \right) = \Psi_\alpha(\zeta) \ . \tag{8}$$

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

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \, . \tag{9}$$

Now integrating over T with respect to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$ and then integrating over Y with respect to $p(y)\nu(\mathrm{d}y)$ in (9) yields

$$\Psi_\alpha(\mu) \geqslant h_\alpha(\zeta, \mu) + A_\alpha \, . \tag{10}$$

Combining this result with (8) gives (7). The case of equality is obtained using the strict convexity of f_α in (8) and (9) which shows that g is constant μ-a.e. so that $\zeta = \mu$. □

We now plan on setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Lemma 5 and obtain that one iteration of the (α, Γ)-descent yields $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu)$. Based on the lower-bound obtained in Lemma 5, a sufficient condition is to prove that taking $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ in (6) implies $A_\alpha \geqslant 0$. For this purpose, let us denote by Δ_α an interval of $\mathbb{R}$ such that for all $\theta \in \mathsf{T}$, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $b_{\mu,\alpha}(\theta) + \kappa$ and $\mu(b_{\mu,\alpha}) + \kappa \in \Delta_\alpha$ and let us make an assumption on (Γ, κ).

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

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

We now state our first main theorem.

THEOREM 1. *Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that (5) holds and $\Psi_\alpha(\mu) < \infty$. Then, the two following assertions hold.*

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

PROOF. To prove (i), we set $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ in (6) and we will show that $A_\alpha \geqslant 0$. Then, the proof is concluded by setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Lemma 5 as

$$\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu) - A_\alpha \leqslant \Psi_\alpha(\mu) \, . \tag{11}$$

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

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

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

where we used that $\mu[\kappa(1 - g)] = 0$ in the last equality. Setting $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,1} + \kappa))$ for all $v \in \Delta_1$, we have $g = \tilde{\Gamma} \circ (b_{\mu,1} + \kappa)$. Let us thus consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and let V be the random variable $V(\theta) = b_{\mu,1}(\theta) + \kappa$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and we can write

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

Under (A2) with $\alpha = 1$, $v \mapsto \log \tilde{\Gamma}(v) + v$ and $v \mapsto 1 - \tilde{\Gamma}(v)$ are increasing on Δ_1 which implies $A_1 \geqslant 0$.

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

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

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

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

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

Let us now show (ii). The *if* part is obvious. As for the *only if* part, $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) = \Psi_\alpha(\mu)$ combined with (11) yields

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

which is the case of equality in Lemma 5. Therefore, $\mathcal{I}_\alpha(\mu) = \mu$. □

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

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

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

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

Improving upon Lemma 5. In the following Lemma, we derive an explicit lower-bound for $\Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu)$ in terms of the variance of $b_{\mu,\alpha}$. Let us thus consider the probability space $(\mathsf{T}, \mathcal{T}, \mu)$ and denote by $\mathbb{V}\mathrm{ar}_\mu$ the associated variance operator.

LEMMA 6. *Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that (5) holds and $\Psi_\alpha(\mu) < \infty$. Then,*

$$\frac{L_{\alpha,1}}{2} \mathbb{V}\mathrm{ar}_\mu \left(b_{\mu,\alpha} \right) \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \,, \tag{12}$$

where

$$L_{\alpha,1} := \inf_{v \in \Delta_\alpha} \left\{ [(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \right\} \times \inf_{v \in \Delta_\alpha} -\Gamma'(v) \,. \tag{13}$$

The proof of Lemma 6 is deferred to Appendix A.1.

Lemma 6 can be interpreted in the following way: provided that $L_{\alpha,1} > 0$, (12) states that the case of equality is reached if and only if the variance of the gradient $b_{\mu,\alpha}$ equals zero. Such a result, which holds for any transform

function Γ satisfying (A2), quantifies the improvement after one step of the (α, Γ)-descent.

Interestingly, monotonicity properties akin to Lemma 6 have previously been derived under stronger smoothness assumptions in the context of Projected Gradient Descent steps. For example, in the particular case where the objective function f is assumed to be β-smooth on $\mathbb{R}$, for all $u \in \mathbb{R}$ it holds (see for example [39, Equation 3.5]) that

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

This result is then used to obtain improved convergence rates compared to regular Projected Gradient Descent. Consequently, we are next interested in proving a rate of convergence for the Exact (α, Γ)-descent by leveraging Lemma 6.

3.3. *Convergence.* Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$. We want to study the limiting behavior of the Exact (α, Γ)-descent for the iterative sequence of probability measure $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (4). To do so, we first introduce the two following useful quantities

$$L_{\alpha,2}^{-1} := \inf_{v\in\Delta_\alpha} (-\log\Gamma)'(v) , \tag{14}$$

$$L_{\alpha,3}^{-1} := \inf_{v\in\Delta_\alpha} \Gamma(v) , \tag{15}$$

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

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

We are now able to derive our second main result.

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

(i) The sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (4) is well-defined and the sequence $(\Psi_\alpha(\mu_n))_{n\in\mathbb{N}^\star}$ is non-increasing.

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

$$\frac{1}{N}\sum_{n=1}^{N}\Psi_\alpha(\mu_n) - \Psi_\alpha(\mu^\star) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\delta_1\right] , \tag{16}$$

where $\mu^\star$ *is such that* $\Psi_\alpha(\mu^\star) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta)$ *and where we have defined* $\delta_1 = \Psi_\alpha(\mu_1) - \Psi_\alpha(\mu^\star)$ *and* $KL(\mu^\star||\mu_1) = \int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu^\star}{\mathrm{d}\mu_1}\right) \mathrm{d}\mu^\star$.

Proof. We prove the assertions successively.

(i) The proof of (i) simply consists in verifying that we can apply Theorem 1. For all $\mu \in \mathrm{M}_1(\mathsf{T})$, (5) holds as we have

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

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

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

$$\delta_n \leqslant L_{\alpha,2}\left[\int_{\mathsf{T}} \log\left(\frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}\right) \mathrm{d}\mu^\star + \frac{L}{2}\mathbb{V}\mathrm{ar}_{\mu_n}(b_{\mu_n,\alpha}) L_{\alpha,3}\right] . \tag{17}$$

The convexity of f_α implies that

$$\delta_n \leqslant \int_{\mathsf{T}} b_{\mu_n,\alpha}(\mathrm{d}\mu_n - \mathrm{d}\mu^\star) = \int_{\mathsf{T}} (\mu_n(b_{\mu_n,\alpha}) - b_{\mu_n,\alpha})\mathrm{d}\mu^\star . \tag{18}$$

In addition, the concavity of $-\log\Gamma$ implies that for all $u, v \in \Delta_\alpha$,

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

i.e

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

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

$$v - u \leqslant \frac{\log\Gamma(u) - \log\Gamma(v)}{(-\log\Gamma)'(v)} . \tag{19}$$

We can apply (19) with $u = b_{\mu_n,\alpha}(\theta)$ and $v = \mu_n(b_{\mu_n,\alpha})$ which yields

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

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

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

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

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

Consequently, we obtain

(20)

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

Next, we show that

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

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

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

which in turn implies

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

Finally, we obtain

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

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

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

which combined with (20) implies (17). To conclude, we apply Lemma 6 to $g = \frac{\mathrm{d}\mu_{n+1}}{\mathrm{d}\mu_n}$ and combining with (17), we obtain

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

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

$$\frac{1}{N}\sum_{n=1}^{N}\Psi_\alpha(\mu_n) - \Psi_\alpha(\mu^\star) \leqslant \frac{L_{\alpha,2}}{N}\Bigg[KL(\mu^\star||\mu_1) - KL(\mu^\star||\mu_{N+1}) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}(\delta_1 - \delta_{N+1})\Bigg]$$

and we recover (16) using that $KL(\mu^\star||\mu_{N+1}) \geqslant 0$ and that $\delta_{N+1} \geqslant 0$.

□

Remark 7. *Let us comment on (16). For all $N \in \mathbb{N}^\star$, define $\bar{\mu}_N = \frac{1}{N}\sum_{n=1}^{N}\mu_n$. Then, the convexity of the mapping $\mu \mapsto \Psi_\alpha(\mu)$ yields*

$$\Psi_\alpha\left(\bar{\mu}_N\right) - \Psi_\alpha(\mu^\star) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\delta_1\right] .$$

However, since $(\Psi_\alpha(\mu_n))_{n\in\mathbb{N}^\star}$ is non-increasing under the assumptions of Theorem 2 we also have

$$\Psi_\alpha(\mu_N) - \Psi_\alpha(\mu^\star) \leqslant \frac{L_{\alpha,2}}{N}\left[KL(\mu^\star||\mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\delta_1\right] ,$$

which illustrates the fact that μ_n is improved at each step and that we shall use μ_N directly instead of $\bar{\mu}_N$.

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

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

- *(i) Kullback-Leibler divergence ($\alpha = 1$): $\Gamma(v) = e^{-\eta v}$, $\eta \in (0,1)$ and κ is any real number (Entropic Mirror Descent);*
- *(ii) Reverse Kullback-Leibler ($\alpha = 0$) and α-Divergence with $\alpha \in \mathbb{R}\backslash\{0,1\}$:*

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

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

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ *be such that* $\Psi_\alpha(\mu_1) < \infty$. *Then, the sequence* $(\mu_n)_{n\in\mathbb{N}^\star}$ *defined by* (4) *is well-defined and the sequence* $(\Psi_\alpha(\mu_n))_{n\in\mathbb{N}^\star}$ *is non-increasing with a convergence rate characterized by* (16).

The proof of Theorem 3 can be found in Appendix A.2.

Remark 8. *The assumption* $|b|_{\infty,\alpha} < \infty$ *can be discarded in Theorem 3 as* $|b|_\alpha := \sup_{n\in\mathbb{N}^\star,\theta\in\mathsf{T}} |b_{\mu_n,\alpha}(\theta)| < \infty$. *Indeed, using that for all* $\alpha \in \mathbb{R}$ *and for all* $u \in \mathbb{R}_{>0}$, $uf'_\alpha(u) = \alpha f_\alpha(u) + (u-1)$, *we can write that for all* $n \in \mathbb{N}^\star$,

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

Under (A1), *we have* $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$, *which settles the case* $\alpha = 0$. *As for the case* $\alpha \in \mathbb{R}\setminus\{0\}$, *we obtain from Lemma 3 that the r.h.s is finite if and only if* $\Psi_\alpha(\mu_n)$ *is finite, which is implied by the assumption* $\Psi_\alpha(\mu_1) < \infty$ *in Theorem 3.*

As a consequence, we can redefine Δ_α *as an interval of* $\mathbb{R}$ *which contains*

$$\{b_{\mu_n,\alpha}(\theta) + \kappa, \mu_n(b_{\mu_n,\alpha}) + \kappa \;:\; n \in \mathbb{N}^\star, \theta \in \mathsf{T}\}\;,$$

that is, we can take

$$\Delta_\alpha = \begin{cases} [-|b|_1 + \kappa, |b|_1 + \kappa] & \text{if } \alpha = 1, \\ [\frac{1}{1-\alpha} + \kappa, |b|_\alpha + \kappa], & \text{if } \alpha > 1, \\ [-|b|_\alpha + \kappa, \frac{1}{1-\alpha} + \kappa], & \text{otherwise.} \end{cases}$$

The case of the Power descent for $\alpha < 1$ is trickier. We introduce the following additive set of assumptions

(A4) (i) T is a compact metric space and $\mathcal{T}$ is the associated Borel σ-field;

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

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

Here, condition (A4)-(iii) implies that $\Psi_\alpha(\mu)$ and $b_{\mu,\alpha}(\theta)$ are uniformly bounded with respect to μ and θ, which is rather weak condition under

(A4)-(i) since we consider a supremum taken over a compact set (and T will always be chosen as such in practice). We then have the following theorem, which states that the possible weak limits of $(\mu_n)_{n\in\mathbb{N}^\star}$ correspond to the global infimum of Ψ_α.

THEOREM 4. *Assume (A1) and (A4). Let $\alpha < 1$, $\kappa \leqslant 0$ and set $\Gamma(v) = [(\alpha-1)\,v+1]^{\eta/(1-\alpha)}$ for all $v \in \Delta_\alpha$. Then, for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, any $\eta > 0$ satisfies (5) and $\Psi_\alpha(\zeta) < \infty$.*

Let $\eta \in (0,1]$. Further assume that there exists $\mu_1, \mu^\star \in \mathrm{M}_1(\mathsf{T})$ such that the (well-defined) sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ defined by (4) weakly converges to $\mu^\star$ as $n \to \infty$. Then the following assertions hold

(i) $(\Psi_\alpha(\mu_n))_{n\in\mathbb{N}^\star}$ *is non-increasing,*
(ii) $\mu^\star$ *is a fixed point of* $\mathcal{I}_\alpha$*,*
(iii) $\Psi_\alpha(\mu^\star) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta)$.

The proof of Theorem 4 is deferred to Appendix A.3. Intuitively, we expect $\mu^\star$ to be a fixed point of $\mathcal{I}_\alpha$ based on Theorem 1. The core difficulty of the proof is then to prove Assertion (iii) and to do so, we proceed by contradiction: we assume there exists $\bar{\mu} \in \mathrm{M}_{1,\mu_1}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star) > \Psi_\alpha(\bar{\mu})$ and we contradict the fact that $(\mu_n)_{n\in\mathbb{N}^\star}$ converges to a fixed point.

The impact of Theorem 3 and Theorem 4 is twofold:

1. We proved an $O(1/N)$ convergence rate for the Entropic Mirror Descent under minimal assumptions. This is an improvement compared to standard Mirror Descent results, which under similar assumptions only provide an $O(1/\sqrt{N})$ convergence rate and assume a decaying learning rate (see [39, Theorem 4.2.]). Furthermore, while accelerated versions of the Mirror Descent (e.g Mirror Prox, [39, Theorem 4.4.]) also yield an $O(1/N)$ convergence rate, they require the objective function to be sufficiently smooth, an assumption that does not hold for many standard kernel families (e.g. Gaussian) for the α-divergence family.
2. We showed that we are able to go beyond the typical Entropic Mirror Descent framework by introducing the Power Descent for which $\Gamma(v) = [(\alpha-1)\,v+1]^{\frac{\eta}{1-\alpha}}$. In particular, we obtain that $(\Psi_\alpha(\mu_n))_{n\in\mathbb{N}^\star}$ decreases for any $\eta \in (0,1]$. When $\alpha < 1$, we have the convergence towards the optimal value $\Psi_\alpha(\mu^\star)$ and when $\alpha > 1$, we obtain an $O(1/N)$ convergence rate.

The results we obtained are summarized in Table 2 below.

TABLE 2
Examples of allowed (Γ, κ) *in the* (α, Γ)*-descent according to Theorem 3 and Theorem 4.*

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

4. Stochastic (α, Γ)-descent. As Algorithm 1 typically involves an intractable integral in the Expectation step, we now turn to a stochastic version of this algorithm. Let $M \in \mathbb{N}^\star$ and let $\mu \in \mathrm{M}_1(\mathsf{T})$.

Algorithm 2: *Stochastic* (α, Γ)*-descent one-step transition*

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

Algorithm 2 uses μk as a sampler instead of $k(\theta, \cdot)$. Indeed, as our algorithm optimises over μ, sampling with respect to μk gives preference to the interesting regions of the parameter space. Furthermore, picking a sampler that is independent from θ is less costly from a computational point of view. Given $\mu_1 \in \mathrm{M}_1(\mathsf{T})$, the stochastic version of the ideal iterative scheme defined by (4) is then given by

$$\mu_{n+1} = \hat{\mathcal{I}}_{\alpha,M}(\mu_n)\,, \qquad n \in \mathbb{N}^\star\,. \tag{21}$$

Let us now focus on the particular case of the Power Descent and the Entropic Mirror Descent.

Power Descent. Consider i.i.d random variables $Y_1, Y_2, ...$ with common density μk w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote by $\mathbb{E}$ the associated expectation operator. We are then able to establish the following result.

PROPOSITION 9. *Assume (A1). Let* $\alpha \in \mathbb{R} \setminus \{1\}$, $\eta > 0$, κ *be such that* $(\alpha - 1)\kappa \geqslant 0$ *and set* $\Gamma(v) = [(\alpha-1)\,v + 1]^{\eta/(1-\alpha)}$ *for all* $v \in \Delta_\alpha$. *Let*

$\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu) < \infty$, (5) holds and

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta)\mathbb{E}\left[\left\{\frac{k(\theta, Y_1)}{\mu k(Y_1)}\left(\frac{\mu k(Y_1)}{p(Y_1)}\right)^{\alpha-1} + (\alpha-1)\kappa\right\}^{\frac{\eta}{1-\alpha}}\right] < \infty\,. \tag{22}$$

Then,

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

The proof is deferred to Appendix B.4. The crux of the proof consists in applying a Dominated Convergence Theorem to non-negative real-valued $(\mathcal{T}\otimes\mathcal{F}, \mathcal{B}(\mathbb{R}_{\geqslant 0}))$-measurable functions, which requires to consider a Generalized version of the Dominated Convergence Theorem (Lemma 17) and an Integrated Law of Large Numbers (Lemma 18).

Entropic Mirror Descent. In the particular case of the Entropic Mirror descent, we obtain a stronger result. Indeed, it can be established that Algorithm 2 converges at an $O(1/\sqrt{N})$ rate in expectation. Denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space.

PROPOSITION 10. *Assume (A1). Let $\eta > 0$ and set $\Gamma(v) = e^{-\eta v}$ for all $v \in \Delta_\alpha$. Assume that for any $N \in \mathbb{N}^\star$, $\theta \in \mathsf{T}$,*

$$\hat{b}_{\mu_N,\alpha,M}(\theta) \leqslant \tilde{L}_N \text{ such that } \mathbb{E}[\tilde{L}_N^2] \leqslant \sigma^2. \tag{23}$$

Then, for any $N \in \mathbb{N}^\star$, taking $\eta := \frac{1}{\sigma}\sqrt{\frac{2KL(\mu^\star||\mu_1)}{N}}$ yields

$$\mathbb{E}\left[\Psi_\alpha\left(\frac{1}{N}\sum_{n=1}^N \mu_n\right) - \Psi_\alpha(\mu^\star)\right] \leqslant \sigma\sqrt{\frac{2KL(\mu^\star||\mu_1)}{N}}.$$

This result can be proven by adapting [39, Theorem 6.1.] to the Infinite-Dimensional case in the particular case of the negative entropy mirror map.

Mixture Models. We now address the case where μ_1 corresponds to a weighted sum of Dirac measures. This case is of particular interest to us since as we shall see, for any kernel K of our choice, the (α, Γ)-descent procedure simplifies and provides an update formula for the mixture weights of the corresponding mixture model $\mu_1 K$.

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

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

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

$$\begin{cases} \mu_1 = \mu_{\boldsymbol{\lambda}}, \\ \mu_{n+1} = \hat{\mathcal{I}}_{\alpha,M}(\mu_n)\ , \qquad n \in \mathbb{N}^\star\ , \end{cases}$$

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

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(\hat{b}_{\mu_n,\alpha,M}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(\hat{b}_{\mu_n,\alpha,M}(\theta_i) + \kappa)}\ , \tag{24}$$

with $Y_{1,n}, ..., Y_{M,n}$ drawn independently from $\mu_n k$ and

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M}\sum_{m=1}^M \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha\left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})}\right)\ .$$

In this particular framework, most of the computing effort at each step lies within the computation of the vector $(\hat{b}_{\mu_n,\alpha,M}(\theta_j))_{1\leqslant j\leqslant J}$. As it turns out, these computations can also be used to obtain an estimate of the Evidence Lower Bound (resp. the Renyi-Bound [19]) when $p(y) = p(y, \mathscr{D})$. Indeed, these two quantities, which assess the convergence of the algorithm and provide a bound on the log-likelihood (see [19, Theorem 1]) are defined the following way : for any variational density q, we have

$$\begin{aligned} \mathcal{L}_1(q; \mathscr{D}) &:= \int_{\mathsf{Y}} \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{L}_\alpha(q; \mathscr{D}) &:= \frac{1}{1-\alpha}\log\left(\int_{\mathsf{Y}} \left(\frac{p(y, \mathscr{D})}{q(y)}\right)^{1-\alpha} q(y)\nu(\mathrm{d}y)\right)\ , \end{aligned}$$

and we see that in our case, we can write

$$\begin{aligned} \mathcal{L}_1(\mu_n k, \mathscr{D}) &= -\sum_{j=1}^J \lambda_{j,n} b_{\mu_n,\alpha}(\theta_j) \\ \mathcal{L}_\alpha(\mu_n k, \mathscr{D}) &= \frac{1}{1-\alpha}\log\left((\alpha-1)\sum_{j=1}^J \lambda_{j,n} b_{\mu_n,\alpha}(\theta_j) + 1\right)\ . \end{aligned} \tag{25}$$

In addition, if there is a need for very large J, one can approximate the summation appearing in $\mu_{\boldsymbol{\lambda}} k$ using subsampling. Finally, we obtain Algorithm 3 below.

Algorithm 3: *Mixture Stochastic* (α, Γ)*-descent*

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

Set $\boldsymbol{\lambda} = [\lambda_{1,1}, ..., \lambda_{J,1}]$.
while *the bound has not converged* **do**
Sampling step : Draw independently M samples $Y_1, ..., Y_M$ from $\mu_{\boldsymbol{\lambda}} k$.
Expectation step : Compute $\boldsymbol{B}_{\boldsymbol{\lambda}} = (b_j)_{1 \leqslant j \leqslant J}$ where

$$b_j = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_m)}{\mu_{\boldsymbol{\lambda}} k(Y_m)} f'_\alpha \left(\frac{\mu_{\boldsymbol{\lambda}} k(Y_m)}{p(Y_m)} \right) \tag{26}$$

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

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

end

REMARK 11. *Note that we recover the mixture weights update rules from the Population Monte Carlo algorithm applied to reverse Kullback-Leibler minimisation [40] by considering the Power descent with* $\alpha = 0$ *and* $\eta = 1$. *We have thus embedded this special case into a more general framework.*

To summarise, we have several levels of generality in our algorithm: we are free to choose the kernel K, the α-divergence being optimised and we have identified couples (Γ, κ) which ensure the convergence of our algorithm.

An important remark is that Algorithm 3 does not require any information on how the $\{\theta_1, ..., \theta_J\}$ have been obtained in order to infer the optimal weights as it draws information from samples that are generated from $\mu_{\boldsymbol{\lambda}} k$. Since the algorithm leaves $\{\theta_1, ..., \theta_J\}$ unchanged throughout the optimisation of the mixture weights (we call it an *Exploitation Step*), a natural idea is to combine Algorithm 3 with an *Exploration step* that modifies the parameter set, which gives Algorithm 4 below.

Algorithm 4: *Complete Exploitation-Exploration Algorithm*

Input: p: measurable positive function, α: α-divergence parameter, (Γ, κ): chosen as per Table 1, q_0: initial sampler, K: Markov transition kernel, $(M_t)_t$: number of samples, $(J_t)_t$: dimension of parameter set.

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

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

while *not converged* **do**

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

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

end

Note that this algorithm is very general, as any Exploration Step can be envisioned. We now move on to numerical experiments in the next section.

5. Numerical experiments. In this part, we want to assess how Algorithm 4 performs on both toy and real-world examples. To do so, we first need to specify the kernel K and an algorithm for the Exploration Step.

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

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

Exploration Step. At time t, we resample among $\{\theta_{1,t}, ..., \theta_{J_t,t}\}$ according to the optimised mixture weights $\boldsymbol{\lambda}$. The obtained sample $\{\theta_{1,t+1}, ..., \theta_{J_{t+1},t+1}\}$ is then perturbed stochastically using the Gaussian transition kernel K_{h_t}, which gives us our new parameter set. The hyperparameter h_t is adjusted according to the number of particles so that $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space (the optimal rate in nonparametric estimation when the function is at least 2-times continuously differentiable and the kernel has order 2 [41]).

Next, we are interested in the choice of α. The hyperparameter α allows us to choose between *mass-covering* divergences which tend to cover all the

modes ($\alpha \ll 0$) and *mode-seeking* divergences that are attracted to the mode with the largest probability mass ($\alpha \gg 1$), the case $\alpha \in (0,1)$ corresponding to a mix of the two worlds (see Appendix C.1 for additional details on how these properties are expressed in the (α,Γ)-descent).

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

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

and as the dimension grows, the conditions of support are often not met in practice, meaning that there exists $A \in \mathcal{Y}$ such that $p(A, \mathscr{D}) = 0$ and $\mu k(A) > 0$. This implies that whenever $\alpha > 1$ we might have that $\hat{b}_{\mu,\alpha,M}(\theta) = \infty$ and that the α-divergence (or equivalently the Renyi-bound as defined in (25)) is infinite, which is the sort of behavior we would like to avoid. Thus, we restrict ourselves to the case $\alpha \leqslant 1$ in the following numerical experiments. Note that the limiting case $\alpha = 1$, corresponding to the commonly-used Forward Kullback-Leibler objective function, also suffers from this poor behavior, but is still considered in the experiments as a reference.

We now move on to our first example where we investigate the impact of different choices of Γ.

5.1. *Toy Example.* The target p is a mixture of two d-dimensional Gaussian densities multiplied by a positive constant Z such that

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

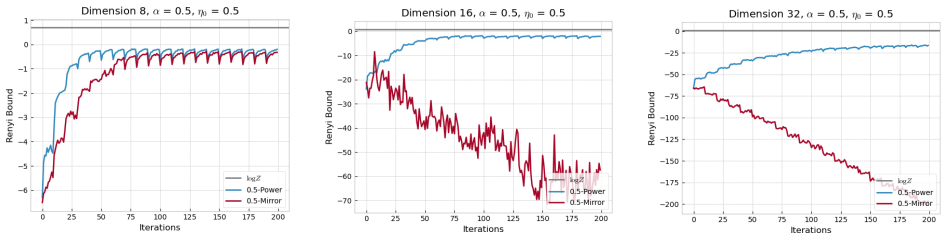
where $\boldsymbol{u_d}$ is the d-dimensional vector whose coordinates are all equal to 1, $s = 2$ and $Z = 2$. $(J_t)_t$ and (M_t) are kept constant equal to $J = M = 100$, $\kappa = 0$ and the initial weights are set to be $[1/J, ..., 1/J]$. The number of inner iterations in the (α,Γ)-descent is set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

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

For each of them, we run $T = 20$ iterations of Algorithm 4 and we replicate the experiment 100 times for $d = \{8, 16, 32\}$.

The results for the 0.5-Mirror and 0.5-Power descent are displayed on Figure 1. A first remark is that we are able to observe the monotonicity property from Theorem 2 (the Renyi-Bound varies like $\Psi_\alpha(\mu_n)^{\alpha-1}$) for the 0.5-Power descent, the jumps in the Renyi-Bound corresponding to an update of the parameter set. Furthermore, we see that the 0.5-Mirror descent (which would have been the default choice based on the existing optimisation literature) converges more slowly than the 0.5-Power descent in dimension 8. An even more striking aspect however is that, as the dimension grows, the 0.5-Mirror descent is unable to learn and the algorithm diverges.

Figure 1: Plotted is the average Renyi-Bound for the 0.5-Power and 0.5-Mirror descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.



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

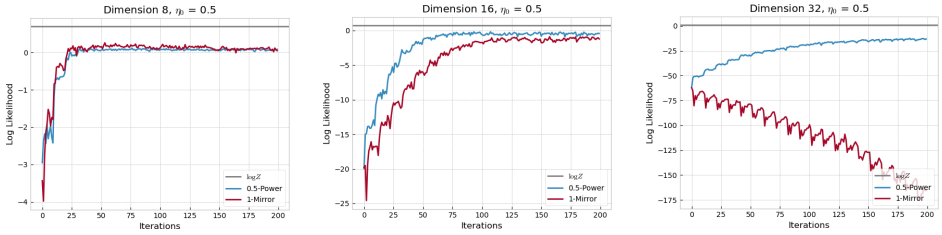
$$\text{Mirror}: \quad \lambda_{j,n} \propto \exp\left(\frac{\eta}{1-\alpha}\left((\alpha-1)b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j) + (\alpha-1)\kappa\right)\right)$$
$$\text{Power}: \quad \lambda_{j,n} \propto \exp\left(\frac{\eta}{1-\alpha}\log\left((\alpha-1)b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j) + (\alpha-1)\kappa\right)\right).$$

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

Finally, we compare how the 0.5-Power and 1-Mirror descent perform at approximating the log-likelihood in dimension $d = \{8, 16, 32\}$. The results are plotted on Figure 2. Again, the 0.5-Power descent comes across as faster

and more stable compared to the 1-Mirror descent as the dimension grows. Furthermore, it also does not fail in dimension 32, unlike the 1-Mirror descent.

Figure 2: Plotted is the average Log-likelihood for 0.5-Power and 1-Mirror descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.



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

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

5.2. *Bayesian Logistic Regression.* We consider Bayesian Logistic Regression for binary classification using the same setting as [42], which assigns the regression weights w with a Gaussian prior $p_0(w|\beta) = \mathcal{N}(w, \beta^{-1})$ and $p_0(\beta) = \text{Gamma}(\beta, 1, 0.01)$. The inference is applied on the posterior density $p(y|\mathscr{D})$ with $y = [w, \log \beta]$.

We test our algorithm for the *Covertype*[1] dataset (581,012 data points and 54 features). Computing $p(y, \mathscr{D})$ constitutes the major computation bottleneck here, since $p(y, \mathscr{D}) = p_0(y) \prod_i p(x_i|y)$ with a very large number of data points. We can conveniently address this problem by approximating $p(y, \mathscr{D})$ with subsampled mini-batches. We adopt this strategy here and consider mini-batches of size 100.

We set $\alpha = 0.5$, $N = 1$, $T = 500$, $\kappa = 0$, $J_0 = M_0 = 20$ and $J_{t+1} = M_{t+1} = J_t + 1$ for $t = 1...T$ in Algorithm 4. The initial weights in the (α, Γ)-descent are set to $\boldsymbol{\lambda}_{init,t} = [1/J_t, ..., 1/J_t]$ and the learning rate is set to $\eta_0 = 0.05$.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

One thing that is very specific to the Exploration step that we used to run our experiments (and sampling-based Exploration steps algorithms in general) is that the particles $\{\theta_{1,t}, ..., \theta_{J_t,t}\}$ are sampled from a known distribution at each Exploration step. This means that we are able to infer information on $\{\theta_{1,t}, ..., \theta_{J_t,t}\}$ using Importance Sampling (IS) weights. We thus compare the Power (α, Γ)-descent with a state-of-the-art Adaptive Importance Sampling-based (AIS) algorithm (see for example [43, 44, 45, 46]).

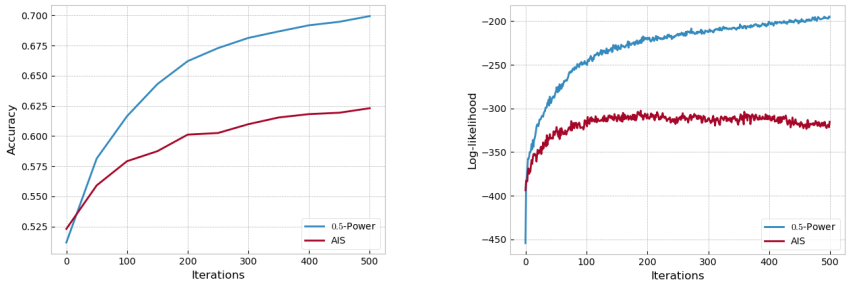
We initialise $\{\theta_{1,0}, ..., \theta_{J_0,0}\}$ by sampling J_0 points from the prior $p_0(y) = p_0(\beta)p_0(w|\beta)$ and set $q_0 = p_0$. Given q_t at time t, we draw J_t i.i.d samples $(\theta_{j,t})_{1 \leqslant j \leqslant J_t}$ from q_t and we define $q_{t+1}(y) = \sum_{j=1}^{J_t} \lambda_{j,t}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

$$\lambda_{j,t} \propto \begin{cases} \frac{p(\theta_{j,t}, \mathscr{D})}{q_t(\theta_{j,t})} & \text{(AIS)} , \\ \Gamma(\hat{b}_{\mu_{\boldsymbol{\lambda}_{init,t}}, \alpha, M}(\theta_{j,t}) + \kappa) & \text{(Power)} . \end{cases} \tag{27}$$

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

We replicate the experiments 100 times. The Accuracy and Log-likelihood averaged over the 100 trials for both algorithms are displayed on Figure 3 and we see that the 0.5-Power descent outperforms the AIS algorithm.

Figure 3: Plotted are the average Accuracy and Log-likelihood computed over 100 replicates for Bayesian Logistic Regression on the Covertype dataset for the 0.5-Power descent and the AIS algorithm.



6. Conclusion and perspectives. The (α, Γ)-descent is a novel gradient-based algorithm which operates on measures and leads at each step to a systematic decrease in the α-divergence for a rich family of values of Γ.

Our framework allows us to recover the Entropic MD with improved convergence rates and to introduce an alternative to the Entropic MD called the Power descent, which converges to an optimum with known convergence rates when $\alpha > 1$. Furthermore, our procedure provides a simple method to optimise the mixture weights of any given mixture model and can be applied without any information on the underlying distribution of the variational parameters. This renders our algorithm compatible with many choices of parameters updates and applicable to a wide range of Machine Learning tasks. We demonstrate empirically in the mixture case the benefit of going beyond the Entropic MD framework by using the Power descent algorithm instead, which is a more scalable alternative.

To conclude, we state several directions to extend our work on both a theoretical and a practical level.

Convergence rate. An interesting area of research consists in proving a convergence rate for a general Γ function satisfying (A2) in both the Exact and Stochastic cases as well as investigating the optimal rate policy for η.

Exploration Step. The (α, Γ)-descent allows us to extend the parameter set and to work with a population of particles $\{\theta_1, ..., \theta_J\}$ instead of just one particle θ. In this regard, many methods could be envisioned as an Exploration step and combined with the (α, Γ)-descent.

Variance Reduction. One may want to resort to more advanced Monte Carlo methods in the estimation of $b_{\mu_n,\alpha}$ for variance reduction purposes, such as reusing the past samples in the approximation of $b_{\mu_n,\alpha}$.

In an earlier version of this work, we were also able to establish conditions to obtain a systematic decrease in the α-divergence [48]. Note however that the proof of the monotonicity in [48, Theorem 1] is different from the one presented in this paper. As a result, the monotonicity and convergence results obtained in both papers do not strictly overlap (see in particular Theorem 1, 2 and 3 as well as Proposition 10), which also lead to different numerical experiments.

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] Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

[4] Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[5] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[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] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, August 2019.

[8] John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning*, page 1363–1370, Edinburgh, Scotland, UK, June 2012.

[9] 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.

[10] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. 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 324–333, New York, New York, USA, 20–22 Jun 2016. PMLR.

[11] Mingzhang Yin and Mingyuan Zhou. Semi-implicit 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 5660–5669, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

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

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

[24] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

[25] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

[26] Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. Stochastic expectation propagation. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2323–2331. Curran Associates, Inc., 2015.

[27] Guillaume Dehaene and Simon Barthelme. Expectation Propagation in the large-data limit. *Journal of the Royal Statistical Society: Series B*, 80(Part 1):197–217, 2017.

[28] David M. Blei, Andrew Y. Ng, and Michael Jordan. Latent dirichlet allocation. volume 3, pages 993–1022, 2003.

[29] Tetsuzo Morimoto. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar Tud. Akad. Mat. Kutat Int.*, page 85–108, 1963.

[30] Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

[31] Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed Nash equilibria of generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2810–2819, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[32] Tommi S. Jaakkola and Michael I. Jordan. Improving the mean field approximation via the use of mixture distributions. *Jordan M.I. (eds) Learning in Graphical Models. NATO ASI Series (Series D: Behavioural and Social Sciences), Springer*, 89, 1998.

[33] Samuel Gershman, Matthew D. Hoffman, and David M. Blei. Nonparametric variational inference. Edinburgh, Scotland, UK, 2012.

[34] 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.

[35] Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta di-

vergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, Jan 2011.

[36] Igal Sason. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, May 2018.

[37] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

[38] Bruce G. Lindsay. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Ann. Statist.*, 22(2):1081–1114, 06 1994.

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

[40] R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.*, 35(1):420–448, 02 2007.

[41] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 12 1982.

[42] Samuel Gershman, Matt Hoffman, and David Blei. Nonparametric variational inference. In *Proceedings of the 29 th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[43] Man-Suk Oh and James O Berger. Adaptive importance sampling in monte carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.

[44] Teun Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

[45] Nicolas Chopin. Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Ann. Statist.*, 32(6):2385–2411, 12 2004.

[46] Bernard Delyon and François Portier. Safe and adaptive importance sampling: a mixture approach, 2019. Available online: https://arxiv.org/abs/1903.08507 (accessed on 20 March 2020).

[47] H.L. Royden and P. Fitzpatrick. *Real Analysis (4th Edtion)*. Prentice Hall, 2010.

[48] Kamélia Daudel, Randal Douc, François Portier, and François Roueff. The f-divergence expectation iteration scheme, 2019. Available online: https://arxiv.org/abs/1909.12239 (accessed on 20 March 2020).

SUPPLEMENTARY MATERIAL

APPENDIX A

A.1. Proof of Lemma 6.

PROOF OF LEMMA 6. On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, consider the random variable $U(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ and let V be an independent copy of U. For all $u \in \Delta_\alpha$, define $\tilde{\Gamma}(u) = \Gamma(u)/\mathbb{E}[\Gamma]$. Let us now prove that

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

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

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

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

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

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

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

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

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

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

Combining with (7) yields (12).

□

A.2. Proof of Theorem 3. We start with a side note on Δ_α. A typical choice for Δ_α is

$$\Delta_\alpha = [-|b|_{\infty,\alpha} + \kappa, |b|_{\infty,\alpha} + \kappa] \,. \tag{28}$$

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

$$\Delta_\alpha = \begin{cases} [\frac{1}{1-\alpha} + \kappa, |b|_{\infty,\alpha} + \kappa], & \text{if } \alpha > 1 \\ [-|b|_{\infty,\alpha} + \kappa, \frac{1}{1-\alpha} + \kappa], & \text{if } \alpha < 1 \end{cases} \tag{29}$$

to underline the fact that for all $v \in \Delta_\alpha$, $(\alpha-1)v + 1 \geqslant (\alpha-1)\kappa$. Unless specified otherwise, we let Δ_α be as in (29) whenever $\alpha \in \mathbb{R} \setminus \{1\}$.

Proof of Theorem 3. The proof consists in verifying that we can apply Theorem 2 in each of the cases mentioned in Theorem 3. Let us recall the different conditions that must be met:

1. $0 < \inf_{v\in\Delta_\alpha} \Gamma(v)$ and $\sup_{v\in\Delta_\alpha} \Gamma(v) < \infty$.
2. The function $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

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

3. $L_{\alpha,1} = \inf_{v\in\Delta_\alpha} \{[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v)+1\} \inf_{v\in\Delta_\alpha} -\Gamma'(v) > 0$.
4. The function $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ is L-smooth and the function $-\log\Gamma$ is concave increasing.
5. $L_{\alpha,2} = (\inf_{v\in\Delta_\alpha}(-\log\Gamma)'(v))^{-1} > 0$.

(i) Kullback-Leibler divergence ($\alpha = 1$): $\Gamma(v) = e^{-\eta v}$, $\eta \in (0,1)$, any real κ. Since the update formula does not depend on κ, there is no constraint on κ and we assume that $\kappa = 0$ for simplicity.
- Condition 1 is satisfied since $|b|_{\infty,1}$ is finite.
- Condition 2 is satisfied with $\Gamma'(v) = -\eta e^{-\eta}$ and $(\log\Gamma)'(v) = -\eta$.
- Condition 3 is satisfied with $L_{1,1} \geqslant (1-\eta)\eta e^{-\eta|b|_{\infty,1}}$.
- Condition 4 is satisfied.
- Condition 5 is satisfied with $L_{1,2} = \frac{1}{\eta}$.

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

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

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

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

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

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

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

which is satisfied for $\eta \in (0,1]$.

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

□

A.3. Proof of Theorem 4. In the following, we use the notation $\mu_n \Rightarrow \mu^\star$ for the weak convergence of measures in $\mathrm{M}_1(\mathsf{T})$. For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, for all $\theta \in \mathsf{T}$, define

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

We first derive four useful lemmas.

LEMMA 12. *Assume (A1) and (A4). Suppose that $\mu_n \Rightarrow \mu^\star$. Then the following assertions hold.*

(i) For all $y \in \mathsf{Y}$, $\mu_n k(y)$ tends to $\mu^\star k(y)$ as $n \to \infty$.
(ii) For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, the function $\theta \mapsto g_\zeta(\theta)$ is continuous. Furthermore for all $\theta \in \mathsf{T}$, $g_{\mu_n}(\theta)$ tends to $g_{\mu^\star}(\theta)$ as $n \to \infty$.
(iii) There exist $0 < m_- < m_+ < \infty$ such that, for all $\zeta \in \mathrm{M}_1(\mathsf{T})$ and $\theta \in \mathsf{T}$, $g_\zeta(\theta) \in [m_-, m_+]$.
(iv) For all continuous, positive and bounded function h,

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

PROOF. We prove the assertions successively.

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

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

$$g_\zeta(\theta) = \int_{\mathsf{Y}} a_\zeta(\theta, y)\nu(\mathrm{d}y) \ ,$$

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

$$|a_\zeta(\theta, y)| \leqslant \sup_{\theta' \in \mathsf{T}} k(\theta', y) \times \sup_{\theta'' \in \mathsf{T}} \left(\frac{k(\theta'', y)}{p(y)}\right)^{\alpha-1} \ , \tag{30}$$

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

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

i.e $\lim_{n\to\infty} a_{\mu_n}(\theta, y) = a_{\mu^\star}(\theta, y)$. The bound (30) and (A4)-(iii) provide a domination criterion and we get that $g_{\mu_n}(\theta)$ tends to $g_{\mu^\star}(\theta)$ as $n \to \infty$, which concludes the proof of (ii).

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

$$\begin{aligned} m_- &:= \int_{\mathsf{Y}} \inf_{\theta' \in \mathsf{T}} k(\theta', y) \times \inf_{\theta'' \in \mathsf{T}} \left(\frac{k(\theta'', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) \ , \\ m_+ &:= \int_{\mathsf{Y}} \sup_{\theta' \in \mathsf{T}} k(\theta', y) \times \sup_{\theta'' \in \mathsf{T}} \left(\frac{k(\theta'', y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) \ . \end{aligned} \tag{31}$$

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

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

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta)\Gamma(b_{\mu^\star,\alpha}(\theta) + \kappa)h(\theta) = \int_{\mathsf{T}} \mu^\star(\mathrm{d}\theta)\Gamma(b_{\mu^\star,\alpha}(\theta) + \kappa)h(\theta) \ . \tag{32}$$

Next we show that

$$\lim_{n\to\infty} \int_{\mathsf{T}} \mu_n(\mathrm{d}\theta) \left|\Gamma(b_{\mu_n,\alpha}(\theta) + \kappa) - \Gamma(b_{\mu^\star,\alpha}(\theta) + \kappa)\right| h(\theta) = 0 \tag{33}$$

ie

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

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

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

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

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

which is integrable w.r.t ν by (A4)-(iii). Moreover, by (i) and by continuity of $u \mapsto u^{\alpha-1}$, we have $\lim_{n\to\infty} a_n(y) = 0$, and (33) follows by dominated convergence. Finally, combining (32), (33) and

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

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

LEMMA 13. *Assume (A1). Let $\mu^\star, \mu \in \mathrm{M}_1(\mathsf{T})$ and assume that there exists $\bar{\mu} \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi_\alpha(\bar{\mu}) < \Psi_\alpha(\mu^\star)$. Then, there exists $\delta > 1$ such that*

$$\bar{\mu}(g_{\mu^\star} > \delta \mu^\star(g_{\mu^\star})) > 0 . \tag{34}$$

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

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

that is

$$\int_{\mathsf{T}} [\zeta - \zeta'](\mathrm{d}\theta) g_{\zeta'}(\theta) \geqslant (\alpha - 1) \left(\Psi_\alpha(\zeta) - \Psi_\alpha(\zeta') \right) . \tag{35}$$

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

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

Using (35), we get that, for all $\delta > 1$,

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

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

LEMMA 14. *Assume (A1). Let $\mu^\star \in \mathrm{M}_1(\mathsf{T})$ be a fixed point of $\mathcal{I}_\alpha$ and let $\eta > 0$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ and assume that there exists $\bar{\mu} \in \mathrm{M}_{1,\mu}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star) > \Psi_\alpha(\bar{\mu})$. Then, there exists $\delta > 1$ such that*

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

PROOF. Note that (5) holds for any $\eta > 0$ and ζ (in particular $\zeta = \mu^\star$) by Lemma 12-(iii). As $\mu^\star$ is a fixed point of $\mathcal{I}_\alpha$, $g_{\mu^\star}$ is $\mu^\star$-almost all constant. Consequently, $\mu^\star(g_{\mu^\star})^{\eta/1-\alpha} = \mu^\star(g_{\mu^\star}^{\eta/1-\alpha}) = \mu^\star(\Gamma(b_{\mu^\star,\alpha} + \kappa))$. For all $\delta > 1$, $\delta' := \delta^{(1-\alpha)/\eta} > 1$ and

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

We conclude by applying Lemma 13. □

LEMMA 15. *Assume (A1) and (A4). Let $\eta > 0$, let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ and define the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ according to (4). Suppose that $\mu_n \Rightarrow \mu^\star$ for some fixed point $\mu^\star \in \mathrm{M}_1(\mathsf{T})$ of $\mathcal{I}_\alpha$. Further assume there exists $\bar{\mu} \in \mathrm{M}_{1,\mu_1}(\mathsf{T})$ such that $\Psi_\alpha(\mu^\star) > \Psi_\alpha(\bar{\mu})$. Then, there exist $\delta > 1$ and $n \in \mathbb{N}^*$ such that*

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

PROOF. First note that the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ is well-defined for any $\eta > 0$ by Lemma 12-(iii), which implies $\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa)) > 0$ for all $n \in \mathbb{N}^\star$. For all $\zeta \in \mathrm{M}_1(\mathsf{T})$, set $h_\zeta(\theta) = \Gamma(b_{\zeta,\alpha}(\theta) + \kappa)$. We further have that

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

Furthermore, applying (ii) and (iv) in Lemma 12, we have, for all $\theta \in \mathsf{T}$, $\lim_{n\to\infty} h_{\mu_n}(\theta) = h_{\mu^\star}(\theta)$ and $\lim_{n\to\infty} \mu_n(h_{\mu_n}) = \mu^\star(h_{\mu^\star})$. Hence, for all $\theta \in \mathsf{T}$,

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

The proof is concluded by applying Lemma 14. □

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

Lemma 12-(iii) is exactly the first result we want to obtain, that is: for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, any $\eta > 0$ satisfies (5) for ζ. Furthermore, $|\Psi_\alpha(\zeta)| < \infty$ by (A4)-(iii).

Assume that $(\mu_n)_{n\in\mathbb{N}^\star}$ weakly converges to $\mu^\star \in \mathrm{M}_1(\mathsf{T})$. First note that Lemma 12-(iii) implies that for any $\eta > 0$ the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ is well-defined and $\mu^\star$ satisfies (5). Using Theorem 1, we obtain that the sequence $(\mu_n)_{n\in\mathbb{N}^\star}$ is decreasing for all $\eta \in (0, 1]$, which gives Assertion (i).

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

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

$$\Psi_\alpha(\zeta) = \int_{\mathsf{Y}} a_\zeta(y)\nu(\mathrm{d}y) \, . \tag{36}$$

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

$$|a_\zeta(y)| \leqslant \left(\sup_{\theta\in\mathsf{T}} \left| f_\alpha\left(\frac{k(\theta, y)}{p(y)}\right)\right|\right) p(y) \, , \tag{37}$$

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

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

By applying twice Lemma 12-(iv) with $h(\theta) = 1$ and $h(\theta) = k(\theta, y)$, we have that for all $y \in \mathsf{Y}$,

$$\lim_{n\to\infty} [\mathcal{I}_\alpha(\mu_n)k](y) = [\mathcal{I}_\alpha(\mu^\star)k](y) \, . \tag{38}$$

Now, since f_α is C^1, we obtain from Lemma 12-(i) and (38) respectively that for all $y \in \mathsf{Y}$, $\lim_{n\to\infty} a_{\mu_n}(y) = a_{\mu^\star}(y)$ and $\lim_{n\to\infty} a_{\mathcal{I}_\alpha(\mu_n)}(y) = a_{\mathcal{I}_\alpha(\mu^\star)}(y)$. Combining with (37) and (36) we can thus apply the Dominated Convergence Theorem to obtain

$$\lim_{n\to\infty} \Psi_\alpha(\mu_n) = \Psi_\alpha(\mu^\star) \tag{39}$$

and

$$\lim_{n\to\infty} \Psi_\alpha(\mu_{n+1}) = \lim_{n\to\infty} \Psi_\alpha(\mathcal{I}_\alpha(\mu_n)) = \Psi_\alpha(\mathcal{I}_\alpha(\mu^\star)) \; . \tag{40}$$

Finally, (39) and (40) together yield $\Psi_\alpha(\mu^\star) = \Psi_\alpha \circ \mathcal{I}_\alpha(\mu^\star)$, which in turn implies that $\mu^\star$ is a fixed point of $\mathcal{I}_\alpha$ according to Theorem 1-(ii).

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

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

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

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

where for all $\zeta \in \mathrm{M}_1(\mathsf{T})$, for all $\theta \in \mathsf{T}$, $h_\zeta(\theta) := \Gamma(b_{\zeta,\alpha}(\theta) + \kappa)$. There exists, according to Lemma 15, for a well chosen $\delta > 1$ and a sufficiently large n_0 such that $\bar{\mu}(B_{n_0}) > 0$.

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

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

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

APPENDIX B

B.1. Lemma 16 : statement and proof. Recall that $Y_1, Y_2, ...$ are i.i.d random variables with common density μk w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we denote by $\mathbb{E}$ the associated expectation operator.

LEMMA 16. *Assume (A1). Let $\alpha \in \mathbb{R} \setminus \{1\}$, $\eta > 0$ and κ be such that $(\alpha - 1)\kappa \geqslant 0$. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\mu(|b_{\mu,\alpha}|) < \infty$ and*

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}\left[\left\{\frac{k(\theta, Y_1)}{\mu k(Y_1)}\left(\frac{\mu k(Y_1)}{p(Y_1)}\right)^{\alpha-1} + (\alpha-1)\kappa\right\}^{\frac{\eta}{1-\alpha}}\right] < \infty \,. \tag{41}$$

Then,

$$\lim_{M\to\infty} \mu(\Gamma(\hat{b}_{\mu,\alpha,M} + \kappa)) = \mu(\Gamma(b_{\mu,\alpha} + \kappa)) \,, \quad \mathbb{P} - \text{a.s.} \tag{42}$$

PROOF. Set $g(\theta, y) = \frac{k(\theta,y)}{\mu k(y)}(\frac{\mu k(y)}{p(y)})^{\alpha-1} + (\alpha-1)\kappa$, $\phi = \frac{\eta}{1-\alpha}$ and $h(u) = (\alpha-1)u + (\alpha-1)\kappa + 1$. Note that $\mathbb{E}[g(\theta, Y_1)] = h(b_{\mu,\alpha}(\theta))$ and $h^\phi = \Gamma$.

(i) We start with the case $\phi \notin [0, 1]$. Our goal is to apply Lemma 17, which is a generalized version of the Dominated Convergence Theorem. To do so, first note that $h(\hat{b}_{\mu,\alpha,M}(\theta))^\phi$ is positive and combining with the convexity of the mapping $u \mapsto u^\phi$, we have for all $M \in \mathbb{N}^\star$ and for all $\theta \in \mathsf{T}$,

$$0 \leqslant h(\hat{b}_{\mu,\alpha,M}(\theta))^\phi \leqslant M^{-1} \sum_{m=1}^{M} [g(\theta, Y_m)]^\phi \,. \tag{43}$$

Since $\mu(|b_{\mu,\alpha}|) < \infty$, the LLN for μ-almost all $\theta \in \mathsf{T}$ yields

$$\lim_{M\to\infty} \hat{b}_{\mu,\alpha,M}(\theta) = b_{\mu,\alpha}(\theta) \,. \tag{44}$$

Now applying successively (a) the LLN for μ-almost all $\theta \in \mathsf{T}$ (as stated in Lemma 18), which is valid under (41), (b) Fubini's Theorem and (c) again the LLN

$$\begin{aligned} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} \{g(\theta, Y_m)\}^\phi &\overset{(a)}{=} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}\left[\{g(\theta, Y_1)\}^\phi\right] \\ &\overset{(b)}{=} \mathbb{E}\left[\int_{\mathsf{T}} \mu(\mathrm{d}\theta)[g(\theta, Y_1)]^\phi\right] \overset{(c)}{=} \lim_{M\to\infty} \int_{\mathsf{T}} \mu(\mathrm{d}\theta) M^{-1} \sum_{m=1}^{M} [g(\theta, Y_m)]^\phi \end{aligned} \tag{45}$$

That is

$$\mu\left(\lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} \{g(\cdot, Y_m)\}^\phi\right) = \lim_{M\to\infty} \mu\left(M^{-1} \sum_{m=1}^{M} [g(\cdot, Y_m)]^\phi\right) < \infty$$

Combining with (43) and (44), we apply Lemma 17 and obtain

$$\mu\left(h(b_{\mu,\alpha})^{\phi}\right) = \mu\left(\lim_{M\to\infty} h(\hat{b}_{\mu,\alpha,M})^{\phi}\right) = \lim_{M\to\infty} \mu(h(\hat{b}_{\mu,\alpha,M})^{\phi}) \ ,$$

that is

$$\mu\left(\Gamma(b_{\mu,\alpha}+\kappa)\right) = \lim_{M\to\infty} \mu(\Gamma(\hat{b}_{\mu,\alpha,M}+\kappa)) \ .$$

(ii) We now turn to the case $\phi \in (0,1]$. Let $M' > 0$. Since

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(M^{-1}\sum_{m=1}^{M} g(\theta,Y_m)\mathbf{1}_{\{g(\theta,Y_m)\leqslant M'\}}\right)^{\phi} \leqslant \mu(h(\hat{b}_{\mu,\alpha,M})^{\phi}) \ ,$$

the LLN for μ-almost all $\theta \in \mathsf{T}$ (Lemma 18) and the Dominated Convergence Theorem yields

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\mathbb{E}[g(\theta,Y_1)\mathbf{1}_{\{g(\theta,Y_1)\leqslant M'\}}]\right)^{\phi} \leqslant \liminf_{M\to\infty} \mu(h(\hat{b}_{\mu,\alpha,M})^{\phi}) \ . \tag{46}$$

Using now $(u+v)^{\phi} \leqslant u^{\phi} + v^{\phi}$ and then Jensen's inequality for the concave mapping $u \mapsto u^{\phi}$,

$$\begin{aligned}\mu(h(\hat{b}_{\mu,\alpha,M})^{\phi}) \leqslant \int_{\mathsf{T}} \mu(\mathrm{d}\theta) &\left(M^{-1}\sum_{m=1}^{M} g(\theta,Y_m)\mathbf{1}_{\{g(\theta,Y_m)\leqslant M'\}}\right)^{\phi} \\ &+ \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta) M^{-1}\sum_{m=1}^{M} g(\theta,Y_m)\mathbf{1}_{\{g(\theta,Y_m)> M'\}}\right)^{\phi}\end{aligned}$$

By invoking the LLN for μ-almost all $\theta \in \mathsf{T}$ (Lemma 18) and the Dominated Convergence Theorem for the first term of the rhs and the LLN combined with Fubini for the second term, we get

$$\begin{aligned}\limsup_{M\to\infty} \mu(h(\hat{b}_{\mu,\alpha,M})^{\phi}) \leqslant \int_{\mathsf{T}} \mu(\mathrm{d}\theta) &\left(\mathbb{E}[g(\theta,Y_1)\mathbf{1}_{\{g(\theta,Y_1)\leqslant M'\}}]\right)^{\phi} \\ &+ \left(\int_{\mathsf{T}} \mu(\mathrm{d}\theta)\mathbb{E}[g(\theta,Y_1)\mathbf{1}_{\{g(\theta,Y_1)> M'\}}]\right)^{\phi}\end{aligned}$$

Letting M' go to infinity both in this inequality and in (46) completes the proof of (42).

□

B.2. General Dominated Convergence Theorem. We state and prove a generalized version of the Dominated Convergence Theorem, adapted from [47, Theorem 19]. We provide here a full proof for the sake of completeness.

LEMMA 17 (General Dominated Convergence Theorem). *Let* $\zeta \in \mathrm{M}_1(\mathsf{T})$. *Assume there exist* (a_M), (b_M), (c_M) *three sequences of* $(\mathcal{T}, \mathcal{B}(\mathbb{R}))$*-measurable functions such that the limits* $\lim_{M\to\infty} a_M(\theta)$, $\lim_{M\to\infty} b_M(\theta)$, $\lim_{M\to\infty} c_M(\theta)$ *exist for* ζ*-almost all* $\theta \in \mathsf{T}$ *and*

$$\zeta|\lim_{M\to\infty} a_M| + \zeta|\lim_{M\to\infty} c_M| < \infty \,.$$

Assume moreover that for all $M \in \mathbb{N}^\star$ *and for* ζ*-almost all* $\theta \in \mathsf{T}$

$$a_M(\theta) \leqslant b_M(\theta) \leqslant c_M(\theta)$$

and

$$\zeta(\lim_{M\to\infty} a_M) = \lim_{M\to\infty} \zeta(a_M) \tag{47}$$

$$\zeta(\lim_{M\to\infty} c_M) = \lim_{M\to\infty} \zeta(c_M) \,. \tag{48}$$

Then,

$$\zeta(\lim_{M\to\infty} b_M) = \lim_{M\to\infty} \zeta(b_M) \,.$$

PROOF. We apply Fatou's Lemma combined with (47) and (48) to the two non-negative, $(\mathcal{T}, \mathcal{B}(\mathbb{R}))$-measurable functions $\theta \mapsto b_M(\theta) - a_M(\theta)$ and $\theta \mapsto c_M(\theta) - b_M(\theta)$ and we obtain

$$\zeta(\liminf_{M\to\infty} b_M) \leqslant \liminf_{M\to\infty} \zeta(b_M)$$
$$\zeta(\liminf_{M\to\infty} -b_M) \leqslant \liminf_{M\to\infty} \zeta(-b_M)$$

which proves the lemma, as $\liminf_{M\to\infty} b_M(\theta) = \limsup_{M\to\infty} b_M(\theta)$ for ζ-almost all $\theta \in \mathsf{T}$. □

B.3. Integrated Law of Large Numbers. Let $Y_1, Y_2, \ldots$ be i.i.d. random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let f be a non-negative real-valued $(\mathcal{T} \otimes \mathcal{F}, \mathcal{B}(\mathbb{R}_{\geqslant 0}))$-measurable function. We are interested in showing

$$\int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1} \sum_{m=1}^{M} f(\theta, Y_m) = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \mathbb{E}[f(\theta, Y_1)] \tag{49}$$

for $\zeta \in \mathrm{M}_1(\mathsf{T})$ satisfying $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] < \infty$. While this result follows easily if we can show that

$$\mathbb{P}\left(\forall \theta \in \mathsf{T},\ \lim_{M\to\infty} M^{-1}\sum_{m=1}^{M} f(\theta, Y_m) = \mathbb{E}[f(\theta, Y_1)]\right) = 1 \tag{50}$$

unfortunately the LLN only yields

$$\mathbb{P}\left(\lim_{M\to\infty} M^{-1}\sum_{m=1}^{M} f(\theta, Y_m) = \mathbb{E}[f(\theta, Y_1)]\right) = 1$$

for ζ-almost all $\theta \in \mathsf{T}$. The following lemma allows to show (49) without resorting to the much stronger identity (50).

LEMMA 18. *Let $\zeta \in \mathrm{M}_1(\mathsf{T})$ and assume that $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] < \infty$. Then, $\mathbb{P}-$*a.s.

$$\int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1}\sum_{m=1}^{M} f(\theta, Y_m) = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)]\ .$$

PROOF. Set

$$B = \left\{(\theta,\omega) \in \mathsf{T}\times\Omega\ :\ \lim_{M\to\infty} M^{-1}\sum_{m=1}^{M} f(\theta, Y_m(\omega)) = \mathbb{E}[f(\theta, Y_1)]\right\}\ ,$$

Let $\gamma_0 : (\theta,\omega) \mapsto \mathbf{1}_{B^c}(\theta,\omega)$ and $\gamma_1 = 1-\gamma_0$. According to the Fubini Theorem and the LLN for $M^{-1}\sum_{m=1}^{M} f(\theta, Y_m)$ where θ is such that $\mathbb{E}[f(\theta, Y_1)] < \infty$ (which is satisfied for ζ-almost all $\theta \in \mathsf{T}$ by assumption),

$$\mathbb{E}\left[\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\gamma_0(\theta,\cdot)\right] = \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}\left[\gamma_0(\theta,\cdot)\right] = 0\ .$$

Therefore, $\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\gamma_0(\theta,\cdot)$ is $\mathbb{P}-$a.s. null that is, there exists Ω_1 such that $\mathbb{P}(\Omega_1) = 1$ and for all $\omega \in \Omega_1$, $A \mapsto \int_A \zeta(\mathrm{d}\theta)\gamma_0(\theta,\omega)$ is the null-measure on $(\mathsf{T}, \mathcal{T})$, which in turn implies that the measures ζ and $A \mapsto \int_A \zeta(\mathrm{d}\theta)\gamma_1(\theta,\omega)$ coincide. The latter property implies for all $\omega \in \Omega_1$,

$$\begin{aligned}
\int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)] &= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\mathbb{E}[f(\theta, Y_1)]\gamma_1(\theta,\omega) \\
&= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta)\left[\lim_{M\to\infty} M^{-1}\sum_{m=1}^{M} f(\theta, Y_m(\omega))\right]\gamma_1(\theta,\omega) \\
&= \int_{\mathsf{T}} \zeta(\mathrm{d}\theta) \lim_{M\to\infty} M^{-1}\sum_{m=1}^{M} f(\theta, Y_m(\omega))\ .
\end{aligned}$$

□

B.4. Proof of Proposition 9.

Proof of Proposition 9. For the sake of readability, we only treat the case $\kappa = 0$ in the proof of Proposition 9. Note that the case $\kappa \neq 0$ unfolds similarly by replacing $b_{\mu,\alpha}$ by $b_{\mu,\alpha} + \kappa$ everywhere in the proof below. Recall that $\Psi_\alpha(\mu) < \infty$ implies $\mu(|b_{\mu,\alpha}|) < \infty$ (see Remark 8). By the triangular inequality, for all $M \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,

$$\left|\frac{\Gamma(\hat{b}_{\mu,\alpha,M}(\theta))}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))} - \frac{\Gamma(b_{\mu,\alpha}(\theta))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| \leqslant \frac{\Gamma(\hat{b}_{\mu,\alpha,M}(\theta))}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}\left|1 - \frac{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| + \frac{|\Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) - \Gamma(b_{\mu,\alpha}(\theta))|}{\mu(\Gamma(b_{\mu,\alpha}))}$$

Thus,

$$\begin{aligned}\left\|\hat{\mathcal{I}}_{\alpha,M}(\mu) - \mathcal{I}_\alpha(\mu)\right\|_{TV} &= \mu\left(\left|\frac{\Gamma(\hat{b}_{\mu,\alpha,M})}{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))} - \frac{\Gamma(b_{\mu,\alpha})}{\mu(\Gamma(b_{\mu,\alpha}))}\right|\right) \\ &\leqslant \left|1 - \frac{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| + \frac{\mu(|\Gamma(\hat{b}_{\mu,\alpha,M}) - \Gamma(b_{\mu,\alpha})|)}{\mu(\Gamma(b_{\mu,\alpha}))}\end{aligned}$$

For the first term of the rhs, Lemma 16 yields

$$\lim_{M\to\infty}\left|1 - \frac{\mu(\Gamma(\hat{b}_{\mu,\alpha,M}))}{\mu(\Gamma(b_{\mu,\alpha}))}\right| = 0 \tag{51}$$

As for the second term of the rhs, first note that for all $M \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$

$$0 \leqslant |\Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) - \Gamma(b_{\mu,\alpha}(\theta))| \leqslant \Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) + \Gamma(b_{\mu,\alpha}(\theta)) \,, \tag{52}$$

and since $\mu(\Gamma(b_{\mu,\alpha})) < \infty$ the LLN for μ-almost all $\theta \in \mathsf{T}$ yields

$$\lim_{M\to\infty} \Gamma(\hat{b}_{\mu,\alpha,M}(\theta)) = \Gamma(b_{\mu,\alpha}(\theta)) \,. \tag{53}$$

Furthermore, since $\mu(\Gamma(b_{\mu,\alpha})) < \infty$, Lemma 16 and (53) imply

$$\lim_{M\to\infty} \mu\left[\Gamma(\hat{b}_{\mu,\alpha,M}) + \Gamma(b_{\mu,\alpha})\right] = \mu\left[\lim_{M\to\infty}\left(\Gamma(\hat{b}_{\mu,\alpha,M}) + \Gamma(b_{\mu,\alpha})\right)\right] < \infty$$

Combining with (52) and (53), we apply Lemma 17 and obtain

$$\lim_{M\to\infty} \frac{\mu(|\Gamma(\hat{b}_{\mu,\alpha,M}) - \Gamma(b_{\mu,\alpha})|)}{\mu(\Gamma(b_{\mu,\alpha}))} = 0$$

which, along with (51), finishes the proof. □

APPENDIX C

C.1. Mass-covering/Mode-seeking behavior in the Mixture Approximate (α, Γ)-descent. Let us first recall the update formula (24) used in the Mixture Stochastic (α, Γ)-descent. We have

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

with $Y_{1,n}, ..., Y_{M,n}$ drawn independently from $\mu_n k$ and

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha\left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})}\right) .$$

We illustrate the mode-seeking and mass-covering properties in the particular case of the Mirror descent ($\alpha = 1$) and the Power descent.

- Mirror descent : $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$. Then,

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} \log\left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})}\right)$$

and

$$\lambda_{j,n+1} \propto \prod_{m=1}^{M} \left(\frac{p(Y_{m,n})}{\mu_n k(Y_{m,n})}\right)^{\frac{\eta}{M}\frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})}} .$$

Observe that if $p(Y_{m,n}) = 0$ with $k(\theta_j, Y_{m,n}) > 0$ for at least one $m = 1...M$, then the weight is set to 0, which is the mode-seeking behavior. Note that this behavior might prevent learning in practice.

- Power descent: $\Gamma(v) = ((\alpha - 1)v + 1)^{\eta/(1-\alpha)}$. Then,

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M(\alpha - 1)}\sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} \left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})}\right)^{\alpha-1} - \frac{1}{\alpha - 1}$$

and

$$\lambda_{j,n+1} \propto \left(\frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} \left(\frac{p(Y_{m,n})}{\mu_n k(Y_{m,n})}\right)^{1-\alpha} + \kappa(\alpha - 1)\right)^{\frac{\eta}{1-\alpha}} .$$

Here again we observe the mode-seeking behavior when $\alpha > 1$ and the mass-covering behavior when $\alpha < 1$ by considering the case where $p(Y_{m,n}) = 0$ with $k(\theta_j, Y_{m,n}) > 0$ for at least one $m = 1...M$.

LTCI, Télécom Paris
Institut Polytechnique de Paris
19 Place Marguerite Perey, 91120 Palaiseau
E-mail: kamelia.daudel@telecom-paris.fr
francois.portier@telecom-paris.fr

SAMOVAR, Télécom SudParis
Institut Polytechnique de Paris
9 rue Charles Fourier, 91000 Evry
E-mail: randal.douc@telecom-sudparis.eu