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

Kamélia Daudel, Randal Douc, François Portier

The Annals of Statistics
2021, Vol. 49, No. 4, 2250–2270
https://doi.org/10.1214/20-AOS2035

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

BY KAMÉLIA DAUDEL[1,*], RANDAL DOUC[2] AND FRANÇOIS PORTIER[1,†]

[1]*LTCI, Télécom Paris, Institut Polytechnique de Paris,* [*]*kamelia.daudel@gmail.com;* [†]*francois.portier@telecom-paris.fr*
[2]*SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, randal.douc@telecom-sudparis.eu*

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 and derive convergence results. Our framework recovers the Entropic Mirror Descent algorithm and provides an alternative algorithm that we call the Power Descent. Moreover, in its stochastic formulation, 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 Mirror Descent 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) [2, 23] and Expectation Propagation (EP) [31, 37] consider this objective purely as an optimisation problem (which is often nonconvex). 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 minimises a certain objective function, typically the Kullback–Leibler divergence [25] between the posterior and the variational density.

Modern variational methods improved in three major directions [4, 48] (i) Black-Box inference techniques [38, 39] and Hierarchical Variational Inference methods [40, 47] have been deployed, expanding the variational family and rendering Variational methods applicable to a wide range of models (ii) Algorithms based on alternative families of divergences such as the α-divergence [49, 50] and Renyi's α-divergence [41, 45] have been introduced [1, 15, 19, 27, 29, 30, 46] to bypass practical issues linked to the Kullback–Leibler divergence [4, 20, 31] (iii) Scalable methods relying on stochastic optimisation techniques [6, 42] have been developed to enable large-scale learning and have been applied to complex probabilistic models [5, 13, 20, 26].

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 considering

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

Received May 2020; revised October 2020.
MSC2020 subject classifications. Primary 62F15; secondary 62F30, 62F35, 62G07, 62L99.
Key words and phrases. Variational inference, alpha-divergence, Kullback–Leibler divergence, Mirror Descent.

which is a more general form compared to the one found in [47] where μ is parametrised by another parametric model. As for the objective function, we work within the α-divergence family, which admits the forward Kullback–Leibler and the reverse Kullback–Leibler as limiting cases. These divergences belong to the f-divergence family [32, 33] 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 [21], Appendix A, is included in our framework and we obtain an $O(1/N)$ convergence rate under minimal assumptions, which improves on existing results and illustrates the generality of our approach. We also introduce a novel algorithm called the Power Descent, for which we prove convergence to an optimum and obtain an $O(1/N)$ convergence rate when $\alpha > 1$.
- In Section 4, we define the Stochastic version of the Exact (α, Γ)-descent and apply it to the important case of mixture models [17, 22]. The resulting general-purpose algorithm is Black-Box and does not require any information on the underlying distribution of the variational parameters. This algorithm notably enjoys an $O(1/\sqrt{N})$ convergence rate in the particular case of the Entropic Mirror Descent if we know the stopping time of the algorithm (Theorem 5).
- Finally, Section 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 computationally equivalent Adaptive Importance Sampling algorithm for Bayesian Logistic Regression on a large dataset.

Apart from the proofs leading to Theorem 1, which is central to our approach and is used to derive several subsequent results, we have deferred all the proofs to the Supplementary Material, alongside with additional comments.

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, that is, $\mathbb{Q} \preceq \nu$, $\mathbb{P} \preceq \nu$. 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

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

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

wherever they are well defined (and otherwise we write $+\infty$).

As $\lim_{\alpha\to 0} D_\alpha(\mathbb{Q}\,\|\,\mathbb{P}) = D_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q})$ and $\lim_{\alpha\to 1} D_\alpha(\mathbb{Q}\,\|\,\mathbb{P}) = D_{\mathrm{KL}}(\mathbb{Q}\,\|\,\mathbb{P})$ (see, e.g., [45]), the definition of the α-divergence can be extended to 0 and 1 by continuity and we will use the notation $D_0(\mathbb{Q}\,\|\,\mathbb{P}) = D_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q})$ and $D_1(\mathbb{Q}\,\|\,\mathbb{P}) = D_{\mathrm{KL}}(\mathbb{Q}\,\|\,\mathbb{P})$ 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 $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}[u^\alpha - 1 - \alpha(u-1)]$ 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 right-hand side 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 [10, 11, 43, 45] for more details on the α-divergence family.

PROPOSITION 2. *The α-divergence is always nonnegative 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 [18, 28] and the χ^2-divergence [15] 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 example of such a model is Bayesian logistic regression for binary classification.

EXAMPLE 1 (Bayesian logistic regression). We use the same setting as in [17]. We observe the data $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$ which is made of I binary class labels, $c_i \in \{-1, 1\}$, and of L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$. The hidden variables $y = \{\boldsymbol{w}, \beta\}$ consist of L regression coefficients $w_l \in \mathbb{R}$, and a precision parameter $\beta \in \mathbb{R}^+$. We assume the following model:

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

where a and b are hyperparameters (shape and inverse scale, resp.) that we assume to be fixed. We thus have $p(y,\mathscr{D}) = p_0(y)\prod_{i=1}^I p(c_i|\boldsymbol{x}_i, y)$ with $p_0(y) = \prod_{l=1}^L p_0(w_l|\beta)p_0(\beta)$ and as the sigmoid does not admit a conjugate exponential prior, $p(\mathscr{D})$ is intractable in this model.

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

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

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:

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

which does not involve the marginal likelihood $p(\mathscr{D})$ anymore (see, e.g., [4] and [15, 27]). 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 minimisation 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 minimising 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 $\{\theta_1, \ldots, \theta_J\} \in \mathsf{T}^J$ 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 $\Psi_\alpha(\mu)$, that is, $\Psi_\alpha(\mu) = \Psi_\alpha(\mu; p, q, \nu)$, 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:

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

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*

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

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 $\Psi_\alpha(\mu) \geq \tilde{f}_\alpha(\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)) > -\infty$. □

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 nonincreasing 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 in Algorithm 1.

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

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

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

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

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

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

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 [21], 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 probability measures.

In the rest of the section, we investigate some core properties of the aforementioned sequence of probability measures $(\mu_n)_{n\in\mathbb{N}^\star}$. We start by establishing conditions on (Γ, κ) such that the (α, Γ)-descent diminishes $\Psi_\alpha(\mu_n)$ 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:

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

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

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*

$$A_\alpha \le \Psi_\alpha(\mu) - \Psi_\alpha(\zeta). \tag{7}$$

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

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

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 $y \in \mathsf{Y}$, $f_\alpha(v_y) \ge f_\alpha(u_{\theta,y}) + f'_\alpha(u_{\theta,y})(v_y - u_{\theta,y})$, that is,

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

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

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

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 $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leq \Psi_\alpha(\mu)$. 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 $A_\alpha \geq 0$. For this purpose, let us denote by Dom_α an interval of $\mathbb{R}$ such that for all $\theta \in \mathsf{T}$, for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $b_{\mu,\alpha}(\theta) + \kappa$ and $\mu(b_{\mu,\alpha}) + \kappa \in \mathrm{Dom}_\alpha$ and let us make an assumption on (Γ, κ).

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

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

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* $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leq \Psi_\alpha(\mu)$.
(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 $A_\alpha \geq 0$. Then the proof is concluded by setting $\zeta = \mathcal{I}_\alpha(\mu)$ in Lemma 5 as

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

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

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

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

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

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

(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

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

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

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

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 Dom_α which implies $A_\alpha \geq 0$.

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 $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leq \Psi_\alpha(\mu)$ 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 forward Kullback–Leibler divergence as objective function is included in this framework.

One can also readily check that $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ satisfies (A2) for all $\alpha \in \mathbb{R} \setminus \{1\}$, for all κ such that $(\alpha - 1)\kappa \geq 0$ 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.

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*

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

where

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

The proof of Lemma 6 builds on the proof of Theorem 1 and can be found in [12], 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.

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 $\alpha \in \mathbb{R} \setminus \{1\}$	$\Gamma(v) = [(\alpha - 1)v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$(\alpha - 1)\kappa \geq 0$

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, e.g., [7], equation (3.5)) that

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

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

3.3. *Convergence.* Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$. We want to study the limiting behaviour 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:

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

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

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

We are now able to derive our second main result.

THEOREM 2. *Assume* (A1), (A2) *and* (A3). *Further assume that* $L_{\alpha,1}$, $L_{\alpha,2} > 0$ *and that* $0 < \inf_{v\in\mathrm{Dom}_\alpha}\Gamma(v) \leq \sup_{v\in\mathrm{Dom}_\alpha}\Gamma(v) < \infty$. *Moreover, let* $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ *be such that* $\Psi_\alpha(\mu_1) < \infty$. *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 nonincreasing*.

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

$$\Psi_\alpha(\mu_N) - \Psi_\alpha(\mu^\star) \leq \frac{L_{\alpha,2}}{N}\left[\mathrm{KL}(\mu^\star \parallel \mu_1) + L\frac{L_{\alpha,3}}{L_{\alpha,1}}\Delta_1\right], \tag{13}$$

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* $\mathrm{KL}(\mu^\star \parallel \mu_1) = \int_{\mathsf{T}} \log(\frac{\mathrm{d}\mu^\star}{\mathrm{d}\mu_1})\,\mathrm{d}\mu^\star$.

The proof of Theorem 2, which as hinted previously brings into play Lemma 6, is deferred to [12], Appendix A.2. We now wish to comment on the constants appearing in (13) and in particular the two constants $\mathrm{KL}(\mu^\star \parallel \mu_1)$ and Δ_1 (since the remaining constants $L_{\alpha,1}$, $L_{\alpha,2}$, $L_{\alpha,3}$ and L all involve the function Γ, which has not been chosen yet in Theorem 2).

To do so, we consider in Example 2 the finite-dimensional case where μ_1 is a weighted sum of dirac measures. As we shall explain in more detail later on in Section 4, this case is of particular relevance to us as our procedure can then be used to optimise the mixture weights of any given mixture model.

EXAMPLE 2 (Simplex framework). Let $J \in \mathbb{N}^\star$, let $(\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$ and let us consider $\mu_1 = J^{-1}\sum_{j=1}^J \delta_{\theta_j}$. Then $\mu^\star$ is of the form $\sum_{j=1}^J \lambda_j^\star \delta_{\theta_j}$ where $(\lambda_1^\star, \ldots, \lambda_J^\star)$ belongs to the simplex of dimension J. Moreover, the two quantities $\mathrm{KL}(\mu^\star \parallel \mu_1)$ and Δ_1 can easily be

bounded in terms of J. Indeed, using that $\log u \le u - 1$ for all $u > 0$ and that $\sum_{j=1}^{J} \lambda_j^{\star 2} \le 1$, we obtain that

$$\mathrm{KL}(\mu^\star \parallel \mu_1) = \sum_{j=1}^{J} \lambda_j^\star \log \lambda_j^\star + \log J \le \log J.$$

As for Δ_1, we have by convexity that

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

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

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

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

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

(i) *Forward Kullback–Leibler divergence* ($\alpha = 1$): $\Gamma(v) = e^{-\eta v}$, $\eta \in (0, 1)$ *and* κ *is any real number* (*Entropic Mirror Descent*);
(ii) *Reverse Kullback–Leibler* ($\alpha = 0$) *and* α*-Divergence with* $\alpha \in \mathbb{R} \setminus \{0, 1\}$:
   (a) $\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$ *and* κ *is any real number* (*Entropic Mirror Descent*);
   (b) $\Gamma(v) = [(\alpha - 1)v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$, $\alpha > 1$ *and* $\kappa > 0$ (*Power Descent*);

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ *be such that* $\Psi_\alpha(\mu_1) < \infty$. *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 nonincreasing with a convergence rate characterized by* (13).

The proof of Theorem 3 can be found in [12], Appendix A.3. In terms of assumptions, we only require the gradients of the function Ψ_α to be bounded in l_∞-norm, which is a standard assumption, and the objective function to be finite at the starting measure μ_1, that is, $\Psi_\alpha(\mu_1) < \infty$, which again is a mild assumption that can even be discarded for all $\alpha \neq 0$ (see Remark 17 of [12], Appendix D).

Let us now illustrate the benefits of our approach with an example where the different constants appearing in (13) are bounded explicitly and where we compare the convergence rate we obtain with typical Mirror Descent convergence results from the optimisation literature.

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

In the particular case of the Entropic Mirror Descent, the constant κ does not appear in the update formula (4) due to the normalisation, so we can choose it however we want without

impacting the convergence of the algorithm. Notice then that by choosing $\kappa = -3|b|_{\infty,\alpha}$ and based on Example 2, we obtain the following convergence rate for all $\eta \in (0, 1)$:

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

Thus, in the particular case of Example 3, the dominant term in (13) with respect to the dimension J of the simplex is in $\log J$ so that we achieve an overall $O(\frac{\log J}{N})$ convergence rate. Furthermore, the range of possible values for η is stated explicitly, since the result holds for all $\eta \in (0, 1)$.

This is an improvement compared to standard Mirror Descent results, which under similar assumptions only provide an $O(1/\sqrt{N})$ convergence rate and assume an $O(1/\sqrt{N})$ learning rate (see [3] or [7], Theorem 4.2.). Indeed, Projected Gradient Descent and Entropic Mirror Descent typically achieve an $O(\sqrt{J/N})$ and $O(\sqrt{\log(J)/N})$ convergence rate respectively in the Simplex framework. This means that Theorem 3 improves with respect to both N and J compared to Projected Gradient Descent and that it improves with respect to N for the Entropic Mirror Descent with a small cost in terms of the dimension J of the simplex.

Moreover, while accelerated versions of the Mirror Descent (e.g., Mirror Prox, see [34] or [7], Theorem 4.4.) also yield an $O(1/N)$ convergence rate, they require the objective function to be sufficiently smooth, an additional assumption that we have bypassed when deriving our results.

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

(A4) (i) T is a compact metric space and $\mathcal{T}$ is the associated Borel σ-field;
(ii) for all $y \in \mathsf{Y}$, $\theta \mapsto k(\theta, y)$ is continuous;
(iii) we have $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} k(\theta, y) \times \sup_{\theta' \in \mathsf{T}} (\frac{k(\theta', y)}{p(y)})^{\alpha - 1} \nu(\mathrm{d}y) < \infty$.
If $\alpha = 0$, assume in addition that $\int_{\mathsf{Y}} \sup_{\theta \in \mathsf{T}} |\log(\frac{k(\theta, y)}{p(y)})| p(y) \nu(\mathrm{d}y) < \infty$.

Here, condition (A4)(iii) implies that $b_{\mu,\alpha}(\theta)$ and $\Psi_\alpha(\mu)$ 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* $\alpha < 1$, $\kappa \leq 0$ *and set* $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ *for all* $v \in \mathrm{Dom}_\alpha$. *Then, for all* $\zeta \in \mathrm{M}_1(\mathsf{T})$, *any* $\eta > 0$ *satisfies* (5) *and* $\Psi_\alpha(\zeta) < \infty$.

Let $\eta \in (0, 1]$. *Further assume that there exist* $\mu_1, \mu^\star \in \mathrm{M}_1(\mathsf{T})$ *such that the* (*well-defined*) *sequence* $(\mu_n)_{n \in \mathbb{N}^\star}$ *defined by* (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 nonincreasing*,
(ii) $\mu^\star$ *is a fixed point of* $\mathcal{I}_\alpha$,
(iii) $\Psi_\alpha(\mu^\star) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta)$.

The proof of Theorem 4 is deferred to [12], Appendix A.4. 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.

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
	$\alpha > 1$, $\Gamma(v) = [(\alpha-1)v+1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa > 0$
	$\alpha < 1$, $\Gamma(v) = [(\alpha-1)v+1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa \leq 0$

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

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

The results we obtained thus far are summarized in Table 2 below.

As Algorithm 1 typically involves an intractable integral in the Expectation step, we now turn to a Stochastic version of this algorithm.

4. Stochastic (α, Γ)-descent. We start by introducing the notation for the Stochastic version of Algorithm 1. Let $M \in \mathbb{N}^\star$ and let $\mu \in \mathrm{M}_1(\mathsf{T})$. The Stochastic (α, Γ)-descent algorithm one-step transition is defined as follows in Algorithm 2.

Let us now denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space and by $\mathbb{E}$ the associated expectation operator. Given $\hat{\mu}_1 \in \mathrm{M}_1(\mathsf{T})$, the Stochastic version of the Exact iterative scheme defined by (4) is then given by

$$\hat{\mu}_{n+1} = \hat{\mathcal{I}}_{\alpha,M}(\hat{\mu}_n), \quad n \in \mathbb{N}^\star, \tag{14}$$

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

$$\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta, Y_{m,n+1})}{\hat{\mu}_n k(Y_{m,n+1})} f'_\alpha\left(\frac{\hat{\mu}_n k(Y_{m,n+1})}{p(Y_{m,n+1})}\right) \tag{15}$$

with $Y_{1,n+1}, \ldots, Y_{M,n+1} \overset{\text{i.i.d}}{\sim} \hat{\mu}_n k$ conditionally on $\mathcal{F}_n$ and where $\mathcal{F}_1 = \varnothing$ and $\mathcal{F}_n = \sigma(Y_{1,2}, \ldots, Y_{M,2}, \ldots, Y_{1,n}, \ldots, Y_{M,n})$ for $k \geq 2$. Notice that we use $\hat{\mu}_n k$ as a sampler instead of $k(\theta, \cdot)$ in (15). As our algorithm optimises over μ, sampling with respect to $\hat{\mu}_n k$ is not only cheaper computationally, but it also gives preference to the interesting regions of the parameter space.

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

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

A first idea to study this algorithm is to adapt Theorem 2 to the Stochastic case. This can be done for the Entropic Mirror Descent and a bound on $\mathbb{E}[\Psi_\alpha(\hat{\mu}_n) - \Psi_\alpha(\mu^\star)]$ of the form $O(1/N) + O(1/\sqrt{M})$ can be derived for a wide range of constant learning rates η (see [12], Appendix B.1, for the formal statement of the result and its proof). Maintaining an $O(1/N)$ bound however requires $M \geq N^2$, which yields an overall computational cost of order N^3. Another option consists in adapting [35] to our framework. This option involves a learning rate policy $(\eta_n)_{n\in\mathbb{N}}$ and notably yields an $O(1/\sqrt{N})$ bound for a constant policy $\eta_n = \eta_0/\sqrt{N}$, as written in Theorem 5 below.

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

$$B_\alpha := \left(\sup_{\mu\in\mathrm{M}_1(\mathsf{T})} \int_{\mathsf{Y}} \sup_{\theta,\theta'\in\mathsf{T}} \frac{k(\theta,y)^2}{k(\theta',y)} \left| f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\right|^2 \nu(\mathrm{d}y)\right)^{1/2} < \infty, \tag{16}$$

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

$$\mathbb{E}\left[\Psi_\alpha\left(\sum_{n=1}^N w_n \hat{\mu}_n\right) - \Psi_\alpha(\mu^\star)\right] \leq \frac{B_\alpha^2 \sum_{n=1}^N \eta_n^2/2}{\sum_{n=1}^N \eta_n} + \frac{\mathrm{KL}(\mu^\star \parallel \hat{\mu}_1)}{\sum_{n=1}^N \eta_n}, \tag{17}$$

In particular, the decreasing policy $\eta_n = \eta_0/\sqrt{n}$ yields an $O(\log(N)/\sqrt{N})$ bound in (17). Furthermore, the constant policy $\eta_n = \eta_0/\sqrt{N}$ yields an $O(1/\sqrt{N})$ bound in (17), which is minimal for $\eta_0 = B_\alpha^{-1}\sqrt{2\mathrm{KL}(\mu^\star \parallel \hat{\mu}_1)}$.

The proof of Theorem 5 can be found in [12], Appendix B.2, and we give below an example satisfying condition (16).

EXAMPLE 4. Consider the case $\mathsf{Y} = \mathbb{R}^d$ and $\alpha = 1$. Let $r > 0$ and let $\mathsf{T} = \mathcal{B}(0,r) \subset \mathbb{R}^d$. Furthermore, let K_h be a Gaussian transition kernel with bandwidth h and denote by k_h its associated kernel density. Finally, let p be a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant Z such that for all $y \in \mathsf{Y}$, $p(y) = Z \times [0.5\mathcal{N}(y;\theta_1^\star, \boldsymbol{I_d}) + 0.5\mathcal{N}(y;\theta_2^\star, \boldsymbol{I_d})]$, where $\theta_1^\star, \theta_2^\star \in \mathsf{T}$ and $\boldsymbol{I_d}$ is the identity matrix. Then (16) holds and we can apply Theorem 5 (see [12], Appendix B.3, for details).

Notice that the $O(1/\sqrt{N})$ convergence rate from Theorem 5 holds under minimal assumptions on Ψ_α. However, bridging the gap with the $O(1/N)$ convergence rate in Theorem 3 typically requires much stronger smoothness and strong-convexity assumptions on Ψ_α which can be hard to satisfy in practice (see [7], Theorem 6.2, for the statement of this result and [8] for an example in Online Variational Inference). Bypassing any of these assumptions like we did in the ideal case in Theorem 3 in order to improve on Theorem 5 constitutes an interesting area of research which is beyond the scope of this paper.

As for the Stochastic version of Power Descent, we establish the total variation convergence of $\hat{\mathcal{I}}_{\alpha,M}(\mu)$ toward $\mathcal{I}_\alpha(\mu)$ as M goes to infinity for all $\mu \in \mathrm{M}_1(\mathsf{T})$. To do so, consider i.i.d random variables $Y_1, Y_2, \ldots$ with common density μk w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote by $\mathbb{E}$ the associated expectation operator. We then have Proposition 7 below.

PROPOSITION 7. *Assume* (A1). *Let* $\alpha \in \mathbb{R} \setminus \{1\}$, $\eta > 0$, κ *be such that* $(\alpha - 1)\kappa \geq 0$ *and set* $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ *for all* $v \in \mathrm{Dom}_\alpha$. *Let* $\mu \in \mathrm{M}_1(\mathsf{T})$ *be such that* $\Psi_\alpha(\mu) < \infty$, (5) *holds and*

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta)\mathbb{E}\left[\left\{\frac{k(\theta, Y_1)}{\mu k(Y_1)}\left(\frac{\mu k(Y_1)}{p(Y_1)}\right)^{\alpha-1} + (\alpha - 1)\kappa\right\}^{\frac{\eta}{1-\alpha}}\right] < \infty. \tag{18}$$

Then

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

The proof is deferred to [12], Appendix C.4. The crux of the proof consists in applying a dominated convergence theorem to nonnegative real-valued $(\mathcal{T} \otimes \mathcal{F}, \mathcal{B}(\mathbb{R}_{\geq 0}))$-measurable functions, which requires to consider a Generalized version of the dominated convergence theorem [12], Lemma 15, and an integrated law of large numbers [12], Lemma 16.

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

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

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

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

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

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

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_i) + \kappa)}, \tag{19}$$

with $Y_{1,n+1}, \ldots, Y_{M,n+1}$ drawn independently from $\hat{\mu}_n k$ conditionally on $\mathcal{F}_n$ and $\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j)$ is given by (15) for all $j = 1 \ldots J$. This leads to Algorithm 3.

In this particular framework, most of the computing effort at each step lies within the computation of the vector $(\hat{b}_{\hat{\mu}_n,\alpha,M}(\theta_j))_{1\leq j\leq J}$. Interestingly, these computations can also be used to obtain an estimate of the Evidence Lower Bound (resp., the Renyi-bound [27]) when $p(y) = p(y, \mathscr{D})$. These two quantities, which are written explicitly in Remark 18 from [12], Appendix D, allow us to assess the convergence of the algorithm and provide a bound on the log-likelihood (see [27], Theorem 1). Note also that if there is a need for very large J, one can approximate the summation appearing in $\hat{\mu}_n k$ using subsampling.

An important point is that Algorithm 3 does not require any information on how the $\{\theta_1, \ldots, \theta_J\}$ have been obtained in order to infer the optimal weights as it draws information from samples that are generated from $\mu_{\boldsymbol{\lambda}} k$. Since the algorithm leaves $\{\theta_1, \ldots, \theta_J\}$ unchanged throughout the optimisation of the mixture weights (we call it an *Exploitation step*),

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

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

Set $\boldsymbol{\lambda} = [\lambda_{1,1}, \ldots, \lambda_{J,1}]$.
while *not converged* **do**

Sampling step : Draw independently M samples $Y_1, \ldots, Y_M$ from $\mu_{\boldsymbol{\lambda}} k$.
Expectation step : Compute $\boldsymbol{B}_{\boldsymbol{\lambda}} = (b_j)_{1 \leq j \leq J}$ where

$$b_j = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_m)}{\mu_{\boldsymbol{\lambda}} k(Y_m)} f'_\alpha\left(\frac{\mu_{\boldsymbol{\lambda}} k(Y_m)}{p(Y_m)}\right) \tag{20}$$

and deduce $\boldsymbol{W}_{\boldsymbol{\lambda}} = (\lambda_j \Gamma(b_j + \kappa))_{1 \leq j \leq J}$ 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

a natural idea is to combine Algorithm 3 with an *Exploration step* that modifies the parameter set, which gives Algorithm 4.

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

As a side remark, notice also that we recover the mixture weights update rules from the Population Monte Carlo algorithm applied to reverse Kullback–Leibler minimisation [16] by considering the Power Descent with $\alpha = 0$ and $\eta = 1$. We have thus embedded this special case into a more general framework.

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

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 $\theta_{1,0}, \ldots, \theta_{J_0,0}$ from q_0. Set $t = 0$.
while *not converged* **do**

Exploitation step : Set $\Theta = \{\theta_{1,t}, \ldots, \theta_{J_t,t}\}$. Perform Mixture Stochastic (α, Γ)-descent and obtain $\boldsymbol{\lambda}$.
Exploration step : Perform any exploration step of our choice and obtain $\theta_{1,t+1}, \ldots, \theta_{J_{t+1},t+1}$. Set $t = t + 1$.

end

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 $\theta_1, \dots, \theta_J \in \mathsf{T}$, we then work within the approximating family

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

Exploration step. At time $t = 1 \dots T$, we resample among $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ according to the optimised mixture weights $\boldsymbol{\lambda}$. The obtained sample $\{\theta_{1,t+1}, \dots, \theta_{J_{t+1},t+1}\}$ is then perturbed stochastically using the Gaussian transition kernel K_{h_t}, which gives us our new parameter set. The hyperparameter h_t is adjusted according to the number of particles so that $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space (the optimal rate in nonparametric estimation when the function is at least 2-times continuously differentiable and the kernel has order 2 [44]).

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, e.g., [30]).

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

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

and as the dimension grows, the conditions of support are often not met in practice, meaning that there exists $A \in \mathcal{Y}$ such that $p(A, \mathscr{D}) = 0$ and $\mu k(A) > 0$. This implies that whenever $\alpha > 1$ we might have that $\hat{b}_{\mu,\alpha,M}(\theta) = \infty$ and that the α-divergence (or equivalently the Renyi-bound as written in Remark 18 from [12], Appendix D) is infinite, which is the sort of behaviour we would like to avoid. Thus, we restrict ourselves to the case $\alpha \leq 1$ 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 behaviour, but is still considered in the experiments as a reference.

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

5.1. *Toy example.* Following Example 4, the target p is a mixture density of two d-dimensional Gaussian distributions multiplied by a positive constant Z such that

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

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

- 0.5-Mirror Descent: $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,

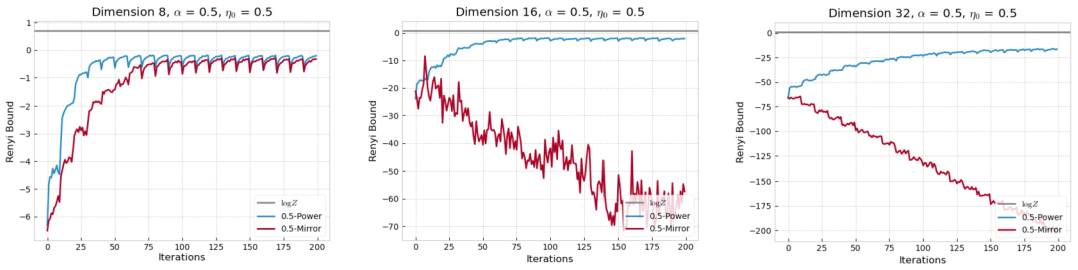

FIG. 1. *Plotted is the average Renyi-bound for the* 0.5*-Power and* 0.5*-Mirror Descent in dimension* $d = \{8, 16, 32\}$ *computed over* 100 *replicates with* $\eta_0 = 0.5$.

- 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.

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

$$\text{Mirror}: \quad \lambda_{j,n} \propto e^{\frac{\eta}{1-\alpha}[(\alpha-1)b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j)+(\alpha-1)\kappa]},$$

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

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 $\{\theta_1, \ldots, \theta_J\}$ 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.

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 $\alpha \leq 1$ and we focus on the Power Descent in our second example.

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

We test our algorithm for the *Covertype* dataset (581,012 data points and 54 features, available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html). Comput-

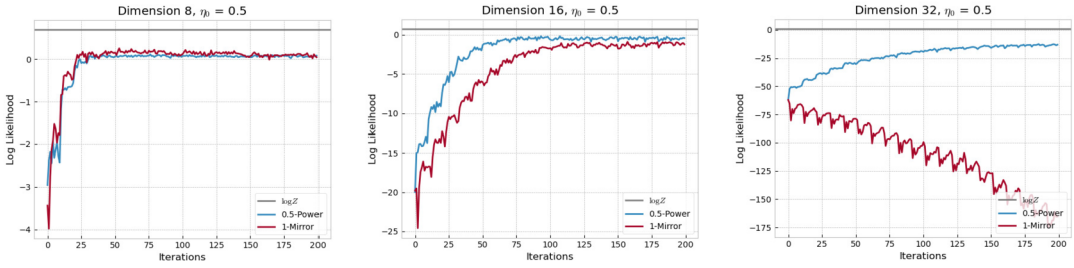

FIG. 2. *Plotted is the average Log-likelihood for* 0.5*-Power and* 1*-Mirror Descent in dimension* $d = \{8, 16, 32\}$ *computed over* 100 *replicates with* $\eta_0 = 0.5$.

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

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

One thing that is very specific to the Exploration step that we used to run our experiments (and sampling-based Exploration steps algorithms in general) is that the particles $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ are sampled from a known distribution at each Exploration step. This means that we are able to infer information on $\{\theta_{1,t}, \dots, \theta_{J_t,t}\}$ using Importance Sampling (IS) weights. We thus compare the Power (α, Γ)-descent with a state-of-the-art Adaptive Importance Sampling-based (AIS) algorithm (see, e.g., [9, 14, 24, 36]).

We initialise $\{\theta_{1,0}, \dots, \theta_{J_0,0}\}$ 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 $(\theta_{j,t})_{1 \le j \le J_t}$ 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

$$\lambda_{j,t} \propto \begin{cases} \dfrac{p(\theta_{j,t}, \mathscr{D})}{q_t(\theta_{j,t})} & \text{(AIS)}, \\ \Gamma\big(\hat{b}_{\mu_{\boldsymbol{\lambda}_{\text{init},t}},\alpha,M}(\theta_{j,t}) + \kappa\big) & \text{(Power)}. \end{cases} \tag{21}$$

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

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

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

Convergence rate. One could seek to establish additional convergence rate results in both the Exact and Stochastic cases, by for example refining the proof of Theorem 2 in the Stochastic case.

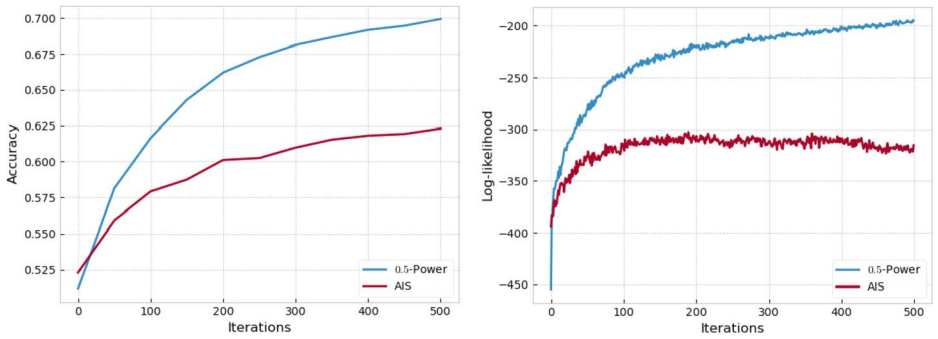

FIG. 3. *Plotted are the average Accuracy and Log-likelihood computed over* 100 *replicates for Bayesian Logistic Regression on the Covertype dataset for the* 0.5*-Power Descent and the AIS algorithm.*

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 $b_{\mu_n,\alpha}$.

Exploration step. Many other methods could be envisioned as an Exploration step and combined with the (α, Γ)-descent.

Acknowledgments. We are grateful to François Roueff for valuable remarks on the paper and to Olivier Fercoq for helpful discussions regarding the Mirror Descent algorithm. We would like to thank the Associate Editor and the referees for insightful comments and suggestions on the paper.

SUPPLEMENTARY MATERIAL

Supplement to: "Infinite-dimensional gradient-based descent for alpha-divergence minimisation" (DOI: 10.1214/20-AOS2035SUPP; .pdf).

- The remaining proofs are provided in the Supplementary Material, namely the proofs of Lemma 6, Theorem 2, 3, 4 and 5, Proposition 7, as well as the proof that Condition (16) is satisfied in Example 4.
- The Supplementary Material also contains the statement and proof for the $O(1/N) + O(1/\sqrt{M})$ bound on $\mathbb{E}[\Psi_\alpha(\hat{\mu}_n) - \Psi_\alpha(\mu^\star)]$ in the particular case of the Stochastic Entropic Mirror Descent and two additional remarks regarding Theorem 3 and the Renyi-bound.

REFERENCES

[1] BAMLER, R., ZHANG, C., OPPER, M. and MANDT, S. (2017). Perturbative black box variational inference. In *Advances in Neural Information Processing Systems* 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 5079–5088. Curran Associates.

[2] BEAL, M. J. (2003). Variational algorithms for approximate Bayesian inference. PhD Thesis.

[3] BECK, A. and TEBOULLE, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* **31** 167–175. MR1967286 https://doi.org/10.1016/S0167-6377(02)00231-6

[4] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

[5] BLEI, D. M., NG, A. Y. and JORDAN, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research* **3** 993–1022.

[6] BOTTOU, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'*2010 (Y. Lechevallier and G. Saporta, eds.) 177–186. Physica-Verlag/Springer, Heidelberg. MR3362066

[7] BUBECK, S. (2015). Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.* **8** 231–357. https://doi.org/10.1561/2200000050

[8] CHÉRIEF-ABDELLATIF, B.-E., ALQUIER, P. and KHAN, M. E. (2019). A generalization bound for online variational inference. In *Proceedings of the* 29*th International Conference on Machine Learning* **101** 662–677.

[9] CHOPIN, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* **32** 2385–2411. MR2153989 https://doi.org/10.1214/009053604000000698

[10] CICHOCKI, A. and AMARI, S. (2010). Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **12** 1532–1568. MR2659408 https://doi.org/10.3390/e12061532

[11] CICHOCKI, A., CRUCES, S. and AMARI, S.-I. (2011). Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **13** 134–170. https://doi.org/10.3390/e13010134

[12] DAUDEL, K., DOUC, R. and PORTIER, F. (2021). Supplement to "Infinite-dimensional gradient-based descent for alpha-divergence minimisation." https://doi.org/10.1214/20-AOS2035SUPP

[13] DEHAENE, G. and BARTHELMÉ, S. (2018). Expectation propagation in the large data limit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 199–217. MR3744718 https://doi.org/10.1111/rssb.12241

[14] DELYON, B. and PORTIER, F. (2019). Safe and adaptive importance sampling: A mixture approach. Available online: https://arxiv.org/abs/1903.08507 (accessed on 20 March 2020).

[15] DIENG, A. B., TRAN, D., RANGANATH, R., PAISLEY, J. and BLEI, D. (2017). Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems* 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 2732–2741. Curran Associates.

[16] DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2007). Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.* **35** 420–448. MR2332281 https://doi.org/10.1214/009053606000001154

[17] GERSHMAN, S., HOFFMAN, M. D. and BLEI, D. M. (2012). Nonparametric variational inference. In *Proceedings of the* 29*th International Conference on Machine Learning*.

[18] HELLINGER, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.* **136** 210–271. MR1580780 https://doi.org/10.1515/crll.1909.136.210

[19] HERNANDEZ-LOBATO, J., LI, Y., ROWLAND, M., BUI, T., HERNANDEZ-LOBATO, D. and TURNER, R. (2016). Black-box alpha divergence minimization. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 1511–1520. PMLR, New York, New York, USA.

[20] HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

[21] HSIEH, Y.-P., LIU, C. and CEVHER, V. (2019). Finding mixed Nash equilibria of generative adversarial networks. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 2810–2819. PMLR, Long Beach, CA.

[22] JAAKKOLA, T. S. and JORDAN, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models* (Jordan, M. I., ed.). *NATO ASI Series* (*Series D*: *Behavioural and Social Sciences*) **89**. Springer.

[23] JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233. https://doi.org/10.1023/A:1007665907178

[24] KLOEK, T. and VAN DIJK, H. K. (1978). Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica* 1–19.

[25] KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968 https://doi.org/10.1214/aoms/1177729694

[26] LI, Y., HERNÁNDEZ-LOBATO, J. M. and TURNER, R. E. (2015). Stochastic expectation propagation. In *Advances in Neural Information Processing Systems* 28 (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 2323–2331. Curran Associates.

[27] LI, Y. and TURNER, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems* 29 (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 1073–1081. Curran Associates.

[28] LINDSAY, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114. MR1292557 https://doi.org/10.1214/aos/1176325512

[29] MINKA, T. (2004). Power EP. Technical Report No. MSR-TR-2004-149.

[30] MINKA, T. (2005). Divergence measures and message passing. Technical Report No. MSR-TR-2005-173.
[31] MINKA, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. *UAI*'01 362–369. Morgan Kaufmann Publishers Inc., San Francisco, CA.
[32] MORIMOTO, T. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* 85–108.
[33] MORIMOTO, T. (1963). Markov processes and the H-theorem. *J. Phys. Soc. Jpn.* **18** 328–331. MR0167200 https://doi.org/10.1143/JPSJ.18.328
[34] NEMIROVSKI, A. (2004). Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.* **15** 229–251. MR2112984 https://doi.org/10.1137/S1052623403425629
[35] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041 https://doi.org/10.1137/070704277
[36] OH, M.-S. and BERGER, J. O. (1992). Adaptive importance sampling in Monte Carlo integration. *J. Stat. Comput. Simul.* **41** 143–168. MR1276184 https://doi.org/10.1080/00949659208810398
[37] OPPER, M. and WINTHER, O. (2000). Gaussian processes for classification: Mean-field algorithms. *Neural Comput.* **12** 2655–2684. https://doi.org/10.1162/089976600300014881
[38] PAISLEY, J., BLEI, D. and JORDAN, M. (2012). Variational Bayesian inference with stochastic search. In *Proceedings of the* 29*th International Conference on Machine Learning* 1363–1370.
[39] RANGANATH, R., GERRISH, S. and BLEI, D. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (S. Kaski and J. Corander, eds.). *Proceedings of Machine Learning Research* **33** 814–822. PMLR, Reykjavik, Iceland.
[40] RANGANATH, R., TRAN, D. and BLEI, D. (2016). Hierarchical variational models. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 324–333. PMLR, New York, New York, USA.
[41] RÉNYI, A. (1961). On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 547–561. Univ. California Press, Berkeley, CA. MR0132570
[42] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586
[43] SASON, I. (2018). On f-divergences: Integral representations, local behavior, and inequalities. *Entropy* **20** Paper No. 383, 32. MR3862573 https://doi.org/10.3390/e20050383
[44] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642
[45] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **60** 3797–3820. MR3225930 https://doi.org/10.1109/TIT.2014.2320500
[46] WANG, D., LIU, H. and LIU, Q. (2018). Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems* 31 (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) 5737–5747. Curran Associates.
[47] YIN, M. and ZHOU, M. (2018). Semi-implicit variational inference. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 5660–5669. PMLR, Stockholmsmässan, Stockholm Sweden.
[48] ZHANG, C., BUTEPAGE, J., KJELLSTROM, H. and MANDT, S. (2019). Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774
[49] ZHU, H. and ROHWER, R. (1995). Bayesian invariant measurements of generalization. *Neural Process. Lett.* **2** 28–31. https://doi.org/10.1007/BF02309013
[50] ZHU, H. and ROHWER, R. (1995). Information geometric measurements of generalisation. Technical Report No. NCRG/4350.