Regularized Rényi divergence minimization through Bregman proximal gradient algorithms

Thomas Guilmeau[1,a], Emilie Chouzenoux[1,b], and Víctor Elvira[2]

[1]Université Paris-Saclay, CentraleSupélec, INRIA, CVN, France
[a] thomas.guilmeau@inria.fr
[b] emilie.chouzenoux@centralesupelec.fr
[2]School of Mathematics, University of Edinburgh, United Kingdom
victor.elvira@ed.ac.uk

Abstract

We study the variational inference problem of minimizing a regularized Rényi divergence over an exponential family, and propose a relaxed moment-matching algorithm, which includes a proximal-like step. Using the information-geometric link between Bregman divergences and the Kullback-Leibler divergence, this algorithm is shown to be equivalent to a Bregman proximal gradient algorithm. This novel perspective allows us to exploit the geometry of our approximate model while using stochastic black-box updates. We use this point of view to prove strong convergence guarantees including monotonic decrease of the objective, convergence to a stationary point or to the minimizer, and geometric convergence rates. These new theoretical insights lead to a versatile, robust, and competitive method, as illustrated by numerical experiments.

Keywords. Variational inference, Rényi divergence, Kullback-Leibler divergence, Exponential family, Bregman proximal gradient algorithms.

MSC2020 Subject Classification. 62F15, 62F30, 62B11, 90C26, 90C30.

1 Introduction

Probability distributions of interest in statistical problems are often intractable. In Bayesian statistics for instance, the targeted posterior distributions often cannot be obtained in closed-form due to intractable normalization constants. Variational inference (VI) methods aim at finding good approximations by minimizing a divergence to the target over a family of parametric distributions [13, 81]. Such procedures can be summarized by the choice of approximating densities, the choice of divergence, and the algorithm used to solve the resulting optimization problem. As an example, the standard VI algorithm uses mean-field approximating densities and minimizes the exclusive Kullback-Leibler (KL) divergence. Assuming that the complete conditionals of the true model are in an exponential family, the optimal mean-field approximation can then be found by a deterministic coordinate-ascent algorithm [45].

The research on VI methods has been very active in the last years (see the review of **(author?)** 81). Majorization techniques have been proposed to cope with large scale models not satisfying conjugacy hypotheses [60, 82, 47]. Another approach in such challenging contexts is to run a stochastic gradient descent, which leads to the so-called black-box VI methods [75, 58, 43, 29]. Black-box methods allow a broad choice of divergence, like the α-divergences [43, 29, 27] and Rényi divergences [58], which are generalizations of the KL divergence depending on a scalar parameter $\alpha > 0$. This parameter can be chosen in order to enforce a mode-seeking or a mass-covering behavior in the approximations. On the contrary, minimizing the exclusive KL divergence may lead to under-estimation of the variance of the target [64, 13].

VI algorithms have also benefited from advances in information geometry, a field that studies statistical models through a differential-geometric lens. Among other results from this field, it has been shown that the Fisher information matrix can play the role of a metric tensor such that the square of the induced Riemannian distance is locally equivalent to the KL divergence [2]. Another useful insight when exponential families are considered is the relation between the KL divergence, Bregman divergences, and dual geometry [69]. These ideas can be leveraged by using the *natural gradient* [3], which amounts to a preconditioning of the standard gradient by the inverse Fisher information matrix. In the VI algorithms investigated in [46, 42, 45, 59], the standard gradient of the evidence lower bound is thus adjusted to take into account the Riemannian geometry of the approximating distributions, leading to simpler updates and improved behavior.

Despite those advances, there are still shortcomings in the development and understanding of VI algorithms, and as such, we identify below two main limitations.

First, to the best of our knowledge, there are still few links between black-box VI algorithms and natural gradient VI algorithms in the literature. On the one hand, the former methods allow to tackle a broad range of targets using various divergence measures but are usually restricted to the use of standard stochastic gradients. On the other hand, the latter methods use the more efficient and robust natural gradients, but are often limited to certain class of divergence, target, and approximating family. In this direction, let us however mention that information-geometric procedures have been deployed along black-box updates in [51, 52, 49], but these works remain restricted to the minimization of the exclusive KL divergence. One can also mention [72] where the minimization of an α-divergence over a mean-field family is studied using the Fisher Riemannian geometry.

Second, convergence studies of VI schemes are mostly empirical for black-box VI schemes [75, 58, 43, 29], and the same arises for schemes based on natural gradients [46, 42, 45, 59]. Indeed, the considered optimization problems are non-convex, making the algorithms hard to analyze. This is in stark contrast with MCMC methods, which can be used alternatively to VI, or optimization procedures, upon which many VI methods are based. MCMC methods based on Langevin diffusion are guaranteed to asymptotically produce samples from the target but also benefit from non-asymptotic convergence guarantees [78, 32] under log-concavity assumptions on the target. Recent works in VI have employed such techniques to prove convergence to the minimizer of the exclusive KL divergence, in the case of a mean-field approximating family [80] or a Gaussian approximating family [57], under the same type of hypotheses.

1.1 Contributions and outline

In this paper, we propose a novel VI algorithm that links black-box VI methods and natural gradient VI methods, while benefiting from solid convergence guarantees. Our algorithm minimizes a versatile composite objective, which is the sum of a Rényi divergence between the target and an exponential family, and a possible regularization term.

In order to solve the minimization problem, we introduce the so-called proximal relaxed moment-matching algorithm, whose iterations are composed of a relaxed moment-matching step, followed by a proximal-like step. A sampling-based implementation using only evaluations of the target unnormalized density is also provided to cover the black-box setting. The convergence of our new algorithm is then studied using the theory of Bregman proximal gradient algorithms. In particular, it exploits an equivalence relationship between Bregman divergences and the KL divergence arising within the space of exponential families.

Bregman proximal gradient algorithms [9, 7, 74, 67] are recent optimization methods arising from the generalization of the powerful proximal minimization schemes from the Euclidean setting [23]. Bregman-based algorithms allow to choose a Bregman divergence that tailors the intrinsic geometry of an optimization problem, more suitably than the standard Euclidean one [7, 74]. Note that stochastic methods have also been generalized in this fashion [40, 79]. Also related are proximal methods on perspective functions [24, 38], where divergences (typically, ϕ-divergence) are directly processed through their proximity operator on the Euclidean metric.

We show in this paper that the connection between VI algorithms and proximal optimization algorithms written in Bregman geometry yields many theoretical and practical insights. To summarize, our main contributions are the following:

- We propose a deterministic VI algorithm for an exponential approximating family. We show that our method can be written as a Bregman proximal gradient algorithm whose Bregman divergence is induced by the KL divergence, and exploits per se the geometry of the approximating family. We propose a stochastic implementation for our method. We show that it can be seen as a stochastic Bregman proximal gradient algorithm in the same geometry, thus bridging the gap between information-geometric and black-box VI methods.

- Our deterministic algorithm is shown to achieve a monotonic decrease of the composite objective, with its fixed points being stationary points of the objective function. Convergence to these stationary points is established. We show that this scheme achieves a geometric convergence rate to the global minimizer, guaranteed to exist and to be unique, in the case when the Rényi divergence identifies with the inclusive KL divergence or when the target belongs to the considered approximating family.

- We explain through a simple counter-example how the convergence of equivalent schemes written in the Euclidean geometry may fail. This theoretical insight is backed by numerical studies highlighting the superior performance and robustness of our scheme over its Euclidean counterpart.

- Our algorithm generalizes many existing moment-matching algorithms. We show through numerical experiments in the Gaussian case how our additional parameters allow to create mass-covering or mode-seeking approximations and compensate high approximation errors.

- Our framework allows a possibly non-smooth regularization term that is handled in our algorithm through a proximal update. This allows in particular to use our algorithms to compute generalized information projections. We explicit the proximal operators of two regularizers that promote the good conditioning of the covariance matrix or the sparsity of the means of the approximating densities.

The paper is organized as follows. In Section 2, we recall basic facts about Rényi divergences and exponential families, before presenting the optimization problem we propose to solve. Then, in Section 3, we outline our algorithm, before providing an alternative black-box implementation for it. In Section 4, we show how these algorithms can be interpreted as Bregman proximal gradient algorithms in the geometry induced by the KL divergence, and state our working assumptions. Theoretical analysis is provided in Section 5. Finally, numerical experiments with Gaussian proposals are presented in Section 6. We discuss our results and possible future research lines in Section 7.

The supplementary material contains four appendices. The proofs of our results are deferred to Appendices A and B, while we construct a proximal operator in Appendix C. Additional numerical experiments are presented in Appendix D.

1.2 Notation

The discrete set $\{n_1, n_1+1, \dots, n_2\}$ defined for $n_1, n_2 \in \mathbb{N}$, $n_1 < n_2$ is denoted by $[\![n_1, n_2]\!]$. Throughout this work, $\mathcal{H}$ is a real Hilbert space of finite dimension n with scalar product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. We denote by $B(\theta, R)$ the closed ball centered at $\theta \in \mathcal{H}$ with radius $R > 0$. The interior of a set C is denoted by $\operatorname{int} C$.

The set of non-negative real numbers if denotes by $\mathbb{R}_+$ and the set of positive real numbers by $\mathbb{R}_{++}$. Similarly, we denote by $\mathbb{R}_-$ and $\mathbb{R}_{--}$ the sets of non-positive and negative real numbers, respectively. Consider the set of matrices of $\mathbb{R}^{d\times d}$. Then, the set of symmetric matrices is denoted by $\mathcal{S}^d$, the set of positive semidefinite matrices is denoted by $\mathcal{S}^d_+$, and the set of positive definite matrices is denoted by $\mathcal{S}^d_{++}$. The identity matrix is denoted by I, $\det(\cdot)$ denotes the determinant operator on matrices and $\|\cdot\|_F$ the Frobenius norm. We use Landau's notation, i.e., for some functions $f, g : \mathcal{H} \to \mathbb{R}$, we write $f(v) = o(g(v))$ if f is such that, for any $\epsilon > 0$, there exists v_0 with $\|v_0\|$ small enough such that $|f(v)| \leq \epsilon |g(v)|$ for any $v \in B(0, \|v_0\|)$. Convex analysis notations are those from [10]. In particular, we denote by $\Gamma_0(\mathcal{H})$ the set of proper convex lower-semicontinuous functions from $\mathcal{H}$ to $\mathbb{R} \cup \{+\infty\}$. The domain of a function $f : \mathcal{H} \to [-\infty, +\infty]$ is $\operatorname{dom} f := \{\theta \in \mathcal{H},\ f(\theta) < +\infty\}$. The indicator function function ι_C of a set $C \subset \mathcal{H}$ is defined for every $\theta \in \mathcal{H}$ by

$$\iota_C(\theta) = \begin{cases} 0 & \text{if } \theta \in C, \\ +\infty & \text{else.} \end{cases}$$

We detail below our notations for measure theory notions. In particular, the Borel algebra of a set $\mathcal{X}$ is denoted by $\mathcal{B}(\mathcal{X})$. $\mathcal{M}(\mathcal{X})$ is the set of measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and $\mathcal{P}(\mathcal{X})$ is the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Given $m_1, m_2 \in \mathcal{M}(\mathcal{X})$, we write $m_1 \ll m_2$ when m_1 is absolutely continuous with respect to m_2. For a given $m \in \mathcal{M}(\mathcal{X})$ and a measurable function $h : \mathcal{X} \to \mathcal{H}$, we denote by $m(h)$ the vector of $\mathcal{H}$ defined by $(m(h))_i = \int_{\mathcal{X}} h_i(x) m(dx)$ for $i \in [\![1, n]\!]$. Finally, $\mathcal{N}(\cdot\,; \mu, \Sigma)$ denotes the density of a Gaussian probability measure with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathcal{S}^d_{++}$.

2 Problem of interest

We propose to reformulate the problem of approximating a target π by a parametric distribution q_θ as a variational minimization problem. In this context, the optimal parameters θ are defined to minimize a divergence to the target. Specifically, we focus here on the case when q_θ lies in an exponential family, and we propose to optimize its parameters θ through the minimization of a Rényi divergence between π and q_θ with a regularization term. In this section, we first recall important definitions regarding Rényi divergences (including the Kullback-Leibler divergence as a special case) and exponential families. We then introduce our variational inference (VI) problem.

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space. Let us consider a measure $\nu \in \mathcal{M}(\mathcal{X})$, with the sets $\mathcal{M}(\mathcal{X}, \nu) := \{m \in \mathcal{M}(\mathcal{X}),\ m \ll \nu\}$ and $\mathcal{P}(\mathcal{X}, \nu) := \{p \in \mathcal{P}(\mathcal{X}),\ p \ll \nu\}$. We are interested in approximating the target probability distribution $\pi \in \mathcal{P}(\mathcal{X}, \nu)$.

2.1 Rényi and Kullback-Leibler divergences

Rényi divergences and Kullback-Leibler (KL) divergence are widely used in statistics as discrepancy measures between probability distributions. To define them, let us consider two probability densities $p_1, p_2 \in \mathcal{P}(\mathcal{X}, \nu)$. We can then define the Rényi and KL divergences between p_1 and p_2 as follows.

Definition 1. The *Rényi divergence* with parameter $\alpha > 0$, $\alpha \neq 1$, between p_1 and p_2 is defined by

$$RD_\alpha(p_1, p_2) = \frac{1}{\alpha - 1} \log \left(\int p_1(x)^\alpha p_2(x)^{1-\alpha} \nu(dx) \right).$$

When the above integral is not well-defined, then $RD_\alpha(p_1, p_2) = +\infty$.

Definition 2. The KL divergence between p_1 and p_2 is defined by

$$KL(p_1, p_2) = \int \log \left(\frac{p_1(x)}{p_2(x)} \right) p_1(x) \nu(dx).$$

When the above integral is not well-defined, then $KL(p_1, p_2) = +\infty$.

The KL divergence is a limiting case of Rényi divergence as shown by [76], since

$$\lim_{\alpha \to 1,\, \alpha \leq 1} RD_\alpha(p_1, p_2) = KL(p_1, p_2).$$

Let us recall the important following property, that explains the term *divergence*:

Proposition 1 ([76]). *For any $\alpha > 0$,*

$$RD_\alpha(p_1, p_2) \geq 0, \text{ and } RD_\alpha(p_1, p_2) = 0 \text{ if and only if } p_1 = p_2,$$

where $RD_1(p_1, p_2)$ is being taken equal to $KL(p_1, p_2)$.

2.2 Exponential families

In this work, we propose to approximate the target $\pi \in \mathcal{P}(\mathcal{X}, \nu)$ by a parametric distribution taken from an exponential family [15, 6].

Definition 3. Let $\Gamma : \mathcal{X} \to \mathcal{H}$ be a Borel-measurable function. The exponential family with base measure ν and sufficient statistics Γ is the family $\mathcal{Q} = \{q_\theta \in \mathcal{P}(\mathcal{X}, \nu),\, \theta \in \Theta\}$ such that

$$q_\theta(x) = \exp\left(\langle \theta, \Gamma(x) \rangle - A(\theta)\right), \forall x \in \mathcal{X}, \tag{1}$$

with A being the log-partition function, such that $\Theta = \operatorname{dom} A \subset \mathcal{H}$, and which reads:

$$A(\theta) = \log\left(\int \exp\left(\langle \theta, \Gamma(x) \rangle\right) \nu(dx)\right), \forall \theta \in \Theta. \tag{2}$$

In the following, for the sake of conciseness, we will say that some family $\mathcal{Q}$ is an exponential family, without stating explicitly the base measure and the sufficient statistics $\mathcal{Q}$ is associated to.

Remark 1. We work here with parameters in the finite-dimensional Hilbert space $\mathcal{H}$, which is slightly more general than considering parameters in $\mathbb{R}^n$. This allows to consider vectors, matrices, or Cartesian products in a unified way. In particular, when symmetric matrices are considered, we work directly with $\mathcal{S}^d$ rather than with its vectorized counterpart $\mathbb{R}^{d(d+1)/2}$.

The goal of our approximation method is thus to find $\theta \in \Theta$ such that q_θ is an optimal approximation of π, in a sense that remains to be precised. Before going further, let us provide an important example of an exponential family.

Example 1. Let $d \geq 1$. Consider the family of Gaussian distributions with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathcal{S}^d_{++}$. This is an exponential family [6], with sufficient statistics $\Gamma : x \longmapsto \left(x, xx^\top\right)^\top$ and Lebesgue base measure that we denote by $\mathcal{G}$ in the following. Its corresponding parameters are $\theta = (\theta_1, \theta_2)^\top$ with $\theta_1 = \Sigma^{-1}\mu$, and $\theta_2 = -\frac{1}{2}\Sigma^{-1}$, while $A(\theta) = \frac{d}{2}\log(2\pi) - \frac{1}{4}\theta_1^\top \theta_2^{-1}\theta_1 - \frac{1}{2}\log\det(-2\theta_2)$. The domain of A is $\Theta = \mathbb{R}^d \times \left(-\mathcal{S}^d_{++}\right)$, which is included in $\mathcal{H} = \mathbb{R}^d \times \mathcal{S}^d$. The scalar product of $\mathcal{H}$ is taken as the sum of the scalar product of $\mathbb{R}^d$ and the one of $\mathcal{S}^d$.

Exponential families recover many other continuous distributions, such as the inverse Gaussian and Wishart distributions, among others. Discrete distributions can also be put under the form (1) when ν is chosen as a discrete measure. Exponential families benefit from a rich geometric structure [2, 69] and have been used as approximating families in many contexts such as VI algorithms [42, 45, 13, 59], or adaptive importance sampling (AIS) procedures [1].

2.3 Proposed approximation approach

We seek to approximate π by a parametric distribution q_θ from an exponential family $\mathcal{Q}$ with base measure ν, such that the domain $\Theta \subset \mathcal{H}$ is non-empty. To measure the quality of our approximations, we define the following family of functions $f_\pi^{(\alpha)}$ for $\alpha > 0$:

$$f_\pi^{(\alpha)}(\theta) := \begin{cases} RD_\alpha(\pi, q_\theta), & \text{if } \alpha \neq 1, \\ KL(\pi, q_\theta), & \text{if } \alpha = 1, \end{cases} \forall \theta \in \Theta. \tag{3}$$

Consider now a *regularizing* term r, which promotes desirable properties on the sought parameters θ. We now define our objective function for some $\alpha > 0$:

$$F_\pi^{(\alpha)}(\theta) := f_\pi^{(\alpha)}(\theta) + r(\theta), \ \forall \theta \in \Theta. \tag{4}$$

We propose to resolve our approximation problem by minimizing (4) over an exponential family $\mathcal{Q}$, i.e., by considering the following optimization problem:

$$\underset{\theta\in\Theta}{\text{minimize}} \ F_\pi^{(\alpha)}(\theta). \tag{$P_\pi^{(\alpha)}$}$$

Problem ($P_\pi^{(\alpha)}$) consists in minimizing $F_\pi^{(\alpha)}$, which is the sum of the Rényi divergence $RD_\alpha(\pi, \cdot)$ and a regularizing function r. This allows to capture or generalize many settings.

Minimizing the KL divergence leads to a particular behavior that may be undesirable in practice. For instance, minimizing $KL(\pi, \cdot)$ induces a mass-covering behavior while minimizing $KL(\cdot, \pi)$ induces a mode-fitting behavior [64, 13]. In contrast, working with a Rényi divergence as a discrepancy measure allows to generalize the KL divergence, recovered when $\alpha = 1$ while allowing to choose the right value of α [58], hence fine-tuning the algorithm's behavior for the application at hand. Moreover, the Rényi divergence with parameter α can be monotonically transformed [76] into the corresponding α-divergence [43, 27], including in particular the χ^2 divergence [29, 1].

Adding a regularization term gives even more possibilities. When r is null or an indicator function, then Problem ($P_\pi^{(\alpha)}$) relates to the computation of the so-called reverse information projection [33, 26] when $\alpha = 1$, which has later been generalized by [56] for $\alpha \neq 1$. A similar setting is used in sparse precision matrix estimation, relying on the KL divergence and a sparsity-inducing regularizer [5]. The problem of computing Bayesian coresets has also been formulated as a KL minimization problem over a set of sparse parameters by [20]. Let us also mention that [73] added a graph regularization term to a KL divergence, to enforce special geometric structure. Finally, the minimization of problems composed of a divergence and an additional term is at the core of the generalized view on variational inference proposed by [53].

3 A proximal relaxed moment-matching algorithm

In this section, we detail our proposed algorithm and its behavior, and discuss its connections with existing works. Our algorithm solves Problem ($P_\pi^{(\alpha)}$) by adapting the parameters θ iteratively. Each iteration is composed of two steps: (i) a relaxed moment-matching step, and (ii) a proximal step, both described in Section 3.1. Then, we provide a black-box implementation of our method based on non-linear importance sampling in Section 3.2. Finally, we discuss in Section 3.3 how our method generalizes existing moment-matching algorithms.

3.1 A proximal relaxed moment-matching algorithm

In order to state our algorithm, we first introduce the notion of geometric average between our target π and the parametric density q_θ.

Definition 4. Consider $\theta \in \Theta$ and $\alpha > 0$. We introduce, whenever it is well-defined, the *geometric average* with parameter α between π and q_θ, denoted by $\pi_\theta^{(\alpha)}$, which is the probability distribution of $\mathcal{P}(\mathcal{X}, \nu)$ defined by

$$\pi_\theta^{(\alpha)}(x) = \frac{1}{\int \pi(y)^\alpha q_\theta(y)^{1-\alpha} \nu(dy)} \left(\pi(x)^\alpha q_\theta(x)^{1-\alpha} \right), \forall x \in \mathcal{X}. \tag{5}$$

Probability densities akin to $\pi_\theta^{(\alpha)}$ have been used for instance in annealing importance sampling [68], in sequential Monte-Carlo schemes [65], or in adaptive importance sampling [16]. The integral in (5) is well-defined if $\alpha \leq 1$ and the supports of π and q_θ have non-empty intersection. Since π and every $q_\theta \in \mathcal{Q}$ are absolutely continuous with respect to ν, and $q_\theta(x) > 0$ for every $x \in \mathcal{X}$, the latter condition is always satisfied within the setting of our study.

Remark 2. If one does not have access to π but only to an unnormalized density $\tilde{\pi}$ such that $\pi(x) = \frac{1}{Z_\pi}\tilde{\pi}(x)$, the geometric average between π and q_θ can be still computed using

$$\pi_\theta^{(\alpha)}(x) = \frac{1}{\int \tilde{\pi}(y)^\alpha q_\theta(y)^{1-\alpha} \nu(dy)} \left(\tilde{\pi}(x)^\alpha q_\theta(x)^{1-\alpha} \right), \forall x \in \mathcal{X}.$$

We are now ready to introduce our *proximal relaxed moment-matching algorithm*, described in Algorithm 1. At iteration k, the first step, Eq. (6) can be viewed as a relaxed form of a moment-matching step, with relaxation step-size τ_{k+1} chosen such that $\tau_{k+1} \in (0, 1]$. The parameter α arises from the Rényi divergence $f_\pi^{(\alpha)}$. The second step, Eq. (7), is a so-called proximal step on the regularization term r (see Section 4.1) that involves again the step-size τ_{k+1}.

Algorithm 1: Proposed proximal relaxed matching algorithm

Choose the step-sizes $\{\tau_k\}_{k\in\mathbb{N}}$, such that $\tau_k \in (0, 1]$ for any $k \in \mathbb{N}$.
Set the Rényi parameter $\alpha > 0$.
Initialize the algorithm with $\theta_0 \in \operatorname{int}\Theta$.
for $k = 0, ...$ **do**

Compute $\theta_{k+\frac{1}{2}}$ such that

$$q_{\theta_{k+\frac{1}{2}}}(\Gamma) = \tau_{k+1}\pi_{\theta_k}^{(\alpha)}(\Gamma) + (1 - \tau_{k+1}) q_{\theta_k}(\Gamma). \tag{6}$$

Update θ_{k+1} following

$$\theta_{k+1} = \operatorname*{arg\,min}_{\theta' \in \Theta} \left(r(\theta') + \frac{1}{\tau_{k+1}} KL(q_{\theta_{k+\frac{1}{2}}}, q_{\theta'}) \right). \tag{7}$$

end

The following example explicits the relaxed moment-matching step of Algorithm 1 when the exponential family is Gaussian.

Example 2. In the case when $\mathcal{Q} = \mathcal{G}$, the first and second order moments $(q_\theta(x), q_\theta(xx^\top))^\top$ are the sufficient statistics of the distribution q_θ. The update (6) reads in this case

$$\begin{cases} q_{\theta_{k+\frac{1}{2}}}(x) & = \tau_{k+1}\pi_{\theta_k}^{(\alpha)}(x) + (1 - \tau_{k+1}) q_{\theta_k}(x), \\ q_{\theta_{k+\frac{1}{2}}}(xx^\top) & = \tau_{k+1}\pi_{\theta_k}^{(\alpha)}(xx^\top) + (1 - \tau_{k+1}) q_{\theta_k}(xx^\top). \end{cases} \tag{8}$$

This shows that (6) consists in matching the first and second order moments of the new distribution $q_{\theta_{k+\frac{1}{2}}}$ with a convex combination between the moments of $\pi_{\theta_k}^{(\alpha)}$ and those of the previous distribution q_{θ_k}. We recall that, for $q_\theta \in \mathcal{G}$, $q_\theta(x) = \mu$ and $q_\theta(xx^\top) = \Sigma + \mu\mu^\top$. Thus, we can further write that (8) is equivalent to

$$\begin{cases} \mu_{k+\frac{1}{2}} & = \tau_{k+1}\pi_{\theta_k}^{(\alpha)}(x) + (1-\tau_{k+1})\mu_k, \\ \Sigma_{k+\frac{1}{2}} & = \tau_{k+1}\pi_{\theta_k}^{(\alpha)}(xx^\top) + (1-\tau_{k+1})\left(\Sigma_k + \mu_k\mu_k^\top\right) - \mu_{k+\frac{1}{2}}\mu_{k+\frac{1}{2}}^\top. \end{cases}$$

We now give an example in order to illustrate the second step of Algorithm 1. This example is rather general and links Eq. (7) with reverse information projections [33, 26]. An example of this step is provided in Appendix C.

Example 3. The proximal step (7) encompasses the notion of projection if the function r is the indicator ι_C of a non-empty closed convex set $C \subset \mathcal{H}$ [10, Example 12.25]. We obtain in this case

$$\theta_{k+1} = \underset{\theta' \in \Theta \cap C}{\arg\min}\, KL(q_{\theta_{k+\frac{1}{2}}}, q_{\theta'}).$$

We recognize that in this case, (7) is the reversed information projection of $q_{\theta_{k+\frac{1}{2}}}$ on the set $\{q_\theta \in \mathcal{Q},\, \theta \in C \cap \Theta\}$, as described in [26, Section 3] for instance.

3.2 A black-box implementation based on non-linear importance sampling

Implementing directly Algorithm 1 might not be possible in practice. In many situations, $\pi_\theta^{(\alpha)}(\Gamma)$ cannot be expressed analytically, and must be approximated. We thus propose a stochastic implementation of Algorithm 1 based on non-linear importance sampling. This new scheme only requires that samples distributed following q_θ are available for any $\theta \in \Theta$, and that there exists $\tilde{\pi} \in \mathcal{M}(\mathcal{X}, \nu)$ and $Z_\pi > 0$ such that for any $x \in \mathcal{X}$, $\pi(x) = \frac{1}{Z_\pi}\tilde{\pi}(x)$ with $\tilde{\pi}(x)$ being easy to compute.

This setting is standard in importance sampling as well as in black-box VI [71] for instance. The proposed stochastic form of Algorithm 1 is motivated by the following alternative form of $\pi_\theta^{(\alpha)}(\Gamma)$:

$$\pi_\theta^{(\alpha)}(\Gamma) = \frac{1}{\int \left(\frac{\tilde{\pi}(y)}{q_\theta(y)}\right)^\alpha q_\theta(y)\nu(dy)} \int \left(\frac{\tilde{\pi}(x)}{q_\theta(x)}\right)^\alpha \Gamma(x) q_\theta(x)\nu(dx). \tag{9}$$

We see here that both integrals in Eq. (9) are expectations with respect to q_θ, with the ratios $\left(\frac{\tilde{\pi}(x)}{q_\theta(x)}\right)^\alpha$ evoking exponentiated importance weights. Therefore, our approximate implementation of Algorithm 1 consists in approximating these integrals with weighted samples from q_θ, which yields Algorithm 2.

Algorithms 1 and 2 are both written assuming that the proximal step can be computed exactly. An example of such computations is provided in Appendix C. However, it may not be the case, depending on r and $\mathcal{Q}$. Still, the minimization step in (7) is easier to solve than Problem ($P_\pi^{(\alpha)}$) since we deal with the KL divergence rather than the Rényi divergence (see Proposition 9 to see why this case leads to better properties) and the target is now $q_{\theta_{k+\frac{1}{2}}} \in \mathcal{Q}$. This property can be exploited by using specialized optimization subroutines to approximate this step. Specific cases have been investigated in the literature. For instance, the proximal algorithm proposed by [11] can be used in the case of Gaussian densities with fixed mean. A graphical lasso solver [5] can also be employed for computation of this step for Gaussian densities with fixed mean and ℓ_1 regularizer. When $r = \iota_C$, the substep of Eq. (7) reduces to a reversed information projection, as discussed in Example 3. Note that in this case, Problem ($P_\pi^{(\alpha)}$) is the Rényi information projection of π on the set $\{q_\theta \in \mathcal{Q},\, \theta \in \Theta \cap C\}$. This shows that Algorithm 1 allows to compute Rényi information projections using a sequence of reversed information projections, that may be more easily available. Rényi information projections have been studied for instance by [56], where possible applications are discussed.

Algorithm 2: Monte Carlo proximal relaxed moment-matching algorithm

Choose the step-sizes $\{\tau_k\}_{k\in\mathbb{N}}$, such that $\tau_k \in (0,1]$ for any $k \in \mathbb{N}$.
Choose the sample sizes $\{N_k\}_{k\in\mathbb{N}}$, such that $N_k \in \mathbb{N}\setminus\{0\}$ for any $k \in \mathbb{N}$.
Set the Rényi parameter $\alpha > 0$.
Initialize the algorithm with $\theta_0 \in \operatorname{int}\Theta$.
for $k = 0, \ldots$ **do**

Sample $x_l \sim q_{\theta_k}$ for $l \in [\![1, N_{k+1}]\!]$.

For every $l \in [\![1, N_{k+1}]\!]$, compute the non-linear importance weight $w_l^{(\alpha)}$ and their normalized counterpart $\bar{w}_l^{(\alpha)}$ through

$$w_l^{(\alpha)} = \left(\frac{\tilde{\pi}(x_l)}{q_{\theta_k}(x_l)}\right)^{\alpha}, \quad \bar{w}_l^{(\alpha)} = \frac{w_l^{(\alpha)}}{\sum_{l=1}^{N_{k+1}} w_l^{(\alpha)}}. \tag{10}$$

Compute $\theta_{k+\frac{1}{2}}$ such that

$$q_{\theta_{k+\frac{1}{2}}}(\Gamma) = \tau_{k+1}\left(\sum_{l=1}^{N_{k+1}} \bar{w}_l^{(\alpha)}\Gamma(x_l)\right) + (1-\tau_{k+1})q_{\theta_k}(\Gamma). \tag{11}$$

Update θ_{k+1} following Eq. (7).

end

3.3 Comparison with existing moment-matching algorithms

Let us now discuss the main features of our algorithms, and their positioning with respect to existing moment-matching algorithms. First, note that a strict moment-matching update of θ_{k+1},

$$q_{\theta_{k+1}}(\Gamma) = \pi(\Gamma), \tag{12}$$

is recovered in Algorithm 1 when $\tau_{k+1} = 1$, $\alpha = 1$ and $r \equiv 0$. Therefore, each update of Algorithm 1 can be viewed as a generalized version of the strict moment-matching update of Eq. (12) with supplementary degrees of freedom, hence its name.

Many algorithms in statistics resort to moment-matching updates. In AIS, the AMIS scheme [25] and the M-PMC scheme [21] rely on such updates. The idea of moment-matching updates with $\tau > 0$ as in (6) can also be found in many contexts, such as VI [52] or covariance learning in adaptive importance sampling [34], although π is not used directly in the latter. However, all the aforementioned works consider KL-based updates, that is with $\alpha = 1$ and no regularization term (i.e., $r \equiv 0$).

Moment-matching updates are often approximated through AIS, as we do in Algorithm 2. Importance sampling estimation of $\pi(\Gamma)$ or $\pi_\theta^{(\alpha)}(\Gamma)$ is for instance used in the AMIS scheme [25, 35] for adaptive importance sampling, where proposals are constructed by matching the moments of the target. AMIS is recovered when $\alpha = 1, \tau_k \equiv 1$ and $r \equiv 0$. However, note that in AMIS, all the past samples are used at each iteration and re-weighted (interpreting that samples are simulated in a multiple IS setting, **(author?)** 36), which is not the case here. In that respect, the APIS algorithm [61] bears some similarity, since it performs adaptation via moment matching with only the samples at each given iteration. Let us also mention the algorithm of [52], where deterministic and stochastic updates are combined to exploit the structure of the target.

When $\alpha = 1$, the weights of Algorithm 2 reduce to standard importance sampling weights, with q_{θ_k} as a proposal distribution. However, for $\alpha \neq 1$, then each weight comes from a non-linear transformation applied to the standard importance sampling weights. A particular type of non-linearity has been studied by [54], where cropped weights have been shown to decrease the variance of the estimator. Some related methodologies for a non-linear transformation of the importance weights can be found in [48, 77, 55]. One can also see the review of [62]. Note that similarly to cropping the weights, raising them at a power $\alpha \leq 1$, is also a concave transformation of the weights, which may make the estimators more robust too (see the bias-variance trade-off of **(author?)** 55, Lemma 1). This is confirmed by our numerical experiments in Section 6. Note that in our case, this transformation comes naturally from the fact that we minimize a Rényi divergence.

In a different context, moment-matching updates have been used by [39] to construct a path between two exponential distributions by averaging their moments, corresponding to $\alpha = 1$. Similarly, geometric paths using distributions similar to $\pi_\theta^{(\alpha)}$ have been used in [68, 65], corresponding to $\tau_k \equiv 0$. This means that our updates in Algorithm 1 use both techniques simultaneously. This is linked to the more general paths between probability distributions proposed by [17], or to the *q-paths* of [63]. Actually, moment-matching and geometric averages both are barycenters between π and q_θ in the sense of the inclusive or exclusive KL divergence [39], indicating that Eq. (6) may have a similar interpretation.

4 Geometric interpretation as a Bregman proximal gradient scheme

Let us show now that Algorithm 1 can be interpreted as a special case of a *Bregman proximal gradient algorithm* [7, 74]. This perspective will be a key element of our convergence analysis in Section 5. We show hereafter that Algorithms 1 and 2 lie within this framework and detail our working assumptions. The proofs are deferred to the supplementary material in Appendix A. Then, we discuss how our algorithms relate with natural gradient methods and black-box schemes.

4.1 Geometric interpretation as a Bregman proximal gradient scheme

In this section, we first recall some notions about Bregman proximal optimization schemes (more details can be found in **(author?)** 9, 7, 74). We then identify the Bregman geometry leading to our algorithms. Finally, we show the equivalence between Algorithm 1 and a Bregman proximal gradient algorithm within this particular geometry under some assumptions that we also explain here.

An essential tool of our analysis is the notion of Bregman divergence, that generalizes the standard Euclidean distance. The Bregman divergence paradigm allows to propose new optimization algorithms by relying on other geometries, with the aim to yield better convergence results and/or simpler updates for a given problem. Each Bregman divergence is constructed from a function satisfying the so-called Legendre property.

Definition 5. A *Legendre function* is a function $B \in \Gamma_0(\mathcal{H})$ that is strictly convex on the interior of its domain $\operatorname{int}\operatorname{dom} B$, and essentially smooth. B is essentially smooth if it is differentiable on $\operatorname{int}\operatorname{dom} B$ and such that $||\nabla B(\theta_k)|| \xrightarrow[k\to+\infty]{} +\infty$ for every sequence $\{\theta_k\}_{k\in\mathbb{N}}$ converging to a boundary point of $\operatorname{dom} B$ with $\theta_k \in \operatorname{int}\operatorname{dom} B$ for every $k \in \mathbb{N}$.

Given a Legendre function B, we define the *Bregman divergence* d_B as

$$d_B(\theta, \theta') := B(\theta) - B(\theta') - \langle \nabla B(\theta'), \theta - \theta' \rangle, \forall (\theta, \theta') \in (\operatorname{dom} B) \times (\operatorname{int}\operatorname{dom} B).$$

We now define the notion of conjugate function [10], which allows to state some useful properties of Legendre functions.

Definition 6. The *conjugate* of a function $f : \mathcal{H} \to [-\infty, +\infty]$ is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$f^*(\theta) = \sup_{\theta' \in \mathcal{H}} \langle \theta', \theta \rangle - f(\theta').$$

Proposition 2 (Section 2.2 in **(author?)** 74)**.** *Let B be a Legendre function. Then we have that*

(i) ∇B *is a bijection from* $\operatorname{int}\operatorname{dom} B$ *to* $\operatorname{int}\operatorname{dom} B^*$*, and* $(\nabla B)^{-1} = \nabla B^*$*,*

(ii) $\operatorname{dom}\partial B = \operatorname{int}\operatorname{dom} B$ *and* $\partial B(\theta) = \{\nabla B(\theta)\}$, $\forall \theta \in \operatorname{int}\operatorname{dom} B$.

Finally, B is a Legendre function if and only if B^* is a Legendre function.

The Bregman divergence $d_B(\theta, \theta')$ measures the gap between the value of the function B and its linear approximation at θ', when both are evaluated at θ. B is strictly convex, meaning that its curve is strictly above its tangent linear approximations. Thus, d_B satisfies the following distance-like property. Note however that d_B is not symmetric nor does it satisfy the triangular inequality in general.

Proposition 3 (Section 2.2 in **(author?)** 74)**.** *Consider a Legendre function B with the associated Bregman divergence d_B. Then, for every $\theta \in \operatorname{dom} B$, $\theta' \in \operatorname{int}\operatorname{dom} B$,*

$$d_B(\theta, \theta') \geq 0,$$
$$d_B(\theta, \theta') = 0 \textit{ if and only if } \theta = \theta'.$$

Each choice for the Legendre function B yields a specific divergence d_B. In particular, Bregman divergences generalize the Euclidean norm, since it is recovered for $B(\theta) = \frac{1}{2}\|\theta\|^2$ [7]. Given these notions, we can now explicit the geometry that will be useful to provide a new interpretation of our Algorithm 1. We will show that the log-partition function defined in (2) is a natural choice to generate a Bregman divergence.

We first make an assumption ensuring that the choice of $\mathcal{Q}$, given the target π, makes the function $f_\pi^{(\alpha)}$ well-posed.

Assumption 1. The exponential family $\mathcal{Q}$ and the target π are such that

(i) $\operatorname{int}\Theta \neq \emptyset$ and $\operatorname{int}\Theta \subset \operatorname{dom} f_\pi^{(\alpha)}$,

(ii) $\mathcal{Q}$ is *minimal* and *steep*, following the definitions of [6, Chapter 8].

Minimality implies in particular that for each distribution in $\mathcal{Q}$, there is a unique vector θ that parametrizes it. Most exponential families are steep. In particular, if Θ is open (in this case, $\mathcal{Q}$ is called *regular*), then $\mathcal{Q}$ is steep [6, Theorem 8.2]. Note that when $\alpha \in (0, 1)$, then $\operatorname{dom} f_\pi^{(\alpha)} = \Theta$ so that Assumption 1 (i) holds. Indeed, $q_\theta(x) > 0$ for every $x \in \mathcal{X}$ and, in particular, $q_\theta(x)$ is positive as soon as $\pi(x) > 0$. This means that the quantity in the logarithm is positive. When $\alpha = 1$, we have

$$KL(\pi, q_\theta) = \int \log(\pi(x))\pi(x)\nu(dx) - \langle \theta, \pi(\Gamma) \rangle + A(\theta),\ \forall \theta \in \Theta.$$

Thus $\operatorname{dom} f_\pi^{(\alpha)} = \Theta$, and Assumption 1 (i) holds if $\int \log(\pi(x))\pi(x)\nu(dx)$ and $\pi(\Gamma)$ are finite. However, Assumption 1 (i) may not be satisfied when $\alpha > 1$.

Proposition 4. *Under Assumption 1 (i), the log-partition A, defined in Eq. (2), is proper, lower semicontinuous and strictly convex. In addition, all the partial derivatives of A exist on* $\operatorname{int}\Theta$*. In particular, its gradient reads*

$$\nabla A(\theta) = q_\theta(\Gamma),\ \forall \theta \in \operatorname{int}\Theta. \tag{13}$$

If Assumption 1 (i)-(ii) is satisfied, then the log-partition function is a Legendre function.

The Bregman divergence induced by the Legendre function A admits a statistical interpretation that has been well-studied in the information geometry community [69]. Indeed, the KL divergence between two distributions from $\mathcal{Q}$ is equivalent to the Bregman divergence d_A between their parameters, as we recall in the next proposition.

Proposition 5 (**(author?)** 69)**.** *Consider $\theta, \theta' \in \operatorname{int}\Theta$ and A the log-partition function defined in* (2)*. Then,*

$$KL(q_\theta, q_{\theta'}) = d_A(\theta', \theta).$$

This proposition links the KL divergence with the notion of Bregman divergence, which is also central to many new algorithms in optimization. We now exploit this connection to analyze Algorithm 1 as an optimization algorithm written with the divergence d_A. We first give an intermediate proposition that shows the differentiability of $f_\pi^{(\alpha)}$, and thus properly justifies the use of the gradients of $f_\pi^{(\alpha)}$ in our following study.

Proposition 6. *Let $\alpha > 0$. The map $f_\pi^{(\alpha)}$ is of class $\mathcal{C}^2$ on $\operatorname{int}\Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$. In particular, for any $\theta \in \operatorname{int}\Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$,*

$$\nabla f_\pi^{(\alpha)}(\theta) = \begin{cases} q_\theta(\Gamma) - \pi(\Gamma) & \text{if } \alpha = 1, \\ q_\theta(\Gamma) - \pi_\theta^{(\alpha)}(\Gamma) & \text{if } \alpha \neq 1. \end{cases}$$

Similarly, for any $\theta \in \operatorname{int}\Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$,

$$\nabla^2 f_\pi^{(\alpha)}(\theta) = \begin{cases} \nabla^2 A(\theta) & \text{if } \alpha = 1, \\ \nabla^2 A(\theta) + (\alpha - 1)\left(\pi_\theta^{(\alpha)}(\Gamma\Gamma^\top) - \pi_\theta^{(\alpha)}(\Gamma)(\pi_\theta^{(\alpha)}(\Gamma))^\top\right) & \text{if } \alpha \neq 1. \end{cases}$$

We now give the definitions of the gradient descent operator for $f_\pi^{(\alpha)}$, of the proximal operator for r, and of the proximal gradient operator for $F_\pi^{(\alpha)} = f_\pi^{(\alpha)} + r$, all within the Bregman metric induced by the log-partition function A.

Definition 7. Consider a positive step-size $\tau > 0$.

(i) The *Bregman proximal operator* of τr is defined as

$$\operatorname{prox}_{\tau r}^A(\theta) := \underset{\theta' \in \operatorname{dom} A}{\arg\min} \left(r(\theta') + \frac{1}{\tau} d_A(\theta', \theta) \right), \forall \theta \in \operatorname{int}\operatorname{dom} A.$$

(ii) When $\nabla A(\theta) - \tau \nabla f_\pi^{(\alpha)}(\theta) \in \operatorname{dom} \nabla A^*$ for every $\theta \in \operatorname{int}\operatorname{dom} A$, the *Bregman gradient descent operator* of $\tau f_\pi^{(\alpha)}$ is well-defined and reads

$$\gamma_{\tau f_\pi^{(\alpha)}}^A(\theta) := \nabla A^* \left(\nabla A(\theta) - \tau \nabla f_\pi^{(\alpha)}(\theta) \right), \forall \theta \in \operatorname{int}\operatorname{dom} A.$$

(iii) The *Bregman proximal gradient operator* of $\tau F_\pi^{(\alpha)}$ is defined by

$$T_{\tau F_\pi^{(\alpha)}}^A(\theta) := \underset{\theta' \in \operatorname{dom} A}{\arg\min} \left(r(\theta') + \langle \nabla f_\pi^{(\alpha)}(\theta), \theta' - \theta \rangle + \frac{1}{\tau} d_A(\theta', \theta) \right), \forall \theta \in \operatorname{int}\operatorname{dom} A.$$

Next, we show that Algorithm 1 is a Bregman proximal gradient algorithm relying on the divergence d_A and that it is well-posed, which brings useful links between statistics, Bregman divergences, and optimization. To do so, let us introduce technical assumptions under which the operators $\gamma_{\tau f_\pi^{(\alpha)}}^A$ and $\operatorname{prox}_{\tau r}^A$ from Definition 7 are well-defined, single-valued, and mapping the set $\operatorname{int}\Theta$ to itself.

Assumption 2. For any $\theta \in \operatorname{int}\operatorname{dom} A$, $\pi_\theta^{(\alpha)}(\Gamma) \in \operatorname{int}\operatorname{dom} A^*$. Equivalently, there exists $\theta^{(\alpha)} \in \operatorname{int}\Theta$ such that $\pi_\theta^{(\alpha)}(\Gamma) = q_{\theta^{(\alpha)}}(\Gamma)$.

In the case where $\alpha = 1$ and $\mathcal{Q} = \mathcal{G}$, Assumption 2 is equivalent to the target π having finite first and second order moments.

Assumption 3. The regularizer r is in $\Gamma_0(\mathcal{H})$, is bounded from below, and is such that $\operatorname{int}\Theta \cap \operatorname{dom} r \neq \emptyset$.

This assumption is standard in the Bregman optimization literature [9], and allows in particular non-smooth regularizers. For instance, Assumption 3 is satisfied by the ℓ_1 norm often used to enforce sparsity [41, Section 3.4], or by indicator functions of non-empty closed convex sets, to impose constraints on the parameters.

We now show how Assumptions 1, 2, and 3 ensure the well-posedness of the operators introduced in Definition 7. We also define the stationary points of $F_\pi^{(\alpha)}$ and show that they coincide with the fixed points of the operators of Definition 7.

Definition 8. Under Assumption 3, we introduce for $\alpha > 0$ the set of *stationary points* of $F_\pi^{(\alpha)}$ as $S_\pi^{(\alpha)} := \{\theta \in \operatorname{int}\Theta \cap \operatorname{dom} f_\pi^{(\alpha)},\, 0 \in \nabla f_\pi^{(\alpha)}(\theta) + \partial r(\theta)\}$.

Proposition 7.

(i) Under Assumptions 1 and 2, if $\tau \in (0, 1]$, the operator $\gamma^A_{\tau f_\pi^{(\alpha)}}$ is well-defined on $\operatorname{int}\Theta$ and $\gamma^A_{\tau f_\pi^{(\alpha)}}(\theta) \in \operatorname{int}\Theta$ for every $\theta \in \operatorname{int}\Theta$.

(ii) Under Assumptions 1 and 3, the domain of $prox^A_{\tau r}$ is $\operatorname{int}\Theta$. On $\operatorname{int}\Theta$, $prox^A_{\tau r}$ is single-valued, and $prox^A_{\tau r}(\theta) \in \operatorname{int}\Theta$ for every $\theta \in \operatorname{int}\Theta$.

(iii) If Assumptions 1, 2, and 3 are satisfied, and $\tau \in (0, 1]$, $T^A_{\tau F_\pi^{(\alpha)}} = prox^A_{\tau r} \circ \gamma^A_{\tau f_\pi^{(\alpha)}}$, and a point $\theta \in \operatorname{int}\Theta$ is a fixed point of $T^A_{\tau F_\pi^{(\alpha)}}$ if and only if $\theta \in S_\pi^{(\alpha)}$.

We now state our main proposition, that provides an optimization-based interpretation for our Algorithm 1. Specifically, we show that Algorithm 1 consists first in a Bregman gradient descent step on $f_\pi^{(\alpha)}$ and then in a Bregman proximal step on the regularization function r, both within the Bregman geometry induced by log-partition function A.

Proposition 8. *Consider a sequence $\{\theta_k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 starting from $\theta_0 \in \operatorname{int}\Theta$. Under Assumptions 1, 2, and 3, for every $k \in \mathbb{N}$, $\theta_k, \theta_{k+\frac{1}{2}} \in \operatorname{int}\Theta$, and we can define equivalently the updates (6) and (7) as*

$$\theta_{k+\frac{1}{2}} = \gamma^A_{\tau_{k+1} f_\pi^{(\alpha)}}(\theta_k), \tag{14}$$

$$\theta_{k+1} = prox^A_{\tau_{k+1} r}\left(\theta_{k+\frac{1}{2}}\right). \tag{15}$$

Furthermore,

$$\theta_{k+1} = T^A_{\tau_{k+1} F_\pi^{(\alpha)}}(\theta_k). \tag{16}$$

Remark 3. Contrary to Algorithm 1, each iteration $k \in \mathbb{N}$ of Algorithm 2 resorts to an approximation of $\pi_{\theta_k}^{(\alpha)}(\Gamma)$. Recall from Proposition 6 that this quantity appears in $\nabla f_\pi^{(\alpha)}(\theta_k) = q_{\theta_k}(\Gamma) - \pi_{\theta_k}^{(\alpha)}(\Gamma)$. Therefore, Algorithm 2 uses a noisy approximation of $\nabla f_\pi^{(\alpha)}(\Gamma)$, that we denote by $\widetilde{G}_\pi^{(\alpha)}(\theta_k)$. Following the result of Proposition 8, which shows that Algorithm 1 is a Bregman proximal gradient algorithm, we can interpret Algorithm 2 as a stochastic Bregman proximal gradient algorithm [79], where

$$\theta_{k+1} = \operatorname{prox}^A_{\tau_{k+1} r}\left(\nabla A^*\left(\nabla A(\theta_k) - \tau_{k+1}\widetilde{G}_\pi^{(\alpha)}(\theta_k)\right)\right).$$

Note however that we do not guarantee here the well-posedness of this stochastic step. Let us mention the work of [4] where a stochastic proximal gradient algorithm, the gradient being approximated by sampling, is investigated in an Euclidean geometry.

4.2 Comparison with existing gradient descent algorithms

In the previous section, we interpret Algorithms 1 and 2 under the framework of Bregman proximal gradient algorithms. Let us use this perspective to explain the links between our algorithms, natural gradients methods, and black-box VI algorithms.

In the work of [58], an alternative objective that does not involve the unknown normalization constant Z_π is constructed from $\theta \longmapsto RD_\alpha(q_\theta, \pi)$. It is called the *variational Rényi bound* and plays a role akin to the evidence lower bound. This objective is then minimized using a stochastic gradient descent algorithm using samples from the proposals. We now explicit this algorithm when an exponential family is used for the proposals. Consider in the following $\alpha \in (0,1)$, and $\theta \in \operatorname{int}\Theta$. Then,

$$\begin{aligned} RD_{1-\alpha}(q_\theta, \pi) &= \frac{1-\alpha}{\alpha} RD_\alpha(\pi, q_\theta) \\ &= -\frac{1}{\alpha}\log\left(\int \pi(x)^\alpha q_\theta(x)^{1-\alpha}\nu(dx)\right) \\ &= -\frac{1}{\alpha}\log\left(\int \tilde{\pi}(x)^\alpha q_\theta(x)^{1-\alpha}\nu(dx)\right) + \log Z_\pi, \end{aligned}$$

where the first equality comes from [76, Proposition 2]. Therefore, minimizing $\theta \longmapsto RD_{1-\alpha}(q_\theta, \pi)$ is equivalent to maximizing

$$\mathcal{L}_\pi^{(\alpha)}(\theta) := \frac{1}{\alpha}\log\left(\int \tilde{\pi}(x)^\alpha q_\theta(x)^{1-\alpha}\nu(dx)\right). \tag{17}$$

Note that, as pointed in [58, Theorem 1], $\mathcal{L}_\pi^{(\alpha)}(\theta) \xrightarrow[\alpha\to 1]{} \log Z_\pi$, and $\mathcal{L}_\pi^{(\alpha)} \leq \log Z_\pi$ for $\alpha \leq 1$, meaning that the marginal likelihood is recovered for $\alpha = 1$.

Now, following computations as in Proposition 6, we obtain $\nabla \mathcal{L}_\pi^{(\alpha)}(\theta) = -\frac{1-\alpha}{\alpha}\nabla f_\pi^{(\alpha)}$. Therefore, the gradient ascent algorithm to maximize $\mathcal{L}_\pi^{(\alpha)}$ on Θ reads $\theta_{k+1} = \theta_k - \tau_{k+1}\nabla f_\pi^{(\alpha)}(\theta_k)$ where the factor $\frac{1-\alpha}{\alpha}$ is absorbed by the step-size.

Hence, the exact implementation of the VRB algorithm appears as an Euclidean analogue of Algorithm 1. In the black-box setting, the quantities $\pi_\theta^{(\alpha)}(\Gamma)$ are approximated at iteration $k \in \mathbb{N}$ using samples from q_{θ_k}, as it is done for Algorithm 2, leading to the VRB update

$$\theta_{k+1} = \theta_k + \tau_{k+1}\left(\sum_{l=1}^{N_{k+1}} \bar{w}_l^{(\alpha)}\Gamma(x_l) - q_{\theta_k}(\Gamma)\right), \tag{18}$$

with weights $\{\bar{w}_l^{(\alpha)}\}_{l=1}^{N_{k+1}}$ computed as in Algorithm 2.

5 Convergence analysis

In this section, we analyze the convergence of Algorithm 1. We rely on its interpretation as a Bregman proximal gradient algorithm from Section 4.1. We explain in Section 5.1 in which sense the Bregman geometry induced by the KL divergence is well-adapted to handle Problem ($P_\pi^{(\alpha)}$). Convergence results are given in Section 5.2 and are compared with existing results in Section 5.3. The proofs can be found in Appendices A-B.

5.1 Properties of Problem ($P_\pi^{(\alpha)}$)

We start by introducing the notions of *relative smoothness* and *relative strong convexity*, which generalize the Euclidean notions of smoothness and strong convexity to the Bregman setting. In the Euclidean setting, having an objective function that satisfies these two notions is desirable to construct efficient algorithms. When these properties are not satisfied, this may indicate that the Euclidean metric is not the best metric to handle the problem and encourages a switch to more adapted Bregman divergences.

Definition 9. Consider a Legendre function B and a differentiable function f.

(i) We say that f is L-relatively smooth with respect to B if there exists $L \geq 0$ such that

$$f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq L d_B(\theta, \theta'),\ \forall(\theta, \theta') \in (\operatorname{dom} B) \times (\operatorname{int} \operatorname{dom} B).$$

(ii) Similarly, we say that f is ρ-relatively strongly convex with respect to B is there exists $\rho \geq 0$ such that

$$\rho d_B(\theta, \theta') \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle,\ \forall(\theta, \theta') \in (\operatorname{dom} B) \times (\operatorname{int} \operatorname{dom} B).$$

These properties give indications about the relation between f and its *tangent approximation at* θ', defined by $\theta \longmapsto f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + L d_B(\theta, \theta')$, where L can be changed for ρ. This tangent approximation majorizes f in the case of relative smoothness, while it minorizes f in the case of relative strong convexity, as illustrated in Fig. 1. In both cases, f and its tangent approximation coincide at θ'.

In the Euclidean case $B(\cdot) = \frac{1}{2}\|\cdot\|^2$, the relative smoothness property is equivalent to the standard smoothness property, that is the Lipschitz continuity of the gradient, and relative strong convexity is equivalent to the strong convexity property [7, 40]. Note also that relative strong convexity implies convexity (which corresponds to $\rho = 0$ in the above). We explain now the interplay between the parameter α of the Rényi divergence and the above notions.

Proposition 9. *Let Assumption 1 be satisfied. The function $f_\pi^{(\alpha)}$, defined in (3), is 1-relatively smooth with respect to A, defined in (2), when $\alpha \in (0, 1]$. Similarly, the function $f_\pi^{(\alpha)}$ is 1-relatively strongly convex with respect to A when $\alpha \in [1, +\infty)$.*

In Proposition 9, the case $\alpha = 1$ plays a special role, as it is the only value for which we have both relative smoothness and relative strong convexity. Indeed, $f_\pi^{(1)}(\theta) = KL(\pi, q_\theta)$ and $d_A(\theta, \theta') = KL(q_{\theta'}, q_\theta)$, which gives the intuition that $f_\pi^{(1)}$ and d_A are functions with similar mathematical behaviors, leading to improved properties.

We now give a result about the existence of minimizers to Problem ($P_\pi^{(\alpha)}$). Again, this result highlights different behaviors depending on the value of α (i.e., if it is lower, equal or higher than one).

Proposition 10. *Let $\alpha > 0$.*

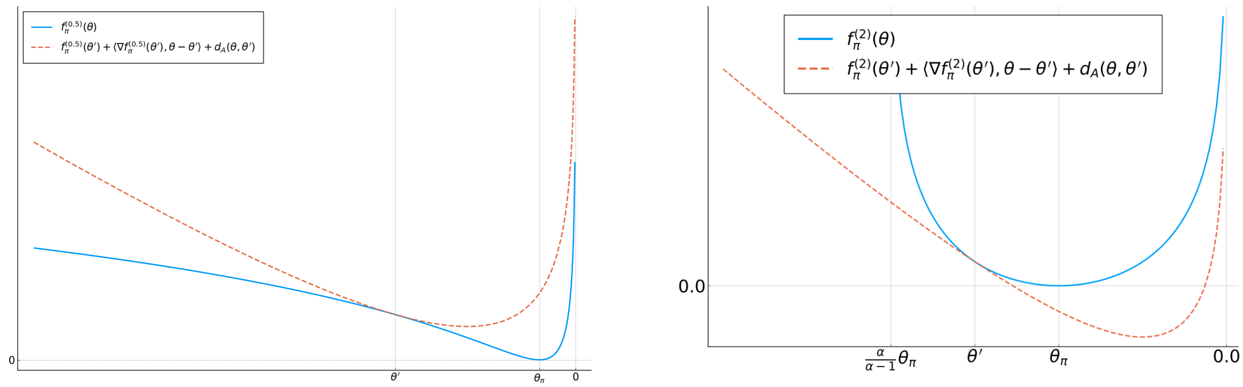
(i) Under Assumptions 1 and 3, the objective function $F_\pi^{(\alpha)}$ is proper (i.e., with nonempty domain), lower semicontinuous, and bounded from below, that is

$$-\infty < \vartheta_\pi^{(\alpha)} := \inf_{\theta \in \Theta} F_\pi^{(\alpha)}(\theta).$$

(ii) If $\alpha \geq 1$ and Assumptions 1, 2, and 3 are satisfied, then $F_\pi^{(\alpha)}$ is coercive and there exists $\theta_* \in \Theta$ such that $F_\pi^{(\alpha)}(\theta_*) = \vartheta_\pi^{(\alpha)}$. Further, it is unique and in $\operatorname{int} \Theta$.

We now introduce an elementary one-dimensional exponential family that we use to illustrate the notions of relative smoothness and strong convexity. We will also use this family to construct counter-examples to various claims.

Example 4. The family of one-dimensional centered Gaussian distributions with variance σ^2 is an exponential family. We denote this family by $\mathcal{G}_0^1$ in the following. It is an exponential family with parameter $\theta = -\frac{1}{2\sigma^2}$ and sufficient statistics $\Gamma(x) = x^2$. Its log-partition function is $A(\theta) = \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(-2\theta)$, whose domain is $\Theta = \mathbb{R}_{--}$.



(a) Relative smoothness illustrated in the case $\alpha = 0.5$ (b) Relative strong convexity shown in the case $\alpha = 2$

Figure 1: Plots of $f_\pi^{(\alpha)}$ and the tangent approximations described in Definition 9, obtained by choosing $\mathcal{Q} = \mathcal{G}_0^1$ and $\pi \in \mathcal{G}_0^1$ equal to some q_{θ_π}.

Figure 1 illustrates the results of Proposition 9 when the exponential family is the family of centered one-dimensional Gaussians $\mathcal{G}_0^1$ and the target as well belongs to this family. One can see that, when $\alpha \leq 1$, relative smoothness is satisfied and $f_\pi^{(\alpha)}$ is above its tangent approximation. On the contrary, $\alpha \geq 1$ yields relative strong convexity, ensuring that $f_\pi^{(\alpha)}$ is above its tangent approximation.

We now give a result about potential failures of the Euclidean smoothness of $f_\pi^{(\alpha)}$. This suggests that the Euclidean metric is not well-suited to minimize $f_\pi^{(\alpha)}$.

Proposition 11. *There exist targets π and exponential families $\mathcal{Q}$ such that the gradient of $f_\pi^{(\alpha)}$ is not Lipschitz on* $\operatorname{dom} f_\pi^{(\alpha)}$*, for $\alpha > 0$.*

Remark 4. The complete proof is in Appendix A. Let us exhibit counter-examples built by using $\mathcal{Q} = \mathcal{G}_0^1$ and targets $q_{\theta_\pi} \in \mathcal{G}_0^1$. Recall that $(f_\pi^{(\alpha)})'$ is Lipschitz continuous on its domain if and only if $(f_\pi^{(\alpha)})''$ is bounded on its domain. In our setting, we have

$$\operatorname{dom} f_\pi^{(\alpha)} = \begin{cases} \Theta & \text{if } \alpha \leq 1, \\ (\frac{\alpha}{\alpha-1}\theta_\pi, 0) & \text{if } \alpha > 1, \end{cases}$$

and $|(f_\pi^{(\alpha)})''(\theta)| \to +\infty$ when $\theta \to 0$, and also when $\theta \to \frac{\alpha}{\alpha-1}\theta_\pi$ for the case $\alpha > 1$.

The counter-example used in the proof of Proposition 11 illustrates why choosing to work in the Bregman geometry induced by A can be beneficial. Indeed, when $\alpha \in (0, 1]$, we have relative smoothness from Proposition 9, while Euclidean smoothness fails. In this case, Euclidean smoothness might be recovered if we restricted $f_\pi^{(\alpha)}$ to some set of the form $[\epsilon, +\infty)$. However, this would create a risk of excluding the target value θ_π.

This counter-example is also a case where Assumption 1 (i) fails for $\alpha > 1$ since $\operatorname{dom} f_\pi^{(\alpha)}$ is strictly included in Θ. One could also restrict the search to a smaller set, but the upper bound of $\operatorname{dom} f_\pi^{(\alpha)}$ would depend on the target true parameters. This prevents from restricting the admissible values of θ in a meaningful way without tight knowledge on the target. Note also that the family $\mathcal{G}_0^1$ has a log-partition function A that is not strongly convex. Finally the family $\mathcal{G}_0^1$ also allows us to show that, even for a log-concave target $\pi \in \mathcal{G}_0^1$, the objective function $f_\pi^{(\alpha)}$ might not be convex, as illustrated in Figure 1a.

Figure 1a shows a situation where the function $f_\pi^{(\alpha)}$ is not convex, but has a unique stationary point, which is the global minimizer. We show now that this situation is implied by having $\pi = q_{\theta_\pi}$ for some $\theta_\pi \in \operatorname{int}\Theta$ and lead to further results.

Proposition 12. *Suppose that Assumption 1 is verified and that there exists $\theta_\pi \in \operatorname{int}\Theta$ with $\pi = q_{\theta_\pi}$. Then Assumption 2 is verified, and the function $f_\pi^{(\alpha)}$ has a unique minimizer, which is θ_π and which is also its only stationary point. Moreover, $\vartheta_\pi^{(\alpha)} = f_\pi^{(\alpha)}(\theta_\pi) = 0$.*

Proposition 12 shows that when $r \equiv 0$ and $\pi = q_{\theta_\pi}$ with $\theta_\pi \in \operatorname{int}\Theta$, the stationary point of Problem ($P_\pi^{(\alpha)}$) is unique and equal to the global minimizer of this problem. The next proposition investigate the behavior of $f_\pi^{(\alpha)}$ around its minimizer θ_π.

Proposition 13. *Consider $\theta \in \operatorname{int}\Theta$ in a neighborhood of θ_π of the form $B(\theta_\pi, \upsilon)$ for some $\upsilon > 0$. Then, for any $\alpha > 0, \alpha \neq 1$, we have that*

(i) $f_\pi^{(\alpha)}$ has a quadratic behavior in the neighborhood of θ_π i.e.,

$$f_\pi^{(\alpha)}(\theta) = \frac{\alpha}{2}\|\theta - \theta_\pi\|^2_{\nabla^2 A(\theta_\pi)} + o(\upsilon^2),$$

(ii) $f_\pi^{(\alpha)}$ satisfies a Polyak-Łojasiewicz inequality around θ_π:

$$f_\pi^{(\alpha)}(\theta) \leq \frac{1}{2\alpha}\|\nabla f_\pi^{(\alpha)}(\theta)\|^2_{\nabla^2 A^*(\nabla A(\theta_\pi))} + o(\upsilon^2).$$

Proposition 13 shows that, in the neighborhood of its minimizer, function $f_\pi^{(\alpha)}$ has a quadratic behavior, hence generalizing a known result in the case $\alpha = 1$ [2]. It also shows that $f_\pi^{(\alpha)}$ satisfies a type of Polyak-Łojasiewicz inequality around the point θ_π. This type of condition has been used to prove geometric rates of convergence to minimizers for a variety of optimization algorithm, while being weaker than strong convexity [50].

5.2 Convergence analysis of Algorithms 1 and 2

We are now ready to present our convergence results for Algorithm 1. We give a first set of results for values of α in $(0, 1]$, and then stronger results when $\alpha = 1$. Results for $\alpha \in (0, 1]$ only exploit the relative smoothness, while the results for $\alpha = 1$ rely on the relative smoothness and the relative strong convexity of $f_\pi^{(1)}$.

We now give our convergence results for Algorithm 1 for $\alpha \in (0, 1]$.

Proposition 14. *Consider a sequence $\{\theta_k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 from $\theta_0 \in \operatorname{int}\Theta$, with $\alpha \in (0, 1]$ and a sequence of step-sizes $\{\tau_k\}_{k\in\mathbb{N}}$ such that $\tau_k \in (0, 1]$. Under Assumptions 1, 2, and 3,*

(i) the sequence $\{F_\pi^{(\alpha)}(\theta_k)\}_{k\in\mathbb{N}}$ is non-increasing,

(ii) if $F_\pi^{(\alpha)}(\theta_{K+1}) = F_\pi^{(\alpha)}(\theta_K)$ for some $K \in \mathbb{N}$, then $\theta_k = \theta_K$ for every $k \geq K$ and θ_K is a stationary point of $F_\pi^{(\alpha)}$,

(iii) $\sum_{k\geq 0} KL(q_{\theta_k}, q_{\theta_{k+1}}) < +\infty$,

(iv) *let* $K \in \mathbb{N}$ *the first iterate such that* $d_A(\theta_K, \theta_{K+1}) \leq \varepsilon$ *for some* $\varepsilon > 0$*. Then* K *is at most equal to* $\frac{1}{\varepsilon}\left(F_\pi^{(\alpha)}(\theta_0) - \vartheta_\pi^{(\alpha)}\right)$.

(v) *if in addition, there exists a non-empty compact set* $C \subset \operatorname{int}\Theta$ *such that* $\theta_k \in C$ *for every* $k \in \mathbb{N}$ *and* r *is continuous on* C*, and a scalar* $\epsilon > 0$ *such that* $\tau_k \in [\epsilon, 1]$ *for every* $k \in \mathbb{N}$*, then every converging subsequence of* $\{\theta_k\}_{k\in\mathbb{N}}$ *converges to a point in* $S_\pi^{(\alpha)}$.

The additional assumption used for Proposition 14(v) is satisfied for instance if $r = \iota_C$, for a compact $C \subset \operatorname{int}\Theta$. The continuity assumption on r is also satisfied by the ℓ_1 norm. In this case, r is also coercive, ensuring that the iterates stay in a compact set. However, this does not ensure that the iterates do not approach the boundary of Θ. Note that when $S_\pi^{(\alpha)} = \{\theta_s\}$ for some $\theta_s \in \operatorname{int}\Theta$, Prop. 14(v) implies that $\theta_k \xrightarrow[k\to+\infty]{} \theta_s$.

We now refine the result of Proposition 14 in the case $\alpha = 1$. In this case, the function $f_\pi^{(\alpha)}$ is also relatively strongly convex and coercive, two properties that are used to give stronger results, including rates of convergence to the global minimizer.

Proposition 15. *Consider a sequence* $\{\theta_k\}_{k\in\mathbb{N}}$ *generated by Algorithm 1 from* $\theta_0 \in \operatorname{int}\Theta$*, with* $\alpha = 1$ *and a sequence of step-sizes* $\{\tau_k\}_{k\in\mathbb{N}}$ *such that* $\tau_k \in [\epsilon, 1]$ *for some* $\epsilon > 0$*. Consider the point* θ_* *defined in Proposition 10. Under Assumptions 1, 2, and 3,*

(i) *the sequence* $\{KL(q_{\theta_k}, q_{\theta_*})\}_{k\in\mathbb{N}}$ *is non-increasing and*

$$KL(q_{\theta_k}, q_{\theta_*}) \leq (1-\epsilon)^k KL(q_{\theta_0}, q_{\theta_*}), \quad \forall k \in \mathbb{N},$$

(ii) *we have that* $F_\pi^{(1)}(\theta_k) \xrightarrow[k\to+\infty]{} F_\pi^{(1)}(\theta_*) = \vartheta_\pi^{(1)}$ *and that*

$$F_\pi^{(1)}(\theta_k) - F_\pi^{(1)}(\theta_*) \leq \frac{(1-\epsilon)^k}{\epsilon} KL(q_{\theta_0}, q_{\theta_*}), \quad \forall k \in \mathbb{N},$$

(iii) *the iterates converge to the solution,* $\theta_k \xrightarrow[k\to+\infty]{} \theta_*$.

We now present a specialized result for the case $\alpha \in (0,1)$, $r \equiv 0$, under the assumption that $\pi = q_{\theta_\pi}$ for some $\theta_\pi \in \operatorname{int}\Theta$. In this case, we are able to derive convergence results that are similar to the case $\alpha = 1$.

Proposition 16. *Consider a sequence* $\{\theta_k\}_{k\in\mathbb{N}}$ *generated by Algorithm 1 from* $\theta_0 \in \operatorname{int}\Theta$*, with* $\alpha \in (0,1)$ *and a sequence of step-sizes* $\{\tau_k\}_{k\in\mathbb{N}}$ *such that* $\tau_k \in [\epsilon, 1]$ *for some* $\epsilon > 0$*. Assume that the iterates stay in a non-empty compact set* $C \subset \operatorname{int}\Theta$*, that* $r \equiv 0$*, that there exists* $\theta_\pi \in \operatorname{int}\Theta$ *such that* $\pi = q_{\theta_\pi}$*, and that Assumption 1 is satisfied. Then,*

(i) $RD_\alpha(\pi, q_{\theta_k}) \xrightarrow[k\to+\infty]{} 0$ *and there exist constants* $C > 0$ *and* $\delta \in (0,1)$ *such that*

$$RD_\alpha(\pi, q_{\theta_k}) \leq C(1-\alpha\delta\epsilon)^k RD_\alpha(\pi, q_{\theta_0}), \quad \forall k \in \mathbb{N},$$

(ii) $\theta_k \xrightarrow[k\to+\infty]{} \theta_\pi$.

5.3 Discussion

Proposition 14 implies a monotonic decrease of $F_\pi^{(\alpha)}$ along iterations of Algorithm 1. This kind of result appears in many statistical procedures [31, 27, 28]. Note that these works encompass more general approximating families than our study, but do not consider our additional regularization term r. In our setting, we are moreover able to give novel and more precise results on the convergence of the sequence of iterates. The result of Proposition 14 (iii), which is a type of *finite length* property of the sequence of iterates, is not common for a statistical procedure, to our knowledge. This result can be used to practically assess the convergence of our algorithms as the condition $KL(q_{\theta_k}, q_{\theta_{k+1}}) \leq \varepsilon$ can be employed as a stopping criterion in Algorithms 1 and 2. Proposition 14 (iv) provides estimates on the number of iterations needed to reach a certain level of stationarity between iterates, while Proposition 14 (v) establishes convergence to the set of stationary points.

We are also able to show the geometric rate of convergence to the global minimizer of Problem ($P_\pi^{(\alpha)}$) when $\alpha = 1$ in Proposition 15 and when $\pi \in \mathcal{Q}$ and $r \equiv 0$ for $\alpha < 1$ in Proposition 16. Note that the result for $\alpha = 1$ is established under minimal assumptions on π as we only need $\pi(\Gamma)$ to be well-defined (see Assumption 2). In comparison, similar rates of convergence are established in the case of the objective function $KL(\cdot, \pi)$ in [57, 80] under strong log-concavity and log-smoothness assumptions on π. We are not aware of any VI algorithm achieving geometric rates in the case of the Rényi divergence. Let us however mention that, in the context of sampling algorithms, a geometric convergence of the probability distribution of the samples to the minimizer of $RD_\alpha(\cdot, \pi)$ for $\alpha \geq 1$ is proven by [78] under log-smoothness assumption on the target and a weaker version of log-concavity. It is difficult to compare the assumption $\pi \in \mathcal{Q}$ with log-concavity or log-smoothness assumptions, as some exponential families might have multi-modal members or can be written over a discrete space $\mathcal{X}$.

More generally, we avoided in our analysis any assumption that would not be satisfied by the one-dimensional Gaussian target described in Example 4. Therefore, we are facing a situation where $\nabla f_\pi^{(\alpha)}$ is not Lipschitz, A is not strongly convex, and $\operatorname{dom} A$ is not closed, which contrasts with common assumptions from the literature on optimization schemes based on Bregman divergences [7, 74, 14, 37, 40] or in the statistical literature [1, 51, 58, 18]. Note that the Euclidean smoothness of $KL(\cdot, \pi)$ is proven for instance by [30, 57] under a log-smoothness assumption on the target. However, when Euclidean smoothness is not satisfied and a standard gradient descent method is used, tuning the step-size, or learning rate, cannot be done using the Lipschitz constant of the gradients. In Section 6, we compare the VRB method of [58], which can be seen as an Euclidean counterpart to our Algorithm 2 (see Section 4.2). We show that the lack of information about the Lipschitz constant creates instabilities and poor performance, in contrast to our proposed method where the step-sizes can be chosen following the results presented in Propositions 14 and 15.

Our convergence analysis is restricted to $\alpha \in (0, 1]$, which is also the case in the analysis of [28], considering the minimization of the α-divergence D_α over wider families. The convergence proof techniques used in this work actually share some common points with ours. In particular, because of the 1-relative smoothness of $f_\pi^{(\alpha)}$ with respect to A, we have from Definition 9 that

$$f_\pi^{(\alpha)}(\theta) - f_\pi^{(\alpha)}(\theta') \leq \langle q_{\theta'}(\Gamma) - \pi_{\theta'}^{(\alpha)}(\Gamma), \theta - \theta' \rangle + KL(q_{\theta'}, q_\theta). \tag{19}$$

This is to be compared with [28, Proposition 1], which, in our setting, would read

$$\Psi_\pi^{(\alpha)}(\theta) - \Psi_\pi^{(\alpha)}(\theta') \leq -\frac{1}{\alpha} \int \pi(x)^\alpha q_{\theta'}(x)^{1-\alpha} \log\left(\frac{q_\theta(x)}{q_{\theta'}(x)}\right) \nu(dx). \tag{20}$$

Note that here, $q_\theta, q_{\theta'}$ are not necessarily from an exponential family and that we used $\Psi_\pi^{(\alpha)}(\theta) = D_\alpha(\pi, q_\theta)$, while $D_\alpha(q_\theta, \pi)$ was considered by [28] (this does not affect the results as $D_\alpha(\pi, q_\theta) = D_{1-\alpha}(q_\theta, \pi)$ for

$\alpha \in [0,1]$). When q_θ and $q_{\theta'}$ are in an exponential family $\mathcal{Q}$, Eq. (20) can be further rewritten as

$$\Psi_\pi^{(\alpha)}(\theta) - \Psi_\pi^{(\alpha)}(\theta') \leq \frac{Z_{\pi_{\theta'}^{(\alpha)}}}{\alpha} \left(\langle q_{\theta'}(\Gamma) - \pi_{\theta'}^{(\alpha)}(\Gamma), \theta - \theta' \rangle + KL(q_{\theta'}, q_\theta) \right), \tag{21}$$

with $Z_{\pi_{\theta'}^{(\alpha)}} = \int \pi(x)^\alpha q_{\theta'}(x)^{1-\alpha} \nu(dx)$. We recognize now that the right-hand side of Eq. (21) is equal to the one of (19) up to a positive multiplicative constant. Even if [28, Proposition 1] is derived directly without using Bregman divergences, our analysis gives a geometric interpretation to it. Moreover, our interpretation allows to use the modern Bregman proximal gradient machinery, allowing to prove convergence results that are more precise while including the additional regularization term r. Indeed, the convergence result in [28] only shows a monotonic decrease of the objective without regularization, although in wider variational families.

6 Numerical experiments

In this section, we investigate the performance of our methods through numerical simulations in a black-box setting and compare them with existing algorithms. We focus our study on Algorithm 2, that we call the *relaxed moment-matching (RMM)* algorithm when $r \equiv 0$ and the *proximal relaxed moment-matching (PRMM)* otherwise. We also consider VRB algorithm from [58], whose implementation for an exponential family is described by Equation (18). It is shown in Section 4.2 that the VRB algorithm can be interpreted as an Euclidean version of our novel RMM algorithm. However, when $\alpha \in (0,1]$, $f_\pi^{(\alpha)}$ is not smooth relatively to the Euclidean distance (see Proposition 11) while it is smooth relatively to the Bregman divergence d_A (see Proposition 9). Therefore, the comparison between the RMM and PRMM algorithms with the VRB method might allow to assess the use of the Bregman divergence instead of the Euclidean distance on a numerical basis. We also use this comparison to assess the role of the regularizer, which is a feature of our approach, but not of the one of [58].

Additional numerical experiments are presented in Appendix D. In particular, the influence of the parameters α and τ and of the regularizer r is studied in Appendix D.1 using a Gaussian toy example. In Appendix D.2, we provide additional comparison between the RMM and the VRB algorithms. We now turn to a Bayesian regression task, which allows us to compare the RMM, PRMM and VRB algorithms on a realistic problem and understand better the interest of using the Bregman geometry. We also use this example to show how our PRMM algorithm allows to compensate for a misspecified prior by adding a regularizer.

We consider a problem of non-linear regression, where we try to infer a regression vector $\beta \in \mathbb{R}^{d+1}$ from J measurements $y \in \mathbb{R}^J$, $X \in \mathbb{R}^{J \times d}$ under Gaussian noise. The non-linearity mimics the effect of a neural network with one single hidden layer,

$$\Phi_\beta(x) = \phi\left(\sum_{i=1}^d \beta_i x_i + \beta_0\right), \forall x \in \mathbb{R}^d,$$

where $\beta = (\beta_i)_{0 \leq i \leq d+1} \in \mathbb{R}^{d+1}$ is the regression vector, with the component β_0 playing the role of the bias. The function ϕ is the activation function and is taken here as the sigmoid function

$$\phi(s) = \frac{1}{1+e^{-s}}, \forall s \in \mathbb{R}.$$

Given a ground truth vector $\overline{\beta} \in \mathbb{R}^{d+1}$, and a feature set X, we assume, for every $j \in \{1, \ldots, J\}$,

$$y_j \sim \mathcal{N}\left(y_j; \Phi_{\overline{\beta}}(X_j), \sigma^2\right),$$

with $X_{j,:}$ the j-th line of X, and $X_j = X_{j,:}^\top \in \mathbb{R}^d$. Assuming i.i.d. realizations, this leads to the likelihood expression for a given $\beta \in \mathbb{R}^{d+1}$,

$$p(y|\beta) = \prod_{j=1}^{J} \mathcal{N}\left(y_j; \Phi_\beta(X_j), \sigma^2\right).$$

Our goal is to explore the posterior distribution on β, $p(\beta|y) \propto p(y|\beta)p(\beta)$ where knowledge on the regression vector β is encoded in a prior density $p(\beta)$. In the following, we drop the dependence on the data, so that our target reads

$$\pi(\beta) := p(\beta|y) \text{ and } \tilde{\pi}(\beta) := p(\beta|y)p(\beta).$$

The RMM, PRMM, and VRB algorithm are tested on synthetic data. First, a regression vector $\overline{\beta}$ is sampled from a spike-and-slab distribution

$$p_0(\beta) = \mathcal{N}(\beta_0; 0.0, 1.0) \prod_{i=1}^{d} \left(\rho\delta_0(\beta_i) + (1-\rho)\mathcal{N}(\beta_i; 0.0, 1.0)\right).$$

which places a non-zero probability on β_i being zero, for $i \in [\![1, d]\!]$. Regression vectors are sampled until we find $\bar{\beta} \sim p_0$ with at least one zero and one non-zero component.

Then, for every $j \in [\![1, J]\!]$, we sample vectors X_j uniformly in the square $[-s, s]^d$ and draw the observation y_j as stated before. Test data $y^{\text{test}} \in \mathbb{R}^{J_{test}}$ and $X \in \mathbb{R}^{J_{test} \times d}$ are also generated in this manner. We consider a Gaussian prior on β, $p(\beta) = \mathcal{N}(\beta; 0, I)$.

Since $\overline{\beta}$ is not sampled from the prior $p(\beta)$, there is a mismatch between the data we feed the algorithms and the posterior model. In the following, we show that the choice of a suitable regularizer in our VI method can allow to cope with this issue.

We run experiments using the VRB and RMM algorithms, as well as the PRMM algorithm, using the family of Gaussian densities with diagonal covariance matrix, whose parametrization is detailed in Appendix C. For the PRMM algorith, we use the regularizer $r(\theta) = \eta\|\theta_1\|_1$ with $\eta \geq 0$. This can be understood as the Lagrangian relaxation [44] with multiplier $\eta \geq 0$ of the constraint $\sum_{i=1}^{d} \|\theta_1\|_1 \leq c$, for $c \geq 0$ such that the constrained set is non empty.

Our ℓ_1-like regularizer enforces sparsity on all the components of the mean μ, except the component μ_0. The aim is to mimic the sparse structure of $\overline{\beta}$ that was simulated from p_0. The computation of the corresponding Bregman proximal operator for this choice of r is detailed in Appendix C. This approach is related to the stochastic proximal gradient algorithm of [4], except that our algorithm exploits a Bregman geometry.

The algorithms are run for $K = 100$ iterations, with a constant number of samples $N = 500$. Two values of α are tested, namely, $\alpha = 1.0$ and $\alpha = 0.5$. The VRB algorithm is run with $\tau = 10^{-3}$ while the PRMM algorithm is run with $\tau = 10^{-1}$. These choices correspond to the most favorable step-size for each algorithm, as indicated by our experiments in Appendix D. The algorithms are run 10^3 times. We choose $\eta = 1.0$ in the following. In the subsequent experiments, we set $d = 5$, $J = 100$, $J^{\text{test}} = 50$, $\sigma^2 = 0.5$, and $s = 5.0$.

In order to asses the performance of the algorithm, we track the variational Rényi bound, defined Eq. (17), that is estimated at each iteration $k \in \mathbb{N}$ through

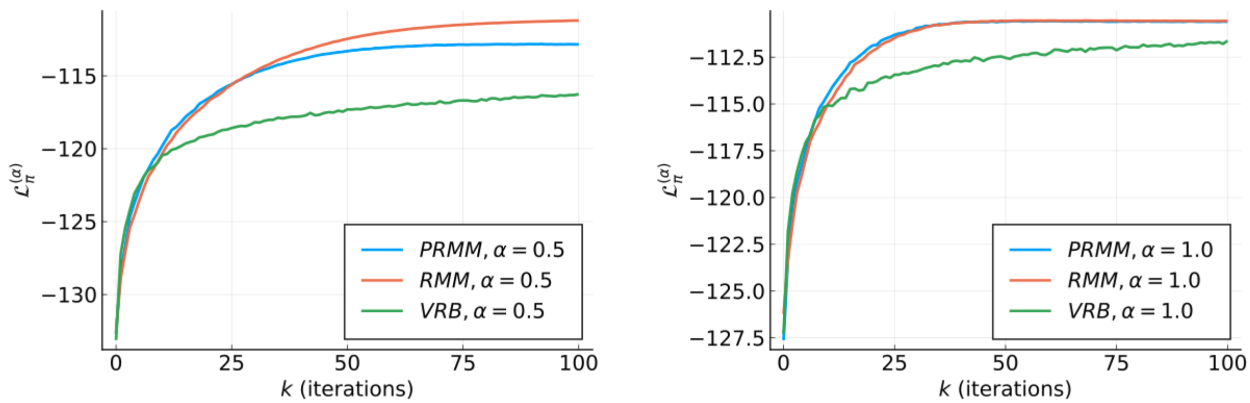
$$\mathcal{L}_\pi^{(\alpha)}(\theta_k) \approx \frac{1}{\alpha} \log\left(\frac{1}{N_{k+1}} \sum_{l=1}^{N_{k+1}} w_l^{(\alpha)}\right). \tag{22}$$

We also consider the F1 score that each algorithm achieves in the prediction of the zeros of the true regression vector $\overline{\beta}$. It is computed at each iteration $k \in \mathbb{N}$, by seeing how the zeros of μ_k match those of $\overline{\beta}$.

Additionally, since we provide not only a pointwise estimate of $\overline{\beta}$, but an approximation of the full target π, we also test the quality of the distributional approximation by sampling a regression vector β from the final proposal q_{θ_K}. This is done by computing

$$\mathrm{MSE}^{\mathrm{test}}(\beta) := \sum_{j=1}^{J^{\mathrm{test}}} \left(y_j^{\mathrm{test}} - \Phi_\beta(X_j^{\mathrm{test}}) \right)^2 .$$

By sampling N_β^{test} vectors $\beta \sim q_{\theta_K}$ and analyzing the distribution of the values $\{\mathrm{MSE}^{\mathrm{test}}(\beta_l)\}_{l=1}^{N_\beta^{\mathrm{test}}}$, we can get a sense of the quality of the approximated density q_{θ_K} in terms of both location and scale. At each run, the final distribution q_{θ_K} is tested by sampling $N_\beta^{\mathrm{test}} = 100$ values of β to assess the test error.



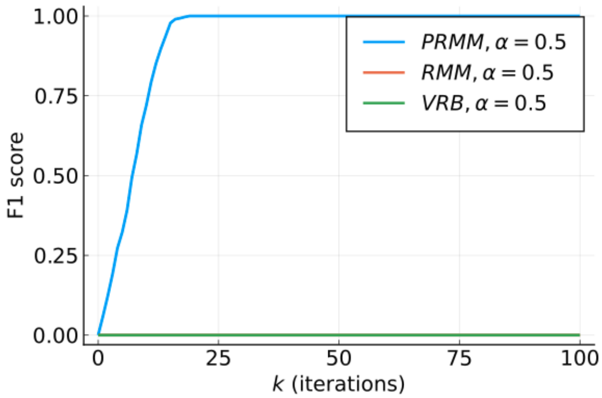
(a) $\alpha = 0.5$ (b) $\alpha = 1.0$

Figure 2: Approximated Rényi bound, averaged over 10^3 runs with $N = 500$ samples per iteration.

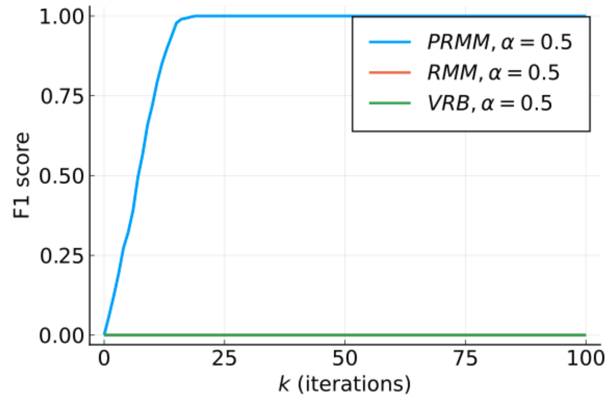
Figure 2 shows the increase of the approximated variational Rényi bound described in Eq. (22). As discussed in Section 4.2, an increase in the Rényi bound $\mathcal{L}_\pi^{(\alpha)}(\theta)$ shows a decrease in the Rényi divergence $RD_\alpha(\pi, q_\theta)$, so these plots show that the three method decrease the Rényi divergence. However, our methods are able to reach higher values at a faster rate than the VRB method, illustrating the improvement coming from using the Bregman geometry rather than the Euclidean one.

Figure 3 shows the F1 score achieved by each algorithm in the retrieval of the zeros of the true regression vector. The RMM and VRB algorithms are not able to recover any zeros, which is to be expected since they do not include any sparsity-inducing mechanisms. However, the PRMM algorithm is able to recover in this example the zero components of the regression vector in a few number of iterations and in most of the runs. Note also that it does not create false positives neither. This illustrates that adding a regularizer in the VI method itself can enforce sparsity although the prior of our model did not enforce it.

The box plots of Fig. 4 assess the quality of the variational approximation of the posterior obtained by each method, by evaluating how regression vectors sampled from the approximations are able to reconstruct the test data. We see that the PRMM and RMM algorithms yields reconstruction errors that are less spread and at a lower level than the ones coming from the VRB algorithm. This is in accordance with the plots of Fig. 2. This shows the higher performance coming from using a more adapted geometry. Note that errors are more spread for the PRMM algorithm than for the RMM algorithm. This may be due to the proximal step, which creates bigger eigenvalues for the covariance matrix (see Appendix C for details).



(a) $\alpha = 0.5$



(b) $\alpha = 1.0$

Figure 3: F1 score in the prediction of the zeros of $\overline{\beta}$ by the zeros of $\{\mu_k\}_{k=0}^K$, averaged over 10^3 runs with 500 samples per iteration.

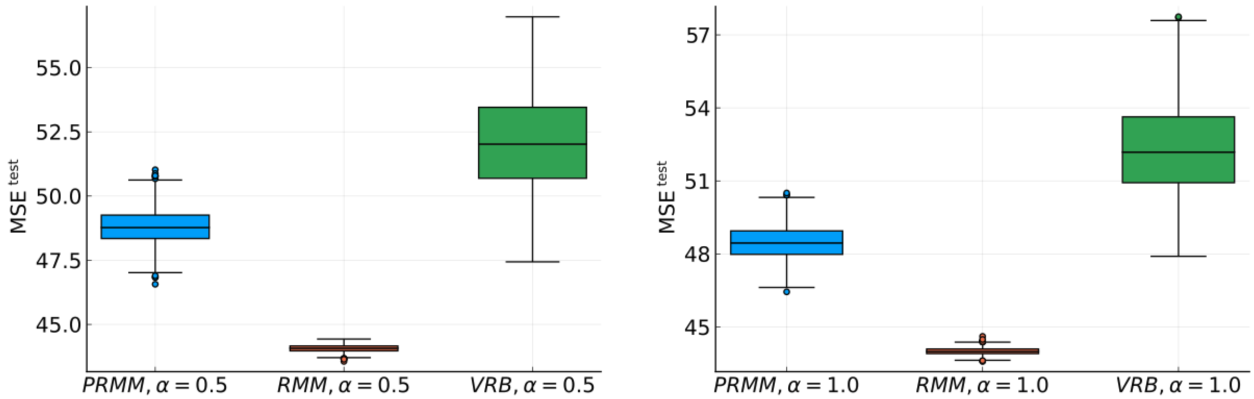
7 Conclusion and perspectives

We introduced in this work the proximal relaxed moment-matching algorithm, which is a novel VI algorithm minimizing the sum of a Rényi divergence and a regularizing function over an exponential family. We provided a black-box implementation which allows to bridge the gap between information-geometric VI methods and black-box VI algorithms, while generalizing several existing moment-matching algorithms. We also rewrote our algorithm as a Bregman proximal gradient algorithm whose Bregman divergence is equivalent to the Kullback-Leibler divergence.

Using this novel perspective, we established strong convergence guarantees for our exact algorithm. For $\alpha \in (0, 1]$, we established the monotonic decrease of the objective function, a finite-length property of the sequence of iterates, and subsequential convergence to a stationary point. In the particular case $\alpha = 1$, we also established the geometric convergence of the iterates towards the optimal parameters. We also exhibited a simple counter-example for which the corresponding Euclidean schemes may fail to converge, showing the necessity of resorting to an adapted geometry. These findings are backed by numerical results showing the versatility of our methods compared to more restricted moment-matching updates. Indeed, our parameters allow to tune the algorithms speed and robustness but also the features of the approximating densities. Comparison of our algorithms with their Euclidean counterparts also showed their robustness and good performance.

This confirmed the benefits of using a regularized Rényi divergence and the underlying geometry of exponential families, but also opened several research avenues.

First, although we proved the convergence of Algorithm 1, work remains to be done to establish the convergence of Algorithm 2. In particular, it would be interesting to understand the interplay between α, the step-sizes $\{\tau_k\}_{k\in\mathbb{N}}$ and the sample sizes $\{N_k\}_{k\in\mathbb{N}}$. Approximation errors can arise for a too low sample size, or when the proximal step is computed with limited precision. Analyzing the algorithm stability with error propagation quantification in such cases is another interesting challenge. Then, another venue of improvement would be the use of more complex optimization schemes, such as block updates or accelerated schemes. Variance reduction techniques as used in some black-box VI algorithms could also be used to improve our Algorithms. Finally, studying optimization schemes over mixtures of distributions from an exponential family could be a natural extension in order to tackle multimodal targets. Similarly, extending our analysis to values $\alpha > 1$ would allow to use the χ^2 divergence, which plays an important role for the



(a) $\alpha = 0.5$

(b) $\alpha = 1.0$

Figure 4: Box plots of the values $\mathrm{MSE}^{\text{test}}$, showing the reconstruction errors on the test data.

analysis of importance sampling schemes.

Acknowledgements

T.G. and E.C. acknowledge support from the ERC Starting Grant MAJORIS ERC-2019-STG-850925. The work of V. E. is supported by the *Agence Nationale de la Recherche* of France under PISCES (ANR-17-CE40-0031-01), the Leverhulme Research Fellowship (RF-2021-593), and by ARL/ARO under grant W911NF-22-1-0235.

Supplementary material

A Results about $F_\pi^{(\alpha)}$

A.1 Proof of Proposition 4

Proof of Proposition 4. The domain of A is non-empty by Assumption 1. Also, since $\int \exp(\langle\theta, \Gamma(x)\rangle)\nu(dx) > 0$ for any $\theta \in \Theta$, we have that $A(\theta) > -\infty$ for every θ in its domain, so A is proper. The set $\Theta = \mathrm{dom}\, A$ is convex, and the function A is lower semi-continuous on $\mathcal{H}$ and strictly convex on Θ by [15, Theorem 1.13]. The derivability property comes from [15, Theorem 2.2], and the expression of the gradient follows from simple computations.

Because of the steepness assumption on $\mathcal{Q}$, A is steep. With the differentiability properties of the above, this means that A is essentially smooth, showing that A is Legendre. □

A.2 Proof of Proposition 6

Proof of Proposition 6. For the case $\alpha = 1$, note that $f_\pi^{(1)}$ can be written as

$$f_\pi^{(1)}(\theta) = \int \log(\pi(x))\pi(x)\nu(dx) - \langle\theta, \pi(\Gamma)\rangle + A(\theta), \quad \forall\theta \in \Theta \cap \mathrm{dom}\, f_\pi^{(\alpha)}, \tag{23}$$

where $\Theta = \operatorname{dom} A$, and A defined in Eq. (2). The results come from the properties of A, given in Proposition 4.

We now turn to the case $\alpha \neq 1$. For every $\theta \in \Theta$, it is possible to decompose $f_\pi^{(\alpha)}$ as in

$$f_\pi^{(\alpha)}(\theta) = A(\theta) + \frac{1}{\alpha - 1} \log \left(\int \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx) \right).$$

where the functions $\tilde{h}$ and $\tilde{p}$ defined such that $\tilde{h}(\theta) = \int \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)$ and $\tilde{p}(x, \theta) = \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha}$ for any $\theta \in \operatorname{int} \Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$ and $x \in \mathcal{X}$.

We can show through standard results that at any $\theta \in \operatorname{int} \Theta$, the partial derivatives of $\tilde{h}$ of first and second order exist, are continuous and can be obtained by derivating under the integral sign. Since $\tilde{h}(\theta) > 0$ for all $\theta \in \Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$, and $f_\pi^{(\alpha)} = A + \frac{1}{\alpha-1} \log \circ \tilde{h}$, these results with those of Proposition 4 about A give the following. On $\operatorname{int} \Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$, the map $f_\pi^{(\alpha)}$ admits continuous first and second order partial derivatives that can be obtained by differentiating under the integral sign.

We now turn to the explicit derivation of the gradient $\nabla f_\pi^{(\alpha)}$ and the Hessian $\nabla^2 f_\pi^{(\alpha)}$, whose components are respectively the first and second order partial derivatives. Consider $\theta \in \operatorname{int} \Theta \cap \operatorname{dom} f_\pi^{(\alpha)}$. For $i \in [\![1, n]\!]$, we first compute

$$\frac{\partial \tilde{h}}{\partial \theta_i}(\theta) = (1 - \alpha) \int \Gamma_i(x) \pi(x)^\alpha q_\theta(x)^{1-\alpha} \nu(dx).$$

From there, we obtain

$$\frac{\partial f_\pi^{(\alpha)}}{\partial \theta_i}(\theta) = \frac{\partial A}{\partial \theta_i}(\theta) - \frac{\int \Gamma_i(x) \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)}{\int \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)}. \tag{24}$$

Since $q_\theta(x) = \exp(\langle \theta, \Gamma(x) \rangle) \exp(-A(\theta))$, we finally obtain that

$$\frac{\partial f_\pi^{(\alpha)}}{\partial \theta_i}(\theta) = \frac{\partial A}{\partial \theta_i}(\theta) - \pi_\theta^{(\alpha)}(\Gamma_i).$$

Because $\left(\nabla f_\pi^{(\alpha)}(\theta) \right)_i = \frac{\partial f_\pi^{(\alpha)}(\theta)}{\partial \theta_i}$, this concludes the computations about the gradient of $f_\pi^{(\alpha)}$.

Before computing the second order partial derivatives, we introduce another intermediate quantity. Denote $\tilde{g}_i : \theta \longmapsto \int \Gamma_i(x) \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)$ for $i \in [\![1, n]\!]$. In fact, $\tilde{g}_i(\theta) = \frac{1}{1-\alpha} \frac{\partial \tilde{h}}{\partial \theta_i}(\theta)$, and from Eq. (24), we have $\frac{\partial f_\pi^{(\alpha)}}{\partial \theta_i}(\theta) = \frac{\partial A}{\partial \theta_i}(\theta) - \frac{\tilde{g}_i(\theta)}{\tilde{h}(\theta)}$. We also compute for any $j \in [\![1, n]\!]$

$$\frac{\partial \tilde{g}_i}{\partial \theta_j}(\theta) = (1 - \alpha) \int \Gamma_j(x) \Gamma_i(x) \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx).$$

Using those intermediate results, we obtain for $i, j \in [\![1, n]\!]$ that

$$\frac{\partial^2 f_\pi^{(\alpha)}}{\partial \theta_j \partial \theta_i}(\theta) = \frac{\partial^2 A}{\partial \theta_j \partial \theta_j}(\theta) + (\alpha - 1) \left(\pi_\theta^{(\alpha)}(\Gamma_i \Gamma_j) - \pi_\theta^{(\alpha)}(\Gamma_i) \pi_\theta^{(\alpha)}(\Gamma_j) \right).$$

We conclude about the Hessian by using that $(\nabla^2 f_\pi^{(\alpha)}(\theta))_{i,j} = \frac{\partial^2 f_\pi^{(\alpha)}}{\partial \theta_j, \partial \theta_i}(\theta)$.

$\square$

A.3 Proof of Proposition 7

Proof of Proposition 7.

(i) Since A is Legendre, A^* is also Legendre from Proposition 2, so in particular $\operatorname{dom} A^*$ is convex. This implies that $\operatorname{int}\operatorname{dom} A^*$ is convex. Consider $\theta \in \operatorname{int}\Theta$, then $q_\theta(\Gamma) = \nabla A(\theta) \in \operatorname{int}\operatorname{dom} A^*$. Since by assumption, $\pi_\theta^{(\alpha)}(\Gamma) \in \operatorname{int}\operatorname{dom} A^*$ and the step-size $\tau \in (0, 1]$, then

$$\nabla A(\theta) - \tau \nabla f_\pi^{(\alpha)} = \tau \pi_\theta^{(\alpha)}(\Gamma) + (1-\tau) q_\theta(\Gamma) \in \operatorname{int}\operatorname{dom} A^*.$$

This shows the well-posedness of $\gamma^A_{\tau f_\pi^{(\alpha)}}$. Using results from Proposition 2, this also implies that $\gamma^A_{\tau f_\pi^{(\alpha)}} \in \operatorname{dom}\nabla A = \operatorname{int}\Theta$.

(ii) We conclude about the proximal operator with [9, Proposition 3.21 (vi)], which ensures that $\operatorname{dom}\operatorname{prox}^A_{\tau r} = \operatorname{int}\Theta$, with [9, Proposition 3.23 (v)] which ensures that $\operatorname{ran}\operatorname{prox}^A_{\tau_{k+1} r} \subset \operatorname{int}\operatorname{dom} A$, and with [9, Proposition 3.22 (ii)(d)], showing that $\operatorname{prox}^A_{\tau r}$ is single-valued.

(iii) The third point comes from [37, Lemma 3]. □

A.4 Proof of Proposition 8

Proof of Proposition 8. Every operation is well-defined because of Proposition 7. We now show the equivalence between the moment-matching step (6) and its reformulation (14). From Assumption 1, and Proposition 6, $f_\pi^{(\alpha)}$ is differentiable on $\operatorname{int}\Theta$ and its gradient is $\nabla f_\pi^{(\alpha)}(\theta) = q_\theta(\Gamma) - \pi_\theta^{(\alpha)}(\Gamma)$. Using that $\nabla A(\theta) = q_\theta(\Gamma)$ from Proposition 4 and that $(\nabla A)^{-1} = \nabla A^*$ from Proposition 2, it comes that (6) reads

$$\begin{aligned}\theta_{k+\frac{1}{2}} &= \nabla A^*\left(\tau_{k+1}\pi_{\theta_k}^{(\alpha)}(\Gamma) + (1-\tau_{k+1}) q_{\theta_k}(\Gamma)\right)\\ &= \nabla A^*\left(\nabla A(\theta_k) - \tau_{k+1}\nabla f_\pi^{(\alpha)}(\theta_k)\right),\end{aligned}$$

which shows the result.

Equation (15) is straightforward, and comes from the equivalence between d_A and the KL divergence stated in Proposition 5. Finally, Eq. (16) comes from the two previous points and Proposition 7 (iii). □

A.5 Proof of Proposition 9

Proof of Proposition 9. We prove relative smoothness and relative strong convexity by using the alternative characterizations given in [40, Proposition 2.2] and [40, Proposition 2.3]. $f_\pi^{(\alpha)}$ and A are twice differentiable on $\operatorname{int}\Theta$, so thanks to these results, $f_\pi^{(\alpha)}$ is L-relatively smooth with respect to A if and only if $\nabla^2 f_\pi^{(\alpha} \preccurlyeq L\nabla^2 A$, on $\operatorname{int}\Theta$, and it is ρ-relatively strongly convex with respect to A if and only if $\rho\nabla^2 A \preccurlyeq \nabla^2 f_\pi^{(\alpha}$ on $\operatorname{int}\Theta$.

We first cover the case $\alpha = 1$. In this case, we have that for every $\theta \in \operatorname{int}\Theta$, $\nabla^2 f_\pi^{(1)}(\theta) = \nabla^2 A(\theta)$ from Proposition 6. Therefore, the functions $f_\pi^{(1)} - A$ and $A - f_\pi^{(1)}$ have null Hessian on $\operatorname{int}\Theta$, showing that they are convex, hence the result.

Now, consider $\alpha \neq 1$, then, under Assumption 1, we recall from Proposition 6 that

$$\nabla^2 f_\pi^{(\alpha)}(\theta) = \nabla^2 A(\theta) + (\alpha - 1)\left(\pi_\theta^{(\alpha)}(\Gamma\Gamma^\top) - \pi_\theta^{(\alpha)}(\Gamma)(\pi_\theta^{(\alpha)}(\Gamma))^\top\right), \forall\theta \in \operatorname{int}\Theta.$$

Consider $\theta \in \operatorname{int}\Theta$, we show now that $\pi_\theta^{(\alpha)}(\Gamma\Gamma^\top) - \pi_\theta^{(\alpha)}(\Gamma)(\pi_\theta^{(\alpha)}(\Gamma))^\top$ is positive semidefinite. Consider a vector $\xi \in \mathbb{R}^d$, then

$$\begin{aligned}\langle \xi, \pi_\theta^{(\alpha)}(\Gamma\Gamma^\top), \xi\rangle &= \int (\langle \Gamma(x), \xi\rangle)^2 \pi_\theta^{(\alpha)}(x)\nu(dx)\\ &\geq \left(\int \langle \Gamma(x), \xi\rangle \pi_\theta^{(\alpha)}(x)\nu(dx)\right)^2\\ &= \langle \xi, \pi_\theta^{(\alpha)}(\Gamma)\pi_\theta^{(\alpha)}(\Gamma)^\top \xi\rangle,\end{aligned}$$

where we used Jensen inequality to show the inequality. This shows that

$$\langle \xi, \left(\pi_\theta^{(\alpha)}(\Gamma\Gamma^\top) - \pi_\theta^{(\alpha)}(\Gamma)\pi_\theta^{(\alpha)}(\Gamma)^\top\right) \xi\rangle \geq 0, \, \forall \xi \in \mathbb{R}^d.$$

Therefore, for every $\theta \in \operatorname{int}\Theta$, $\nabla^2(f_\pi^{(\alpha)} - A)(\theta)$ is positive semidefinite if $\alpha \geq 1$, and $\nabla^2(A - f_\pi^{(\alpha)})(\theta)$ is positive semidefinite if $\alpha \leq 1$. This shows that $f_\pi^{(\alpha)} - A$ is convex if $\alpha \geq 1$ and $A - f_\pi^{(\alpha)}$ is convex if $\alpha \leq 1$, giving the results using the characterizations from [40, Proposition 2.2] and [40, Proposition 2.3]. □

A.6 Proof of Proposition 10

Proof of Proposition 10. Consider $\alpha > 0$.

(i) $F_\pi^{(\alpha)}$ is proper because $f_\pi^{(\alpha)}$ is non-negative from Proposition 1, takes finite values for some $\theta \in \Theta$ by Assumption 1, and because r is proper by Assumption 3. The fact that the infimum of ($P_\pi^{(\alpha)}$) is not equal to $-\infty$ comes from the non-negativity of $f_\pi^{(\alpha)}$ and the fact that r is bounded from below from Assumption 3.

We now prove the lower semicontinuity. When $\alpha = 1$, we recall from Eq. (23) that

$$f_\pi^{(1)}(\theta) = H(\pi) - \langle \theta, \pi(\Gamma)\rangle + A(\theta), \, \forall \theta \in \Theta, \tag{25}$$

where $H(\pi) = \int \log(\pi(x))\pi(x)\nu(dx)$. Because A is lower semicontinuous on Θ from Proposition 4, so is $f_\pi^{(1)}$.

Now consider $\alpha \neq 1$. For every $\theta \in \Theta$, it is possible to decompose $f_\pi^{(\alpha)}$ as in

$$f_\pi^{(\alpha)}(\theta) = A(\theta) + \frac{1}{\alpha - 1}\log\left(\tilde{h}(\theta)\right),$$

where $\tilde{h}(\theta) = \int \pi(x)^\alpha \exp(\langle \theta, \Gamma(x)\rangle)^{1-\alpha}\nu(dx)$. The function $\tilde{h}$ is lower semicontinuous due to Fatou's lemma [22, Lemma 18.13] and takes values in $\mathbb{R}_{++}$, thus $\frac{1}{\alpha-1}\log\circ\tilde{h}$ is lower semicontinuous.

(ii) We now turn to the second point, concerning values $\alpha \geq 1$. In the particular case $\alpha = 1$, consider again the decomposition given in Eq. (25). Because of Assumption 2, $\pi(\Gamma) \in \operatorname{int}\operatorname{dom} A^*$. Thanks to [8, Fact 2.11] and Proposition 4, this ensures that $f_\pi^{(1)}$ is coercive. Because of Assumption 1 which ensures the well-posedness of $f_\pi^{(\alpha)}$, we have from [76, Theorem 3] that

$$f_\pi^{(1)}(\theta) \leq f_\pi^{(\alpha)}(\theta), \, \forall \theta \in \operatorname{int}\Theta.$$

This ensures that $f_\pi^{(\alpha)}$ is coercive for $\alpha > 1$. The regularizer r is bounded from below thanks to Assumption 3, so $F_\pi^{(\alpha)}$ is also coercive for $\alpha \geq 1$.

We have proven that $F_\pi^{(\alpha)}$ is lower-continuous and coercive, so there exists $\theta_* \in \operatorname{dom}\Theta$ such that $F_\pi^{(\alpha)}(\theta_*) = \vartheta_\pi^{(\alpha)}$. We now use the optimality conditions that θ_* satisfies to show that $\theta_* \in \operatorname{int}\Theta$. In particular, we have from [10, Theorem 16.2] that

$$0 \in \partial F_\pi^{(\alpha)}(\theta_*). \tag{26}$$

When $\alpha = 1$, we can split the subdifferential of $F_\pi^{(\alpha)}$ as $\partial F_\pi^{(1)}(\theta_*) = \pi(\Gamma) + \partial A(\theta_*) + \partial r(\theta_*)$. This comes from the decomposition (23), Assumption 3 and the convexity and properness of $\theta \longmapsto -\langle\theta, \pi(\Gamma)\rangle$, A and r, and [10, Corollary 16.38]. By the same arguments, when $\alpha > 1$, $\partial F_\pi^{(\alpha)}(\theta_*) = \partial\left(\frac{1}{\alpha-1}\log \circ h_\pi^{(\alpha)}\right)(\theta_*) + \partial A(\theta_*) + \partial r(\theta_*)$

Assume by contradiction that θ_* belongs to the boundary of Θ. Then $\partial A(\theta_*) = \emptyset$, because of Proposition 2, so Eq. (26) implies that $0 \in \emptyset$. This shows that $\theta_* \in \operatorname{int}\Theta$.

Finally, since A is strictly convex on $\operatorname{int}\Theta$ (Proposition 4), so is $F_\pi^{(1)}$, so such θ_* is unique. $\square$

A.7 Proof of Proposition 11

Proof of Proposition 11. Consider the family of one-dimensional centered Gaussian distributions with variance σ^2, that we denote by $\mathcal{G}_0^1$ in the following. It is an exponential family, with parameter $\theta = -\frac{1}{2\sigma^2}$, sufficient statistics $\Gamma(x) = x^2$ and log-partition function $A(\theta) = \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(-2\theta)$, whose domain is $\Theta = \mathbb{R}_{--}$. We show that $(f_\pi^{(\alpha)})'$ is not Lipschitz for $\alpha > 0$ by showing that $(f_\pi^{(\alpha)})''$ is unbounded on Θ.

Consider first the case $\alpha = 1$. From Proposition 6, $(f_\pi^{(\alpha)})''$ is independent of the choice of the target π, and is equal to

$$(f_\pi^{(\alpha)})''(\theta) = A''(\theta) = \frac{1}{2\theta^2}. \tag{27}$$

Now, for $\alpha \neq 1$, consider a target $\pi \in \mathcal{G}_0^1$, meaning that there exists $\theta_\pi \in \Theta$ such that $\pi = q_{\theta_\pi}$. We can compute that $\pi_\theta^{(\alpha)} = q_{\alpha\theta_\pi + (1-\alpha)\theta}$, assuming that θ is such that $\alpha\theta_\pi + (1-\alpha)\theta \in \Theta$. This condition is always satisfied when $\alpha \leq 1$, but when $\alpha > 1$, it is equivalent to having $\theta > \frac{\alpha}{\alpha-1}\theta_\pi$. In the case $\alpha > 1$, $f_\pi^{(\alpha)}$ is not even defined outside of $(\frac{\alpha}{\alpha-1}\theta_\pi, 0)$, showing that $\operatorname{dom} f_\pi^{(\alpha)} = (0, \frac{\alpha}{\alpha-1}\theta_\pi)$ for $\alpha > 1$. In the following, we consider $\theta \in \operatorname{dom} f_\pi^{(\alpha)}$. Then we compute using the result of Proposition 6 that

$$(f_\pi^{(\alpha)})''(\theta) = \frac{1}{2\theta^2} + (\alpha - 1)\left(\int x^4 q_{\alpha\theta_\pi+(1-\alpha)\theta}(x)dx - \left(\int x^2 q_{\alpha\theta_\pi+(1-\alpha)\theta}(x)dx\right)^2\right).$$

To do so, we recall the following formulas

$$\int x^4 \exp(-bx^2)dx = \frac{3\sqrt{\pi}}{4b^{5/2}}, \qquad \int x^2 \exp(-bx^2)dx = \frac{\sqrt{\pi}}{2b^{3/2}},$$

and we note that $A(\theta) = \log\left(\sqrt{-\frac{\pi}{\theta}}\right)$. We first compute

$$\begin{aligned}
\int x^4 q_{\alpha\theta_\pi+(1-\alpha)\theta}(x)dx &= \exp(A(\alpha\theta_\pi + (1-\alpha)\theta))^{-1}\int x^4 \exp(-((\alpha-1)\theta - \alpha\theta_\pi)x^2)dx \\
&= \left(\frac{\pi}{(\alpha-1)\theta - \alpha\theta_\pi}\right)^{-1/2} \frac{3\sqrt{\pi}}{4((\alpha-1)\theta - \alpha\theta_\pi)^{5/2}} \\
&= \frac{3}{4((\alpha-1)\theta - \alpha\theta_\pi)^2},
\end{aligned}$$

and then by similar means $\int x^2 q_{\alpha\theta_\pi+(1-\alpha)\theta}(x)dx = \frac{1}{2((\alpha-1)\theta-\alpha\theta_\pi)}$.

These calculations yield

$$(f_\pi^{(\alpha)})''(\theta) = \frac{1}{2\theta^2} + \frac{\alpha - 1}{2((\alpha-1)\theta - \alpha\theta_\pi)^2}. \tag{28}$$

Equations (27) and (28) show that the absolute value of $\nabla^2 f_\pi^{(\alpha)}$ goes to $+\infty$ when θ approaches 0 or $\frac{\alpha}{\alpha-1}\theta_\pi$, which is in Θ if and only if $\alpha > 1$. $\square$

A.8 Proof of Proposition 12

Proof of Proposition 12. Since $\pi = q_{\theta_\pi}$, we can compute that $\pi_\theta^{(\alpha)} = q_{\alpha\theta_\pi+(1-\alpha)\theta}$. Since Θ is convex (from Proposition 4) and $\theta_\pi, \theta \in \operatorname{int}\Theta$, $\alpha\theta_\pi + (1-\alpha)\theta \in \operatorname{int}\Theta$. Therefore, $\pi_\theta^{(\alpha)}(\Gamma) = \nabla A(\alpha\theta_\pi + (1-\alpha)\theta) \in \operatorname{int}\operatorname{dom} A^*$, since A is Legendre and $\alpha\theta_\pi + (1-\alpha)\theta \in \operatorname{int}\operatorname{dom} A$, showing that Assumption 2 is satisfied.

Recall that $RD_\alpha(\pi, q_\theta) \geq 0$, and it is equal to zero if and only if $\pi = q_\theta$. In our case, this means that $f_\pi^{(\alpha)}(\theta) = 0$ if and only if $\theta = \theta_\pi$. Since $\mathcal{Q}$ is minimal by assumption, this shows the existence and unicity of the minimizer of $f_\pi^{(\alpha)}$.

Consider now a stationary point of $f_\pi^{(\alpha)}$ i.e., $\theta \in \operatorname{int}\Theta$ such that $\nabla f_\pi^{(\alpha)}(\theta) = 0$. This implies that

$$q_\theta(\Gamma) = \pi_\theta^{(\alpha)}(\Gamma), \tag{29}$$

due to the characterization given in Proposition 6. Under our assumptions, Eq. (29) now reads

$$\nabla A(\theta) = \nabla A(\alpha\theta_\pi + (1-\alpha)\theta),$$

which is equivalent to having $\theta = \theta_\pi$ by inverting ∇A on both sides. Hence we have shown that $\nabla f_\pi^{(\alpha)}(\theta) = 0$ if and only $\theta = \theta_\pi$, showing the existence and unicity of the stationary point of $f_\pi^{(\alpha)}$. ☐

A.9 Proof of Proposition 13

Proof of Proposition 13. Under our Assumptions on π, we can compute

$$f_\pi^{(\alpha)}(\theta) = \frac{1}{1-\alpha}\left(\alpha A(\theta_\pi) + (1-\alpha)A(\theta) - A(\alpha\theta_\pi + (1-\alpha)\theta)\right), \forall\theta \in \operatorname{int}\Theta \tag{30}$$

Consider in the following $\theta, \theta' \in \operatorname{int}\Theta$, with $\theta, \theta' \in B(\theta_\pi, \upsilon)$.

(i) We can compute

$$\begin{aligned}
A(\alpha\theta_\pi + (1-\alpha)\theta) &= A(\theta_\pi + (1-\alpha)(\theta - \theta_\pi)) \\
&= A(\theta_\pi) + (1-\alpha)\langle\nabla A(\theta_\pi), \theta - \theta_\pi\rangle + \frac{(1-\alpha)^2}{2}\|\theta - \theta_\pi\|_{\nabla^2 A(\theta_\pi)} + o(\upsilon^2), \\
A(\theta) &= A(\theta_\pi) + \langle\nabla A(\theta_\pi), \theta - \theta_\pi\rangle + \frac{1}{2}\|\theta - \theta_\pi\|_{\nabla^2 A(\theta_\pi)} + o(\upsilon^2).
\end{aligned}$$

Using Eq. (30), these two equalities imply in particular that

$$f_\pi^{(\alpha)}(\theta) = \frac{\alpha}{2}\|\theta - \theta_\pi\|^2_{\nabla^2 A(\theta_\pi)} + o(\upsilon^2). \tag{31}$$

(ii) We now turn to the Polyak-Łojasiewicz inequality. We begin by showing that $d_{f_\pi^{(\alpha)}}(\theta, \theta') = \frac{\alpha}{2}\|\theta - \theta'\|^2_{\nabla^2 A(\theta_\pi)}$ up to higher order terms. We now further compute that

$$\begin{aligned}
\nabla f_\pi^{(\alpha)}(\theta) &= \nabla A(\theta) - \nabla A(\alpha\theta_\pi + (1-\alpha)\theta_\pi) \\
&= \nabla A(\theta_\pi + (\theta - \theta_\pi)) - \nabla A(\theta_\pi + (1-\alpha)(\theta - \theta_\pi) \\
&= \nabla A(\theta_\pi) + \nabla^2 A(\theta_\pi)(\theta - \theta_\pi) - \nabla A(\theta_\pi) - (1-\alpha)\nabla^2 A(\theta_\pi)(\theta - \theta_\pi) + o(\upsilon) \\
&= \alpha\nabla^2 A(\theta_\pi)(\theta - \theta_\pi) + o(\upsilon),
\end{aligned}$$

which yields with Eq. (31) that

$$\begin{aligned}
d_{f_\pi^{(\alpha)}}(\theta,\theta') &= f_\pi^{(\alpha)}(\theta) - f_\pi^{(\alpha)}(\theta') - \langle \nabla f_\pi^{(\alpha)}(\theta'), \theta-\theta'\rangle \\
&= \frac{\alpha}{2}\|\theta-\theta_\pi\|^2_{\nabla^2 A(\theta_\pi)} - \frac{\alpha}{2}\|\theta'-\theta_\pi\|^2_{\nabla^2 A(\theta_\pi)} - \alpha\langle \nabla^2 A(\theta_\pi)(\theta'-\theta_\pi), \theta-\theta'\rangle + o(\upsilon^2) \\
&= \frac{\alpha}{2}\|\theta-\theta'\|^2_{\nabla^2 A(\theta_\pi)} + o(\upsilon^2).
\end{aligned}$$

We thus have that

$$f_\pi^{(\alpha)}(\theta) = f_\pi^{(\alpha)}(\theta') + \langle \nabla f_\pi^{(\alpha)}(\theta'), \theta-\theta'\rangle + \frac{\alpha}{2}\|\theta-\theta'\|^2_{\nabla^2 A(\theta_\pi)} + o(\upsilon^2). \tag{32}$$

When θ' is fixed, the quantity $\theta \longmapsto f_\pi^{(\alpha)}(\theta') + \langle \nabla f_\pi^{(\alpha)}(\theta'), \theta-\theta'\rangle + \frac{\alpha}{2}\|\theta-\theta'\|^2_{\nabla^2 A(\theta_\pi)}$ is a quadratic form that is minimized when the following optimaliy condition is satisfied:

$$\nabla f_\pi(\theta') + \alpha\nabla^2 A(\theta_\pi)(\theta-\theta') = 0. \tag{33}$$

We now compute the inverse of $\nabla^2 A(\theta_\pi)$. Consider $\eta \in \operatorname{int}\operatorname{dom} A^*$. Since A is Legendre, we have that $\nabla A(\nabla A^*(\eta)) = \eta$, therefore differentiating this expression with respect to η yields

$$\nabla^2 A^*(\eta)\nabla^2 A(\nabla A^*(\eta)) = Id.$$

Now take $\eta = \nabla A(\theta_\pi)$ which belongs to $\operatorname{int}\operatorname{dom} A^*$ since $\theta_\pi \in \operatorname{int}\operatorname{dom} A$. We obtain that $\nabla^2 A^*(\nabla A(\theta_\pi))\nabla^2 A(\theta_\pi) = Id$. This shows that the optimality condition of Eq. (33) is equivalent to having

$$\theta - \theta' = -\frac{1}{\alpha}\nabla^2 A^*(\nabla A(\theta_\pi))\nabla f_\pi^{(\alpha)}(\theta').$$

This shows that the right-hand side of Eq. (32) can be minorized to obtain:

$$\begin{aligned}
f_\pi^{(\alpha)}(\theta) &\geq f_\pi^{(\alpha)}(\theta') - \frac{1}{\alpha}\langle \nabla f_\pi^{(\alpha)}(\theta'), \nabla^2 A^*(\nabla A(\theta_\pi))\nabla f_\pi^{(\alpha)}(\theta')\rangle \\
&\quad + \frac{1}{2\alpha}\langle \nabla^2 A^*(\nabla A(\theta_\pi))\nabla f_\pi^{(\alpha)}(\theta'), \nabla^2 A(\theta_\pi)\nabla^2 A^*(\nabla A(\theta_\pi))\nabla f_\pi^{(\alpha)}(\theta')\rangle + o(\upsilon^2) \\
&= f_\pi^{(\alpha)}(\theta') - \frac{1}{2\alpha}\|\nabla f_\pi^{(\alpha)}(\theta')\|^2_{\nabla^2 A^*(\nabla A(\theta_\pi))} + o(\upsilon^2).
\end{aligned}$$

This is true in particular for $\theta = \theta_\pi$, which yields the result.

□

B Convergence analysis of Algorithm 1

In order to prove Propositions 14 and 15, we start with a *sufficient decrease lemma* that reads as follows.

Lemma 1. *Under Assumptions 1, 2, and 3, for $\tau > 0$ and $\alpha \in (0,1]$, we have that for every $\theta \in \operatorname{int}\Theta$,*

$$\tau\left(F_\pi^{(\alpha)}(T^A_{\tau F_\pi^{(\alpha)}}(\theta)) - F_\pi^{(\alpha)}(\theta)\right) \leq -d_A(\theta, T^A_{\tau F_\pi^{(\alpha)}}(\theta)) + (\tau-1)d_A(T^A_{\tau F_\pi^{(\alpha)}}(\theta), \theta). \tag{34}$$

In the particular case where $\alpha = 1$, we further have

$$\begin{aligned}
\tau\left(F_\pi^{(1)}(T^A_{\tau F_\pi^{(1)}}(\theta)) - F_\pi^{(1)}(\theta')\right) \leq (1-\tau)d_A(\theta',\theta) - (1-\tau)d_A(T^A_{\tau F_\pi^{(1)}}(\theta), \theta) \\
- d_A(\theta', T^A_{\tau F_\pi^{(1)}}(\theta)),\ \forall \theta' \in \operatorname{int}\Theta. \qquad (35)
\end{aligned}$$

Proof. Using [74, Lemma 4.1], which still holds in our finite-dimensional Hilbert setting, we get that

$$\tau\left(F_\pi^{(\alpha)}(T^A_{\tau F_\pi^{(\alpha)}}(\theta)) - F_\pi^{(\alpha)}(\theta')\right) \leq d_A(\theta', \theta) - (1-\tau) d_A(T^A_{\tau F_\pi^{(1)}}(\theta), \theta) \\ - d_A(\theta', T^A_{\tau F_\pi^{(1)}}(\theta)) - \tau d_{f_\pi^{(\alpha)}}(\theta', \theta),\ \forall \theta' \in \operatorname{int}\Theta,$$

where $d_{f_\pi^{(\alpha)}}(\theta', \theta) = f_\pi^{(\alpha)}(\theta') - f_\pi^{(\alpha)}(\theta) - \langle \nabla f_\pi^{(\alpha)}(\theta), \theta' - \theta\rangle$.

Equation (34) comes by evaluating the above at $\theta' = \theta$. To get Eq. (35), the strong convexity of $f_\pi^{(1)}$ relatively to A yields

$$d_{f_\pi^{(1)}}(\theta', \theta) \geq d_A(\theta', \theta),\ \forall \theta', \theta \in \operatorname{int}\Theta,$$

showing the result. ☐

We also give a *sequential consistency* lemma, that links the Bregman divergence d_A with the Euclidean distance.

Lemma 2. *Consider two sequences $\{\theta_k\}_{k\in\mathbb{N}}$ and $\{\theta'_k\}_{k\in\mathbb{N}}$ and assume that there exists a compact set $C \subset \operatorname{int}\Theta$ such that $\theta_k, \theta'_k \in C$ for every $k \in \mathbb{N}$. In this case, if $d_A(\theta_k, \theta'_k) \xrightarrow[k\to+\infty]{} 0$, then $\|\theta_k - \theta'_k\| \xrightarrow[k\to+\infty]{} 0$.*

Proof. We introduce the convex hull of C, denoted by $\operatorname{conv} C$ which is the intersection of every convex set containing C. Therefore $\operatorname{conv} C \subset \operatorname{int}\Theta$. Since we are in finite dimension, we also have that $\operatorname{conv} C$ is compact. Thus, $\operatorname{conv} C$ is a convex compact included in $\operatorname{int}\Theta$.

A is proper, strictly convex, and continuous on $\operatorname{conv} C \subset \operatorname{int}\Theta$, therefore, A is *uniformly convex*, following the definition of [10, Definition 10.5] on $\operatorname{conv} C$ [10, Proposition 10.15]. This means that there exists an increasing function $\psi : \mathbb{R}_+ \to [0, +\infty]$ that vanishes only at 0, such that for every $\theta, \theta' \in \operatorname{conv} C$,

$$\psi(\|\theta - \theta'\|) \leq \frac{1}{2}A(\theta) + \frac{1}{2}A(\theta') - A\left(\frac{1}{2}\theta + \frac{1}{2}\theta'\right).$$

Because A is convex on $\operatorname{conv} C$, we prove following the proof of [19, Eq. (1.32)] that for every $\theta, \theta' \in \operatorname{conv} C$,

$$d_A(\theta, \theta') \geq \psi(\|\theta - \theta'\|).$$

Suppose now by contradiction that $d_A(\theta_k, \theta'_k) \xrightarrow[k\to+\infty]{} 0$ while there exists some $\epsilon > 0$ such that $\|\theta_k - \theta'_k\| \geq \epsilon$ for every $k \in \mathbb{N}$. Then we have that

$$d_A(\theta_k, \theta'_k) \geq \psi(\epsilon) > 0,$$

which is a contradiction, hence showing the result. ☐

B.1 Proof of Proposition 14

Proof of Proposition 14. We first give an intermediate result that will be used several times. Using Lemma 1, we can get for any $k \in \mathbb{N}$ that

$$F_\pi^{(\alpha)}(\theta_{k+1}) - F_\pi^{(\alpha)}(\theta_k) \leq -\frac{1}{\tau_{k+1}} d_A(\theta_k, \theta_{k+1}) - \left(\frac{1}{\tau_{k+1}} - 1\right) d_A(\theta_{k+1}, \theta_k). \tag{36}$$

(i) Due to the non-negativity of the Bregman divergences and the hypothesis on the step-size τ_{k+1}, the right-hand side of Eq. (36) is non-negative, yielding the decrease property $F_\pi^{(\alpha)}(\theta_{k+1}) \leq F_\pi^{(\alpha)}(\theta_k)$.

(ii) If $F_\pi^{(\alpha)}(\theta_{K+1}) = F_\pi^{(\alpha)}(\theta_K)$, then, using Lemma 1 and $\tau_{K+1} \leq 1$, $d_A(\theta_K, \theta_{K+1}) \leq 0$. By Proposition 3, this shows that $\theta_{K+1} = \theta_K$. Since $\theta_{K+1} = T^A_{\tau_{K+1} F_\pi^{(\alpha)}}(\theta_K)$, θ_K is a fixed point of $T^A_{\tau_{K+1} F_\pi^{(\alpha)}}$. From Proposition 7, it is a stationary point of $F_\pi^{(\alpha)}$.

(iii) By summing Eq. (36) over $k \in [\![0, K]\!]$, one gets

$$\sum_{k=0}^{K} \left(\frac{1}{\tau_{k+1}} d_A(\theta_k, \theta_{k+1}) + \left(\frac{1}{\tau_{k+1}} - 1 \right) d_A(\theta_{k+1}, \theta_k) \right) \leq F_\pi^{(\alpha)}(\theta_0) - F_\pi^{(\alpha)}(\theta_{K+1}).$$

Because of the hypothesis on the step-sizes, one has for any $k \in \mathbb{N}$ that $1 \leq \frac{1}{\tau_{k+1}}$ and $0 \leq \frac{1}{\tau_{k+1}} - 1$, giving the inequality

$$\sum_{k=0}^{K} d_A(\theta_k, \theta_{k+1}) \leq F_\pi^{(\alpha)}(\theta_0) - \vartheta_\pi^{(\alpha)}.$$

Since the right-hand side of the above is uniform in K, this implies the result.

(iv) Suppose that K is the first iterate such that $d_A(\theta_K, \theta_{K+1}) \leq \varepsilon$. Then for any $k \in [\![0, K-1]\!]$, $d_A(\theta_k, \theta_{k+1}) > \varepsilon$. Consider Eq. (B.1) where the sum goes only from $k = 0$ to $k = K - 1$, then one can show the desired result with

$$\varepsilon K \leq \sum_{k=0}^{K-1} d_A(\theta_k, \theta_{k+1}) \leq F_\pi^{(\alpha)}(\theta_0) - \vartheta_\pi^{(\alpha)}.$$

(v) This proof relies on two notions of subdifferentials: the limiting subdifferential ∂_L [70, Chapter 6] and the Fréchet subdifferential ∂_F [70, Chapter 4]. Our working space $\mathcal{H}$ is a finite-dimensional Hilbert space, which is included in the setting of [70].

Set $k \in \mathbb{N}$. Under Assumptions 1, 2, and 3, since $\theta_0 \in \operatorname{int} \Theta$, Proposition 8 applies and thus $\theta_{k+1} = T^A_{\tau_{k+1} F_\pi^{(\alpha)}}(\theta_k)$. This implies that there exists $g_{k+1} \in \partial r(\theta_{k+1})$ such that

$$\frac{1}{\tau_{k+1}} (\nabla A(\theta_{k+1}) - \nabla A(\theta_k)) + \nabla f_\pi^{(\alpha)}(\theta_k) + g_{k+1} = 0. \tag{37}$$

According to [70, Corollary 4.35],

$$\nabla f_\pi^{(\alpha)}(\theta_{k+1}) + g_{k+1} \in \partial_F F_\pi^{(\alpha)}(\theta_{k+1}). \tag{38}$$

Using Eq. (37) and the assumptions on τ_{k+1},

$$\|\nabla f_\pi^{(\alpha)}(\theta_{k+1}) + g_{k+1}\| \leq \|\nabla f_\pi^{(\alpha)}(\theta_{k+1}) - \nabla f_\pi^{(\alpha)}(\theta_k)\| + \frac{1}{\epsilon} \|\nabla A(\theta_{k+1}) - \nabla A(\theta_k)\|.$$

The additional hypothesis introduced in (iv) ensures that both θ_{k+1} and θ_k belong to C, a compact set included in $\operatorname{int} \Theta$. Since $\nabla^2 f_\pi^{(\alpha)}$ is continuous on C (by Proposition 6) and C is bounded, $\nabla f_\pi^{(\alpha)}$ is Lipschitz on C. The same reasoning applies for ∇A. This shows that there exists a scalar $s > 0$ such that, for every $k \in \mathbb{N}$, there exists $\varrho_{k+1} \in \partial_F F_\pi^{(\alpha)}(\theta_{k+1})$ satisfying

$$\|\varrho_{k+1}\| \leq s \|\theta_{k+1} - \theta_k\|. \tag{39}$$

Now, we deduce from (iii) that $d_A(\theta_{k+1}, \theta_k) \xrightarrow[k \to +\infty]{} 0$. Using Lemma 2, this yields $\|\theta_{k+1} - \theta_k\| \xrightarrow[k \to +\infty]{} 0$, showing that the sequence $\{\varrho_k\}_{k \in \mathbb{N}}$ is such that

$$\varrho_k \in \partial_F F_\pi^{(\alpha)}(\theta_k), \ \forall k \in \mathbb{N}, \ \text{and} \ \varrho_k \xrightarrow[k \to +\infty]{} 0. \tag{40}$$

On the other hand, the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ is contained in the compact set C by assumption. Hence, there exists $\theta_{\lim} \in C$, and a strictly increasing function $\varphi : \mathbb{N} \to \mathbb{N}$ such that $\theta_{\varphi(k)} \xrightarrow[k\to+\infty]{} \theta_{\lim}$. The regularizing term r is continuous on C as assumed in (iv), implying

$$\begin{cases} \theta_{\varphi(k)} \xrightarrow[k\to+\infty]{} \theta_{\lim}, \\ F_\pi^{(\alpha)}(\theta_{\varphi(k)}) \xrightarrow[k\to+\infty]{} F_\pi^{(\alpha)}(\theta_{\lim}), \\ \varrho_{\varphi(k)} \in \partial_F F_\pi^{(\alpha)}(\theta_{\varphi(k)}),\ \varrho_{\varphi(k)} \xrightarrow[k\to+\infty]{} 0. \end{cases}$$

By definition of the limiting subdifferential $\partial_L F_\pi^{(\alpha)}$ [70, Defintion 6.1], this shows that $0 \in \partial_L F_\pi^{(\alpha)}(\theta_{\lim})$. Hence $\theta_{\lim}$ is a stationary point of $F_\pi^{(\alpha)}$ concluding the proof. □

B.2 Proof of Proposition 15

Proof of Proposition 15. We first give an inequality to prove (i)-(ii). Consider iteration k of Algorithm 1, and evaluate Eq. (35) from Lemma 1 at $\theta' = \theta_*$, yielding

$$\tau_{k+1}\left(F_\pi^{(\alpha)}(\theta_{k+1}) - F_\pi^{(\alpha)}(\theta_*)\right) \le (1-\tau_{k+1})d_A(\theta_*,\theta_k) \\ - (1-\tau_{k+1})d_A(\theta_{k+1},\theta_k) - d_A(\theta_*,\theta_{k+1}). \quad (41)$$

(i) Since $\tau_{k+1} \in [\epsilon, 1]$, $F_\pi^{(1)}(\theta_{k+1}) \ge F_\pi^{(1)}(\theta_*)$, and d_A takes non-negative values (from Proposition 3), Eq. (41) gives

$$d_A(\theta_*,\theta_{k+1}) \le (1-\tau_{k+1})d_A(\theta_*,\theta_k), \quad (42)$$

from which we deduce the results since $\tau_{k+1} \in [\epsilon, 1]$.

(ii) Since $\tau_{k+1} \in [\epsilon, 1]$ and d_A takes non-negative values, we get from Eq. (41) that

$$\tau_{k+1}\left(F_\pi^{(1)}(\theta_{k+1}) - F_\pi^{(1)}(\theta_*)\right) \le (1-\tau_{k+1})d_A(\theta_*,\theta_k).$$

With Eq. (42) and the condition on τ_{k+1}, we obtain

$$\left(F_\pi^{(1)}(\theta_{k+1}) - F_\pi^{(1)}(\theta_*)\right) \le \frac{1}{\epsilon} d_A(\theta_*,\theta_{k+1}),$$

from which we conclude using point (i) and Proposition 10.

(iii) Using Proposition 14 (i), we obtain that for every $k \in \mathbb{N}$, $F_\pi^{(1)}(\theta_k) \le F_\pi^{(1)}(\theta_0)$, meaning that the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ is contained in a sub-level set of $F_\pi^{(1)}$. $F_\pi^{(1)}$ is coercive under our assumptions (see the proof of Proposition 10), and it is lower semicontinuous from Proposition 4, so its sub-level sets are compact. This means that we can extract converging subsequences from $\{\theta_k\}_{k\in\mathbb{N}}$.

Consider now such a subsequence $\{\theta_{\varphi(k)}\}_{k\in\mathbb{N}}$, with $\theta_{\varphi(k)} \xrightarrow[k\to+\infty]{} \theta_{\lim}$. $F_\pi^{(1)}$ is lower semicontinuous, so

$$\liminf F_\pi^{(1)}(\theta_{\varphi(k)}) \ge F_\pi^{(1)}(\theta_{\lim}).$$

However, because of (ii), $\liminf F_\pi^{(1)}(\theta_{\varphi(k)}) = F_\pi^{(1)}(\theta_*)$, so we obtain that $F_\pi^{(1)}(\theta_{\lim}) = F_\pi^{(1)}(\theta_*)$. Using Proposition 10, this shows that $\theta_{\lim} = \theta_*$.

We have shown that $\{\theta_k\}_{k\in\mathbb{N}}$ is contained in a compact set and that each of its converging subsequences converges to θ_*, which implies the result.

□

B.3 Proof of Proposition 16

Proof of Proposition 16. We first prove that we can apply the results of Proposition 14. Assumption 2 holds because of the result of Proposition 12. Since $r \equiv 0$, Assumption 3 holds. Therefore, all the hypotheses of Proposition 14 hold, showing the result.

(iii) Because of Proposition 14, every converging subsequence of $\{\theta_k\}_{k\in\mathbb{N}}$ converges to $S_\pi^{(\alpha)}$. However, $S_\pi^{(\alpha)} = \{\theta_\pi\}$ in our case (see Proposition 12). Therefore, all the converging subsequences of $\{\theta_k\}_{k\in\mathbb{N}}$ converge to θ_π, showing that the sequence of iterates converges to θ_π.

(ii) Due to the smoothness of $f_\pi^{(\alpha)}$ relatively to A shown in Proposition 9, we have for any $k \in \mathbb{N}$ that

$$f_\pi^{(\alpha)}(\theta_{k+1}) \leq f_\pi^{(\alpha)}(\theta_k) + \langle \nabla f_\pi^{(\alpha)}(\theta_k), \theta_{k+1} - \theta_k \rangle + d_A(\theta_{k+1}, \theta_k). \tag{43}$$

Since $r \equiv 0$, we have the relation $\nabla A(\theta_{k+1}) = \nabla A(\theta_k) - \tau_{k+1} \nabla f_\pi^{(\alpha)}(\theta_k)$, therefore, Eq. (43) reads

$$f_\pi^{(\alpha)}(\theta_{k+1}) \leq f_\pi^{(\alpha)}(\theta_k) - \frac{1}{\tau_{k+1}} \langle \nabla A(\theta_{k+1}) - \nabla A(\theta_k), \theta_{k+1} - \theta_k \rangle + d_A(\theta_{k+1}, \theta_k). \tag{44}$$

Since $\tau_{k+1} \leq 1$ and $\langle \nabla A(\theta_{k+1}) - \nabla A(\theta_k), \theta_{k+1} - \theta_k \rangle = d_A(\theta_{k+1}, \theta_k) + d_A(\theta_k, \theta_{k+1})$, we get

$$\begin{aligned}
f_\pi^{(\alpha)}(\theta_{k+1}) &\leq f_\pi^{(\alpha)}(\theta_k) - \frac{1}{\tau_{k+1}} \left(d_A(\theta_k, \theta_{k+1}) + d_A(\theta_{k+1}, \theta_k) \right) + \frac{1}{\tau_{k+1}} d_A(\theta_{k+1}, \theta_k) \\
&= f_\pi^{(\alpha)}(\theta_k) - \frac{1}{\tau_{k+1}} d_{A^*}(\nabla A(\theta_{k+1}), \nabla A(\theta_k)).
\end{aligned}$$

Now, we consider some $\upsilon > 0$. Since $\theta_k \to \theta_\pi$, there exists $K \in \mathbb{N}$ such that for any $k \geq K$, $\theta_k \in B(\theta_\pi, \upsilon)$ and $\nabla A(\theta_k) \in B(\nabla A(\theta_\pi), \upsilon)$ (recall that ∇A is continuous on $\operatorname{int} \Theta$). Thus, we can write following the same steps as in the proof of Proposition 13 that for any $k \geq K$,

$$d_{A^*}(\nabla A(\theta_{k+1}), \nabla A(\theta_k)) = \frac{1}{2} \|\nabla A(\theta_{k+1}) - \nabla A(\theta_k)\|^2_{\nabla^2 A^*(\nabla A(\theta_\pi))} + o(\upsilon^2).$$

From there, we obtain that

$$\begin{aligned}
d_{A^*}(\nabla A(\theta_{k+1}), \nabla A(\theta_k)) &= \frac{\tau_{k+1}^2}{2} \|\nabla f_\pi^{(\alpha)}(\theta_k)\|^2_{\nabla^2 A^*(\nabla A(\theta_\pi))} + o(\upsilon^2) \\
&\geq \tau_{k+1}^2 \alpha f_\pi^{(\alpha)}(\theta_k) + o(\upsilon^2),
\end{aligned}$$

where the last inequality comes from Proposition 13. With our previous point, this yields

$$\begin{aligned}
f_\pi^{(\alpha)}(\theta_{k+1}) &\leq (1 - \alpha\tau_{k+1}) f_\pi^{(\alpha)}(\theta_k) + o(\upsilon^2) \\
&< (1 - \alpha\delta\epsilon) f_\pi^{(\alpha)}(\theta_k) + o(\upsilon^2)
\end{aligned}$$

for any constant $\delta \in (0, 1)$. By choosing δ or υ small enough, we finally obtain

$$f_\pi^{(\alpha)}(\theta_{k+1}) \leq (1 - \alpha\delta\epsilon) f_\pi^{(\alpha)}(\theta_k).$$

This means that for any $k \geq K$, we have $f_\pi^{(\alpha)}(\theta_k) \leq (1 - \alpha\delta\epsilon)^{k-K} f_\pi^{(\alpha)}(\theta_K)$. Since the sequence $\{f_\pi^{(\alpha)}(\theta_k)\}_{k\in\mathbb{N}}$ is decreasing, $f_\pi^{(\alpha)}(\theta_K) \leq f_\pi^{(\alpha)}(\theta_0)$, which shows the result with $C = (1 - \alpha\delta\epsilon)^{-K}$.

$\square$

C Computation of a Bregman proximal operator

Consider an orthonormal matrix Q and the family of Gaussian distribution with covariance of the form $\Sigma = Q\,\mathrm{diag}(\sigma_1^2, ..., \sigma_d^2)Q^\top$ and mean $\mu \in \mathbb{R}^d$. It is an exponential family with parameters $\theta = (\theta_1, \theta_2)^\top$, with $\theta_1 = \mathrm{diag}(\frac{1}{\sigma_1^2}, ..., \frac{1}{\sigma_d^2})Q^\top \mu$ and $\theta_2 = -(\frac{1}{2\sigma_1^2}, ..., \frac{1}{2\sigma_d^2})^\top$. Its sufficient statistics is $\Gamma(x) = (Q^\top x, (Q^\top x_1)^2, ..., (Q^\top x_d)^2)$. Its log-partition function is $A(\theta) = -\frac{1}{4}\theta_1^\top(\mathrm{diag}(\theta_2))^{-1}\theta_1 + \frac{d}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^d \log(-2(\theta_2)_i)$, and its natural parameters $\nabla A(\theta)$ are $Q^\top\mu$ and $((Q^\top\mu)_1^2 + \sigma_1^2, ..., (Q^\top\mu)_d^2 + \sigma_d^2)^\top$.

We consider a regularizer that enforces sparsity on some components of the mean. We propose to this end

$$r(\theta) = \sum_{i=1}^d \eta_i \left|(\theta_1)_i\right|, \tag{45}$$

where $\eta_i \geq 0$ for $i \in [\![1, d]\!]$.

Since $\sigma_i^2 > 0$ for all $i \in [\![1, d]\!]$, having a null component in θ_1 means that $Q^\top\mu$ has a null component, promoting sparsity in $Q^\top\mu$. We aim at computing $\breve{\theta} = \mathrm{prox}_{\tau r}^A(\theta)$.

Lemma 3. *Consider the Gaussian family defined above. Consider q_θ in this family, with $\theta \in \operatorname{int}\Theta$ and whose mean and covariance are respectively μ and $Q\,\mathrm{diag}(\sigma_1^2, ..., \sigma_d^2)Q^\top$. If we consider the regularizing function defined in Eq. (45), then $\breve{\theta} = \mathrm{prox}_{\tau r}^A(\theta)$ is such that the mean $\breve{\mu}$ and covariance $Q\,\mathrm{diag}(\breve{\sigma}_1^2, ..., \breve{\sigma}_d^2)Q^\top$ of $q_{\breve{\theta}}$ satisfy for any $i \in [\![1, d]\!]$*

$$(Q^\top\breve{\mu})_i = \begin{cases} 0 & \text{if } (Q^\top\mu)_i \in [-\tau\eta_i, \tau\eta_i], \\ -\tau\eta_i + (Q^\top\mu)_i & \text{if } (Q^\top\mu)_i > \tau\eta_i, \\ \tau\eta_i + (Q^\top\mu)_i & \text{if } (Q^\top\mu)_i < -\tau\eta_i. \end{cases}$$
$$\breve{\sigma}_i^2 = (\sigma_i)^2 + ((Q^\top\mu)_i^2 - (Q^\top\breve{\mu})_i^2).$$

Consider $i \in [\![1, d]\!]$. In the particular case where $\eta_i = 0$, then $\mu_i^* = \mu_i$ and $\breve{\sigma}_i^2 = (\sigma_i)^2$. We can also remark that we always have $\breve{\sigma}_i^2 \geq \sigma_i^2$, with equality if and only if $(Q^\top\mu)_i = 0$. Therefore, the operator $\mathrm{prox}_{\tau r}^A$ modifies q_θ by shrinking certain values of the mean to zero, but it increases the variance. In particular, the bigger the $(Q^\top\mu)_i$, the bigger the variance increase.

When $Q = I$, the exponential family is the family of Gaussian distributions with diagonal covariance. The above results can thus be applied to this family too.

Proof. The regularizing function r is separable, so we study the optimality condition for every $i \in [\![1, d]\!]$. This is justified by [10, Proposition 16.8], which shows that $\partial r(\theta)$ is the Cartesian product of its subdifferentials with respect to each of its variable. Therefore, for $i \in [\![1, d]\!]$, we have

$$\begin{cases} \frac{1}{\tau}((Q^\top\mu)_i - (Q^\top\breve{\mu})_i) & \in \eta_i \partial|\cdot|((\breve{\theta}_1)_i), \\ \frac{1}{\tau}((Q^\top\mu)_i^2 + \sigma_i^2 - ((Q^\top\breve{\mu})_i^2 + \breve{\sigma}_i^2)) & = 0, \end{cases}$$

from which we already deduce the result about the standard deviation.

Because $(\breve{\Sigma}_i)^2 > 0$, the sign of $(\breve{\theta}_1)_i = \frac{1}{(\breve{\Sigma}_i)^2}(Q^\top\breve{\mu})_i$ is the sign of $(Q^\top\breve{\mu})_i$ and we get that

$$(Q^\top\mu)_i - (Q^\top\breve{\mu})_i \in \begin{cases} [-\tau\eta_i, \tau\eta_i] & \text{if } (Q^\top\breve{\mu})_i = 0, \\ \{\tau\eta_i\} & \text{if } (Q^\top\breve{\mu})_i > 0, \\ \{-\tau\eta_i\} & \text{if } (Q^\top\breve{\mu})_i < 0, \end{cases}$$

from which we can obtain the result.

$\square$

D Supplementary numerical experiments

D.1 Understanding the influence of the parameters

To this end, we use Gaussian targets in various dimensions d, with unnormalized density of the form

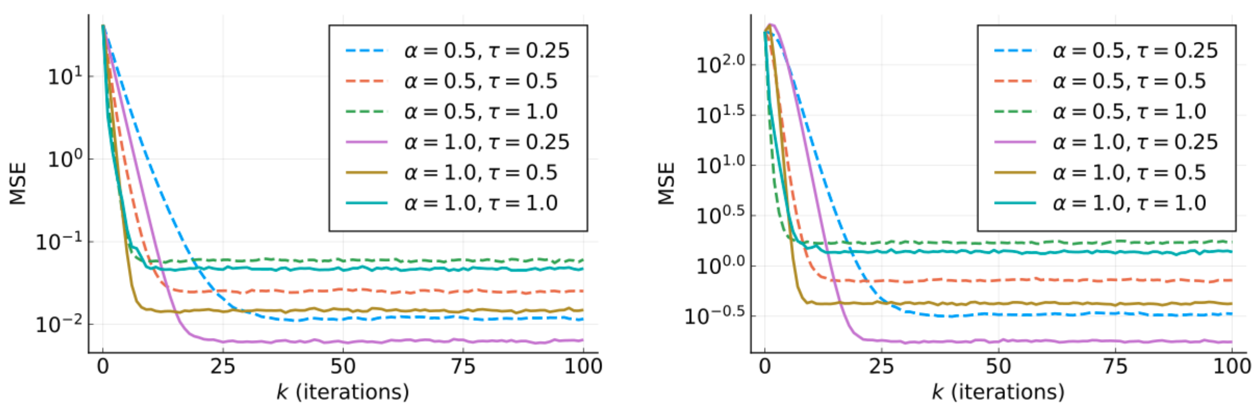
$$\tilde{\pi}(x) = \exp\left(-\frac{1}{2}(x-\bar{\mu})^\top \bar{\Sigma}_\kappa^{-1}(x-\bar{\mu})\right), \forall x \in \mathbb{R}^d. \tag{46}$$

Their means $\bar{\mu}$ are chosen uniformly in $[-0.5, 0.5]^d$ and their covariance matrices $\bar{\Sigma}_\kappa$ are chosen with a condition number equal to κ, following the procedure in [66, Section 5].

We now discuss the influence of α, τ on the practical speed and robustness of Algorithm 2, in its non-regularized version RMM. We recall that this algorithm resorts to importance sampling to approximate the integrals involved in the computation of $\pi_\theta^{(\alpha)}(\Gamma)$, which creates an approximation error linked with the sample size, N. The influence of τ can be understood through the theory on stochastic Bregman gradient descent with fixed step-size. In particular, [40, Theorem 5.3] states that such methods converge to a neighborhood of the optimum, whose size decreases with τ. On the other hand, low values of α amount to a concave transformation of the importance weights, which is known in the importance sampling field to lead to a higher effective sample size [54].

In order to highlight this compromise between speed and robustness, we use the RMM algorithm to approximate the target described in Eq. (46) with $\kappa = 10$. We use a constant number of samples per iteration $N = 500$, for $d \in \{5, 10, 20, 40\}$. It is recommended for importance sampling procedures that the sample size grows as $\exp(d)$ to avoid weight degeneracy [12]. In our setting, d increases while N remains constant, thus creating approximation errors that increase with d.

For each dimension, we test $\alpha \in \{0.5, 1.0\}$ and $\tau \in \{0.25, 0.5, 1.0\}$. We track the square errors $\|\bar{\mu} - \mu_k\|^2$ and $\|\bar{\Sigma}_\kappa - \Sigma_k\|_F^2$, that are averaged over 10^3 independent runs.

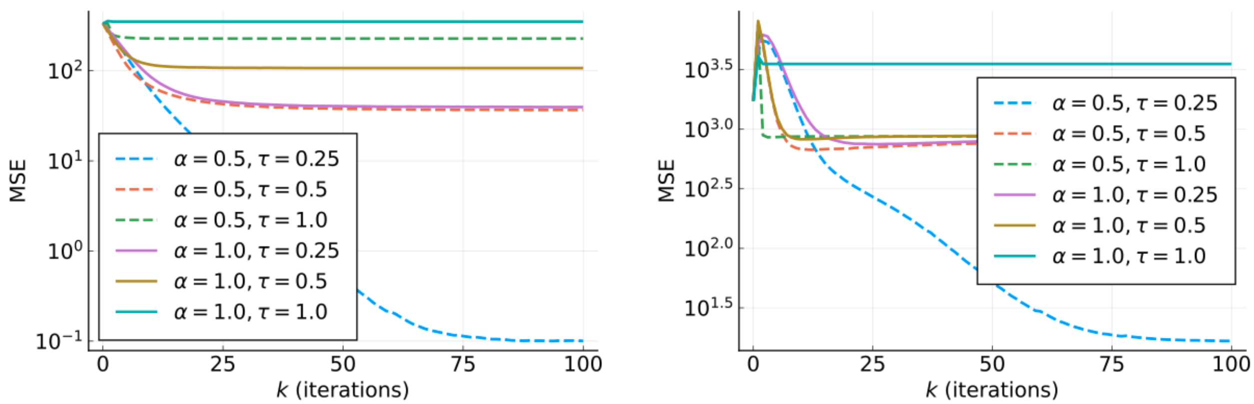


(a) MSE on the mean (b) MSE on the covariance

Figure 5: MSE averaged over 10^3 runs, in dimension $d = 5$.

In dimension $d = 5$, all the choices of parameters lead to convergence, as shown in Fig. 5. We can notice that the lowest values of τ lead to the slowest convergence, but the values reached are lower. On the contrary, when $\tau = 1.0$, the algorithm stops early at higher values.

Finally, for $d = 40$, only the lowest values of α and τ yield a significant decrease of the MSE as shown in Fig. 6. This shows that low values of α and τ can counteract high approximation errors. As expected, the convergence is slower and the final MSE values are higher than in lower dimensions.



(a) MSE on the mean

(b) MSE on the covariance

Figure 6: MSE averaged over 10^3 runs, in dimension $d = 40$.

This study shows that the parameters α and τ should be lowered to compensate for high approximation errors possibly arising in Algorithm 2. On the contrary, when these errors are low, one can increase the values of τ to create faster algorithms.

D.2 Comparison with the variational Rényi bound on a Gaussian target

Our theoretical analysis provides guidelines to choose the step-size τ for our RMM algorithm (Propositions 14 and 15) but also shows that there is no equivalent guarantees for the VRB algorithm (see Proposition 11). In particular, poorly chosen step-sizes could create unstable behaviors. We thus investigate these effects in the following by comparing our novel RMM algorithm with the VRB algorithm on Gaussian targets.

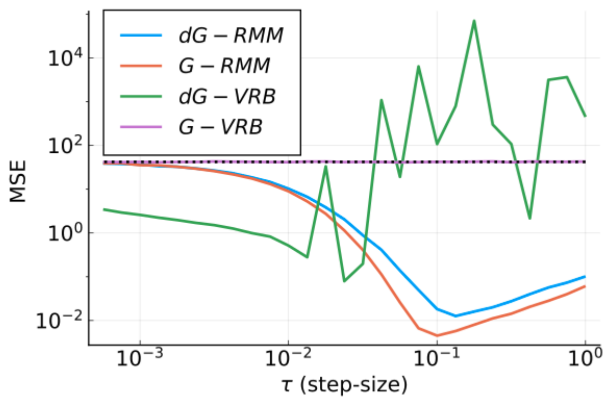
We use Gaussian target from Eq. (46), with $\kappa = 10$, and $d = 5$. Each algorithm is run with constant number of samples $N = 500$, and constant values of the step-size τ. We test values of α corresponding to the Hellinger distance ($\alpha = 0.5$) and the KL divergence ($\alpha = 1.0$). We test two different exponential families: Gaussian with full covariance, and Gaussian with diagonal covariance. For each tested value of τ, 10^3 runs are performed.

Figure 7 shows that the VRB algorithm used with diagonal covariance in the approximation family exhibits two distinct regimes. For sufficiently low values of τ, it is able to improve the estimates compared to initialization, but once τ crosses a certain threshold, the MSE reaches very high values, showing a degradation from the initialization. The VRB algorithm with full covariance in the approximation family is not able to create covariance matrices that are positive definite, hence it stops after initialization. On the contrary, our RMM algorithm does not degrade the values reached at initialization even for the worst settings of τ, and reaches the lowest MSE values for properly chosen step-sizes.

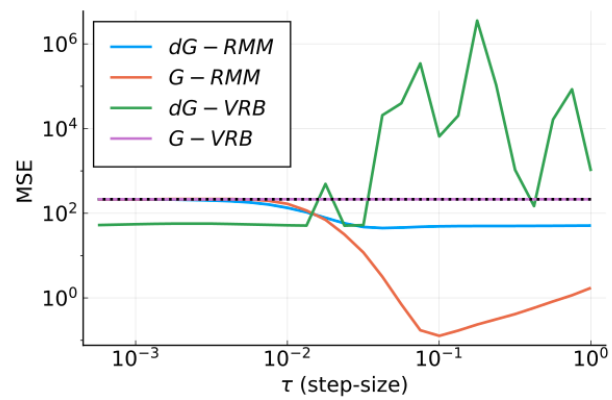
This confirms that the lack of Euclidean smoothness of $f_\pi^{(\alpha)}$ translates numerically into a high level of instability of VRB with respect to the choice of the step-size. On the contrary, the RMM algorithm has a more stable behavior even for poorly chosen step-sizes, confirming the theoretical study of Section 5.

References

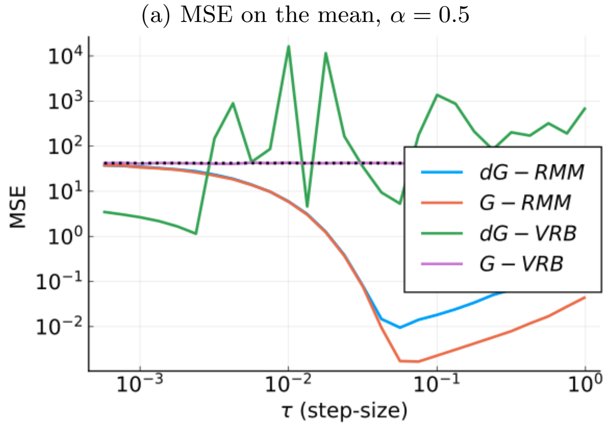
[1] O. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(12), 2021.



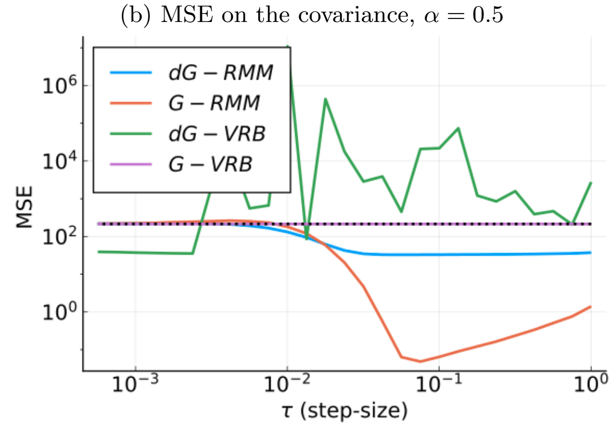
(a) MSE on the mean, $\alpha = 0.5$



(b) MSE on the covariance, $\alpha = 0.5$



(c) MSE on the mean, $\alpha = 1.0$



(d) MSE on the covariance, $\alpha = 1.0$

Figure 7: MSE in the estimation of $\bar{\mu}$ and $\bar{\Sigma}_\kappa$ ($d = 5$) after 100 iterations, against values of τ. For each value of τ, 10^3 runs with 500 samples per iteration are conducted. The dotted black lines represent the MSE at initialization. The prefix dG refer to the family of diagonal Gaussians, while the prefix G refers to Gaussians with full covariance.

[2] S. Amari. *Differential-Geometrical Methods in Statistics*. Springer New York, 1985.

[3] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[4] Y. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

[5] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9(15):485–516, 2008.

[6] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, Ltd, 2014.

[7] H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[8] H. Bauschke and J. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4:27–67, 1997.

[9] H. Bauschke, J. Borwein, and P. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[10] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, 2011.

[11] A. Benfenati, E. Chouzenoux, and J.-C. Pesquet. Proximal approaches for matrix optimization problems: Application to robust precision matrix estimation. *Signal Processing*, 169, 2020.

[12] T. Bengsston, P. Bickel, and B. Li. *Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems*, pages 316–334. Institute of Mathematical Statistics, 2008.

[13] D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for the statistician. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[14] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[15] L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, 1986.

[16] M. F. Bugallo, V. Elvira, and L. Martino. A new strategy for effective learning in population Monte Carlo sampling. In *Asilomar Conference on Signals, Systems and Computers*, pages 1540–1544, 2016.

[17] T. Bui. Connecting the thermodynamic variational objective and annealed importance sampling. Technical report, 2020.

[18] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger. A Bregman learning framework for sparse neural networks. *Journal of Machine Learning Research*, 23(192):1–43, 2022.

[19] D. Butnariu and A. Iusem. *Totally convex functions*, chapter 1. Kluwer Academic Publisher, 2000.

[20] T. Campbell and B. Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems (NeurPIS)*, volume 32, 2019.

[21] O. Cappé, R. Douc, A. Guillin, J. Marin, and C. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.

[22] N. L. Carothers. *Real Analysis.* Cambridge University Press, 2000.

[23] P. Combettes and J.-C. Pesquet. *Proximal Splitting Methods in Signal Processing*, page 185–212. Springer-Verlag, New York, 2010.

[24] P. L. Combettes and C. L. Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications*, 457(2):1283–1306, 2018.

[25] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, December 2012.

[26] I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial.* Now Foundations and Trends, 2004.

[27] K. Daudel, R. Douc, and F. Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49(4):2250–2270, 2021.

[28] K. Daudel, R. Douc, and F. Roueff. Monotonic Alpha-divergence minimization. *Journal of Machine Learning Research*, 62(24):1–76, 2023.

[29] A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[30] J. Domke. Provable smoothness guarantees for black-box variational inference. In *International Conference on Machine Learning (ICML)*, pages 2587–2596, 2020.

[31] R. Douc, A. Guillin, J. Marin, and C. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Annals of Statistics*, 35:420–448, 2007.

[32] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–49, 2019.

[33] R. Dykstra. An iterative procedure for obtaining I-projections onto the intersection of convex sets. *The Annals of Probability*, 13(3):975–984, 1985.

[34] Y. El-Laham, V. Elvira, and M. F. Bugallo. Recursive shrinkage covariance learning in adaptive importance sampling. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 624–628, 2019.

[35] Y. El-Laham, L. Martino, V. Elvira, and M. Bugallo. Efficient adaptive multiple importance sampling. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

[36] V. Elvira, L. Martino, D. Luengo, and M. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

[37] T. Gao, S. Lu, J. Liu, and C. Chu. Randomized Bregman coordinate descent methods for non-Lipschitz optimization. https://arxiv.org/pdf/2001.05202, 2020.

[38] M. E. Gheche, G. Chierchia, and J.-C. Pesquet. Proximity operators of discrete information divergences. *IEEE Transactions on Information Theory*, 64(2):1092–1104, 2018.

[39] R. B. Grosse, C. J. Maddison, and R. R. Salakhutdinov. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

[40] F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *Computational Optimization and Applications*, 79(3):717–766, 2021.

[41] T. Hastie, R. Tibshirani, and J. Firedman. *The Elements of Statistical Learning.* Springer, 2009.

[42] J. Hensman, M. Rattray, and N. Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, 2012.

[43] J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning (ICML)*, volume 48, pages 1511–1520, 2016.

[44] J.-B. Hiriart-Urruty and C. Lemaréchal. *Abstract Duality for Practitioners*, pages 137–193. Springer, 1993.

[45] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

[46] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11(106):3235–3268, 2010.

[47] Y. Huang, E. Chouzenoux, and J.-C. Pesquet. Unrolled variational Bayesian algorithm for image blind deconvolution. *IEEE Transactions on Image Processing*, 32:430–445, 2022.

[48] E. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[49] G. Ji, D. Sujono, and E. Sudderth. Marginalized stochastic natural gradients for black-box variational inference. In *International Conference on Machine Learning (ICML)*, volume 139, pages 4870–4881, 2021.

[50] H. Karimi, J. Nutini, and M. W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2016.

[51] M. Khan, R. Babanezhad, W. Lin, M. Schmidt, and M. Sugiyama. Faster stochastic variational inference using proximal-gradient methods with general divergence functions. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, page 319–328, 2016.

[52] M. Khan and W. Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 878–887, 2017.

[53] J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

[54] E. Koblents and J. Míguez. A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing*, 25(2):407–425, 2013.

[55] A. Korba and F. Portier. Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 11503–11527, 2022.

[56] M. A. Kumar and I. Sason. Projection theorems for the Rényi divergence on α-convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935, 2016.

[57] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems (NeurPIS)*, volume 35, 2022.

[58] Y. Li and R. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[59] W. Lin, M. Khan, and M. Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning (ICML)*, volume 97, pages 3992–4002, 2019.

[60] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet. Majorize–minimize adapted Metropolis–Hastings algorithm. *IEEE Transactions on Signal Processing*, 68:2356–2369, 2020.

[61] L. Martino, V. Elvira, D. Luengo, and J. Corander. An adaptive population importance sampler. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8038–8042, 2014.

[62] L. Martino, V. Elvira, J. Míguez, A. Artés-Rodríguez, and P. Djurić. A comparison of clipping strategies for importance sampling. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 558–562, 2018.

[63] V. Masrani, R. Brekelmans, T. Bui, F. Nielsen, A. Galstyan, G. V. Steeg, and F. Wood. q-paths: Generalizing the geometric annealing path using power means. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161, pages 1938–1947, 2021.

[64] T. Minka. Divergence measures and message passing. Technical report, 2005.

[65] P. D. Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

[66] J. Moré and G. Toraldo. Algorithms for bound contrained quadratic programming problems. *Numerische Mathematik*, 55(4):377–400, 1989.

[67] M. Mukkamala, P. Ochs, T. Pock, and S. Sabach. Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. *SIAM Journal on Mathematics of Data Science*, 2(3):658–682, 2020.

[68] R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[69] F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *IEEE International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.

[70] J.-P. Penot. *Calculus without Derivatives*. Springer, 2013.

[71] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33, pages 814–822, 2014.

[72] A. Saha, K. Barath, and S. Kurtek. A geometric variational approach to Bayesian inference. *Journal of the American Statistical Association*, 115(530):822–835, 2020.

[73] Y. Shao, Y. Zhou, and D. Cai. Variational inference with graph regularization for image annotation. *ACM Transactions on Intelligent Systems and Technology*, 2(2):1–21, 2011.

[74] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[75] M. Titsias and M. Lázaro-Gredilla. Local expectation gradients for black box variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

[76] T. van Erven and P. Harremoes. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions of Information Theory*, 60(7):3797–3820, 2014.

[77] A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. https://arxiv.org/abs/1507.02646, 2015.

[78] S. Vempala and A. Wibosono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[79] X. Xiao. A unified convergence analysis of stochastic Bregman proximal gradient and extra-gradients methods. *Journal of Optimization Theory and Applications*, 188(3):605–627, 2021.

[80] R. Yao and Y. Yang. Mean field variational inference via Wasserstein gradient flow. https://arxiv.org/abs/2207.08074, 2022.

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

[82] Y. Zheng, A. Fraysse, and T. Rodet. Efficient unsupervised variational Bayesian image reconstruction using a sparse gradient prior. *Neurocomputing*, 359:449–465, 2019.