

Regularized Rényi Divergence Minimization through Bregman Proximal Gradient Algorithms

Thomas Guilmeau

Emilie Chouzenoux

Centre pour la Vision Numérique

Université Paris-Saclay, INRIA, CentraleSupélec

9 rue Joliot Curie

91190, Gif-sur-Yvette, France

THOMAS.GUILMEAU@INRIA.FR

EMILIE.CHOUZENOUX@INRIA.FR

Víctor Elvira

School of Mathematics

University of Edinburgh

James Clerk Maxwell Building, The King's Buildings

Peter Guthrie Tait Road

Edinburgh, EH9 3FD, United Kingdom

VICTOR.ELVIRA@ED.AC.UK

Editor: Matthew Hoffman

Abstract

We study the variational inference problem of minimizing a regularized Rényi divergence over an exponential family. We propose to solve this problem with a Bregman proximal gradient algorithm. We propose a sampling-based algorithm to cover the black-box setting, corresponding to a stochastic Bregman proximal gradient algorithm with biased gradient estimator. We show that the resulting algorithms can be seen as relaxed moment-matching algorithms with an additional proximal step. Using Bregman updates instead of Euclidean ones allows us to exploit the geometry of our approximate model. We prove strong convergence guarantees for both our deterministic and stochastic algorithms using this viewpoint, 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.

1 Introduction

Probability distributions of interest in statistical problems are often intractable. In Bayesian statistics, for instance, the targeted posterior distribution often cannot be evaluated in a closed-form nor sampled due to intractable normalization constants. Variational inference (VI) methods aim at finding accurate and tractable approximations by minimizing a divergence to the target over a family of parametric distributions (Blei et al., 2017; Zhang et al., 2019). Such procedures can be summarized by the choice of the approximating density family, the choice of the divergence, and the design of 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 is then found by a deterministic coordinate-ascent algorithm (Hoffman et al., 2013).

The research on VI methods has been very active in the last years (see the review of Zhang et al. 2019). Majorization techniques have been proposed to cope with large scale models not satisfying conjugacy hypotheses (Marnissi et al., 2020; Zheng et al., 2019; Huang et al., 2022). Another approach in such challenging context is to run a stochastic gradient descent, which leads to the so-called black-box VI methods (Ranganath et al., 2014). Black-box methods allow to handle a broad choice of divergence, such as the α -divergences (Hernandez-Lobato et al., 2016; Dieng et al., 2017; Daudel et al., 2021) and Rényi divergences (Li and Turner, 2016), themselves generalizations of the KL divergence depending on a scalar parameter $\alpha > 0$. The latter parameter can be chosen in order to enforce a mode-seeking or a mass-covering behavior in the approximations. For instance, minimizing the exclusive KL divergence leads to under-estimation of the variance of the target (Minka, 2005; Margossian and Saul, 2023).

VI algorithms have also benefited from the recent advances in information geometry, a field that studies statistical models through a differential-geometric lens. Among available 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 (Amari and Nagaoka, 2000). Another useful insight, when exponential families are considered, is the relation between the KL divergence, Bregman divergences, and dual geometry (Nielsen and Nock, 2010). These ideas can be leveraged by using the *natural gradient* (Amari, 1998), which amounts to preconditioning the standard (i.e., Euclidean) gradient by the inverse Fisher information matrix. In the VI algorithms investigated in (Honkela et al., 2010; Hensman et al., 2012; Hoffman et al., 2013; Khan and Nielsen, 2018; Lin et al., 2019), the standard gradient of the evidence lower bound (ELBO) is thus adjusted to take into account the Riemannian geometry of the approximating distributions, leading to simpler updates and an improved behavior. Another related approach to exploit the non-Euclidean geometry of the model is to formulate the gradient updates within the geometry induced by a general divergence. This has been explored for VI in (Khan et al., 2016; Khan and Lin, 2017) and coincides with natural gradients when the exponential family is used.

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

First, to the best of our knowledge, natural gradient methods, and more generally, non-Euclidean optimization methods, are restricted to the minimization of the exclusive KL divergence. This is the case, for instance, of the black-box procedures leveraging natural gradients presented in (Khan and Nielsen, 2018; Lin et al., 2019; Ji et al., 2021), of the black-box methods using the geometry induced by a divergence (Khan et al., 2016; Khan and Lin, 2017), and of the methods presented in (Yao and Yang, 2022; Lambert et al., 2022) leveraging a Wasserstein-based geometry. Let us however mention the work of Saha et al. (2020) that studies the minimization of an α -divergence over a mean-field family using the Fisher Riemannian geometry.

Second, convergence studies of VI schemes are mostly empirical for black-box VI schemes, whether Euclidean geometry (Titsias and Lázaro-Gredilla, 2015; Li and Turner, 2016; Hernandez-Lobato et al., 2016; Dieng et al., 2017), natural gradients (Honkela et al., 2010; Hensman et al., 2012; Hoffman et al., 2013; Khan and Nielsen, 2018; Lin et al., 2019), or non-Euclidean geometry (Khan et al., 2016; Khan and Lin, 2017), are used. Indeed, the considered optimization problems are non-convex with noisy gradient, making the algorithms hard to analyze. This is in stark contrast with MCMC methods, which can be used alternatively to VI. MCMC methods based on Langevin diffusion are guaranteed to asymptotically produce samples from the target but also benefit from non-asymptotic convergence guarantees (Vempala and Wibosono, 2019; Durmus et al., 2019) under log-concavity assumptions on the target. Recent VI works have started to close the gap under smoothness and log-concavity assumptions on the target and for specific approximating families. For instance, VI algorithms leveraging the Wasserstein geometry have been shown to converge to the minimizer of the exclusive KL divergence, in the case of a mean-field approximating family (Yao and Yang, 2022) or a Gaussian approximating family (Lambert et al., 2022). Related are the convergence guarantees for standard stochastic gradient descent algorithms that have been achieved in (Kim et al., 2023; Domke et al., 2023) for location-scale approximating families and exclusive KL divergence.

1.1 Contributions and outline

In this paper, we consider the minimization of a regularized Rényi divergence between the target and an exponential family. We propose a novel black-box VI algorithm that leverages the geometry induced by the Kullback-Leibler divergence and benefits from solid convergence guarantees.

More precisely, we propose a Bregman proximal gradient algorithm. These recent (possibly stochastic) optimization algorithms (Bauschke et al., 2003, 2017; Teboulle, 2018; Mukkamala et al., 2020; Hanzely and Richtárik, 2021; Xiao, 2021) arise from the generalization of the Euclidean proximal minimization schemes (Combettes and Pesquet, 2010). This general approach allows to tailor the intrinsic geometry of optimization problems by choosing a suitable Bregman divergence (Bauschke et al., 2017; Teboulle, 2018).

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

- We propose a Bregman proximal gradient algorithm to minimize a Rényi divergence over an exponential family. The Bregman divergence that we use is induced by the KL divergence and as such, our algorithm exploits the geometry of the approximating family. We also propose a sampling-based stochastic implementation for our method allowing to cope with the black-box setting. The corresponding scheme is a stochastic Bregman proximal gradient method with biased gradient estimations. Our algorithms also allow to add a regularizer function to the divergence to enforce solutions with specific features. We show that our algorithms can be seen as proximal relaxed moment-matching algorithms generalizing existing methods.
- We analyze the convergence of our deterministic and stochastic algorithms using a combination of existing and novel techniques for the study of Bregman proximal gra-

dient methods. In the deterministic setting, we establish monotonic decrease of the objective and show that limit points and fixed points of our algorithms are stationary points. We establish geometric rates of convergence provided that the target belongs to the approximating family or that the Rényi divergence is the inclusive KL divergence. In the stochastic setting, we show that gradients (possibly subgradients) of the objective converge to zero. When the Rényi divergence is the inclusive KL divergence, we also give convergence rates, which depend on the step sizes and number of samples. These results are achieved with mild assumptions on the target and on the sampling procedure and can be applied to existing moment-matching methods.

- We demonstrate the performance of our algorithms when the approximating family is Gaussian. In this case, we show that our algorithm is faster and more robust than algorithms based on the Euclidean geometry. We also show that using an additional regularizing term can efficiently enforce sparse solutions. We also highlight how the choice of Rényi divergence allows to create mass-covering or mode-seeking approximations and to compensate high approximation errors.

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 proposed algorithm, then we give an alternative, stochastic, black-box implementation for it. Theoretical analysis of our algorithms, both deterministic and stochastic, are provided in Section 4. Finally, numerical experiments with Gaussian approximating distributions are presented in Section 5. We discuss our results and possible future research lines in Section 6.

The supplementary material contains four appendices. The proofs of our results are deferred to Appendices A and B, while we construct a sparsity-enforcing 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 $\llbracket n_1, n_2 \rrbracket$. 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 $\text{int } C$. The set of non-negative real numbers is denoted 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 $f, g : \mathcal{H} \rightarrow \mathbb{R}$, 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 (Bauschke and Combettes, 2011). 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 $f : \mathcal{H} \rightarrow [-\infty, +\infty]$ is $\text{dom } f := \{\theta \in \mathcal{H}, f(\theta) < +\infty\}$. The indicator

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 $h : \mathcal{X} \rightarrow \mathcal{H}$, we denote by $m(h)$ the vector of \mathcal{H} defined by $(m(h))_i = \int_{\mathcal{X}} h_i(x) m(dx)$ for $i \in \llbracket 1, n \rrbracket$. 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 van Erven and Harremoës (2014), since

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

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

Proposition 3 (van Erven and Harremoës (2014)) *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 (Brown, 1986; Barndorff-Nielsen, 2014).

Definition 4 *Let $\Gamma : \mathcal{X} \rightarrow \mathcal{H}$ 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*

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

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

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

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 5 *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 (Barndorff-Nielsen, 2014), with sufficient statistics $\Gamma : x \mapsto (x, xx^\top)^\top$ 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 $\Theta = \mathbb{R}^d \times (-\mathcal{S}_{++}^d)$, 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 (Amari and Nagaoka, 2000; Nielsen and Nock, 2010) and have been used as approximating families in many contexts such as VI algorithms (Hensman et al., 2012; Hoffman et al., 2013; Blei et al., 2017; Lin et al., 2019), or adaptive importance sampling (AIS) procedures (Akyildiz and Míguez, 2021).

We now give some background about the geometric structure of the parameters of an exponential family. To this end, we first introduce the concepts of Legendre functions and Bregman divergences.

Definition 6 *A Legendre function is a function $B \in \Gamma_0(\mathcal{H})$ that is strictly convex on the interior of its domain $\text{int dom } B$, and essentially smooth. B is essentially smooth if it is differentiable on $\text{int dom } B$ and such that $\|\nabla B(\theta_k)\| \xrightarrow[k \rightarrow +\infty]{} +\infty$ for every sequence $\{\theta_k\}_{k \in \mathbb{N}}$ converging to a boundary point of $\text{dom } B$ with $\theta_k \in \text{int dom } B$ for every $k \in \mathbb{N}$.*

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

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

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 . Bregman divergences generalize the Euclidean norm, since it is recovered for $B(\theta) = \frac{1}{2}\|\theta\|^2$ (Bauschke et al., 2017). We now recall the definition of conjugate functions (Bauschke and Combettes, 2011), which allows to state some useful properties of Legendre functions.

Definition 7 *The conjugate of a function $f : \mathcal{H} \rightarrow [-\infty, +\infty]$ is the function $f^* : \mathcal{H} \rightarrow [-\infty, +\infty]$ such that*

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

Proposition 8 (Section 2.2 in Teboulle 2018) *Let B be a Legendre function. Then we have that*

- (i) ∇B is a bijection from $\text{int dom } B$ to $\text{int dom } B^*$, and $(\nabla B)^{-1} = \nabla B^*$,
- (ii) $\text{dom } \partial B = \text{int dom } B$ and $\partial B(\theta) = \{\nabla B(\theta)\}$, $\forall \theta \in \text{int dom } B$,
- (iii) B is a Legendre function if and only if B^* is a Legendre function,
- (iv) for every $\theta \in \text{dom } B$, $\theta' \in \text{int dom } B$, $d_B(\theta, \theta') \geq 0$ with equality if and only if $\theta = \theta'$.

Proposition 8 (iv) shows that Legendre functions can be used to create Bregman divergence with a distance-like property. Note, however, that Bregman divergences are not symmetric nor do they satisfy the triangular inequality in general. Proposition 8 also shows that the gradients of Legendre functions are bijections between $\text{int dom } B$ and $\text{int dom } B^*$, with inverse ∇B^* . We will now recall results showing that the log-partition function defined in (2) is a Legendre function under minimal assumptions. This makes the log-partition function a natural choice to generate a Bregman divergence and allows to define a bijection between the parameters and the moments of distributions from considered family.

Proposition 9 (Brown (1986); Barndorff-Nielsen (2014)) *Under the hypothesis that $\text{int } \Theta \neq \emptyset$, 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 $\text{int } \Theta$. In particular, its gradient reads*

$$\nabla A(\theta) = q_\theta(\Gamma), \forall \theta \in \text{int } \Theta. \quad (3)$$

If \mathcal{Q} is minimal and steep (see (Barndorff-Nielsen, 2014, Chapter 8) for more details on these notions), then the log-partition function is a Legendre function.

Minimality ensures that each distribution in \mathcal{Q} can be parametrized only by a unique parameter θ . Steepness is satisfied by most exponential families. It is in particular implied by having $\Theta = \text{dom } A$ being open (Barndorff-Nielsen, 2014, Theorem 8.2). More precisely, the results of Proposition 9 come from (Brown, 1986, Theorem 1.13) for the convexity results, (Barndorff-Nielsen, 2014, Theorem 8.1) for the differentiability result, and (Barndorff-Nielsen, 2014, Eq. (20)) for the steepness part.

The Legendre property on A allows to benefit from the results of Proposition 8, implying in particular that there is a bijection between the parameter θ and the moments $q_\theta(\Gamma)$. This leads to an alternative parametrization of q_θ in terms of its moments, which are often called the dual parameters. The Bregman divergence induced by the Legendre function A admits a statistical interpretation that has been well-studied in the information geometry community (Nielsen and Nock, 2010). 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 10 (Nielsen and Nock 2010) *Consider $\theta, \theta' \in \text{int } \Theta$ and the log-partition function A defined in (2). Then,*

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

This results gives a natural choice of Bregman divergence to use in the context of a Bregman proximal gradient algorithm.

2.3 Considered VI problem

In this work, 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 $\alpha > 0$:

$$f_\pi^{(\alpha)}(\theta) := \begin{cases} RD_\alpha(\pi, q_\theta), & \text{if } \alpha \neq 1, \\ KL(\pi, q_\theta), & \text{if } \alpha = 1, \end{cases} \forall \theta \in \Theta. \quad (4)$$

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

$$F_\pi^{(\alpha)}(\theta) := f_\pi^{(\alpha)}(\theta) + r(\theta), \forall \theta \in \Theta. \quad (5)$$

We propose to resolve our approximation problem by minimizing (5) over an exponential family \mathcal{Q} , i.e., by considering the following optimization problem:

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

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 (Minka, 2005; Blei et al., 2017). 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 (Li and Turner, 2016), hence fine-tuning the algorithm’s behavior for the application at hand. Moreover, the Rényi divergence with parameter α can be monotonically transformed (van Erven and Harremoës, 2014) into the corresponding α -divergence (Hernandez-Lobato et al., 2016; Daudel et al., 2021), including in particular the χ^2 divergence (Dieng et al., 2017; Akyildiz and Míguez, 2021).

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 (Dykstra, 1985; Csiszár and Shields, 2004) when $\alpha = 1$, which has later been generalized by Kumar and Sason (2016) for $\alpha \neq 1$. A similar setting is used in sparse precision matrix estimation, relying on the KL divergence and a sparsity-inducing regularizer (Banerjee et al., 2008). The problem of computing Bayesian coresets has also been formulated as a KL minimization problem over a set of sparse parameters by Campbell and Beronov (2019). Let us also mention that Shao et al. (2011) 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 Knoblauch et al. (2022).

3 A proximal relaxed moment-matching algorithm

Let us now propose our algorithm to solve Problem $(P_\pi^{(\alpha)})$. This algorithm is a Bregman proximal gradient algorithm, for which we first give an exact version and show that it can be interpreted as a relaxed moment-matching algorithm. We then propose a sampling-based implementation of this algorithm, before discussing its links with existing algorithms in the field of computational statistics.

3.1 An exact Bregman proximal gradient algorithm

Problem $(P_\pi^{(\alpha)})$ is a composite problem since the objective is the sum of two terms. We propose to solve this problem using a Bregman proximal gradient algorithm, where the gradient step is used for the Rényi divergence term and the proximal step is used upon the regularizer. We first define these two operators, following (Bauschke et al., 2017), before giving the full algorithm.

Definition 11 *Consider a positive step-size $\tau > 0$.*

- (i) *The Bregman proximal operator of $\theta \mapsto \tau r(\theta)$ is defined for every $\theta \in \text{int dom } A$ by*

$$\text{prox}_{\tau r}^A(\theta) := \arg \min_{\theta' \in \text{dom } A} \left(r(\theta') + \frac{1}{\tau} d_A(\theta', \theta) \right).$$

(ii) The Bregman gradient descent operator of $\theta \mapsto \tau f_\pi^{(\alpha)}(\theta)$ is defined for every $\theta \in \text{int dom } A$ by

$$\gamma_{\tau f_\pi^{(\alpha)}}^A(\theta) := \arg \min_{\theta' \in \text{dom } A} \left(f_\pi^{(\alpha)}(\theta) + \langle \nabla f_\pi^{(\alpha)}(\theta), \theta' - \theta \rangle + \frac{1}{\tau} d_A(\theta', \theta) \right).$$

(iii) The Bregman proximal gradient operator of $\theta \mapsto \tau F_\pi^{(\alpha)}(\theta)$ is defined for every $\theta \in \text{int dom } A$ by

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

These operators leverage the geometry of the KL divergence over the exponential family. Indeed, we use the Bregman divergence induced by the log-partition function which is the KL divergence between two densities from \mathcal{Q} as shown in Proposition 10. We now leverage these operators to provide our algorithm solving Problem $(P_\pi^{(\alpha)})$.

Algorithm 1: Proposed Bregman proximal gradient 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 $\theta_0 \in \text{int } \Theta$.

for $k = 0, \dots$ **do**

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

$$\theta_{k+\frac{1}{2}} = \gamma_{\tau_{k+1} f_\pi^{(\alpha)}}^A(\theta_k) \tag{6}$$

 Update θ_{k+1} following

$$\theta_{k+1} = \text{prox}_{\tau_{k+1} r}^A(\theta_{k+\frac{1}{2}}). \tag{7}$$

end

The Bregman gradient step in Algorithm 1 involves the computation of $\nabla f_\pi^{(\alpha)}$, ∇A , and ∇A^* . We now show that this step can be written explicitly in terms of moments of specific distributions, so as to outline a relaxed moment-matching interpretation of Algorithm 1. One of these distributions combines the target and the proposal distributions as follows.

Definition 12 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

$$\pi_\theta^{(\alpha)}(x) = \frac{1}{\int \pi(y)^\alpha q_\theta(y)^{1-\alpha} \nu(dy)} \left(\pi(x)^\alpha q_\theta(x)^{1-\alpha} \right), \forall x \in \mathcal{X}. \tag{8}$$

Probability densities akin to $\pi_\theta^{(\alpha)}$ have been used for instance in annealed importance sampling (Neal, 2001), in sequential Monte-Carlo schemes (Moral et al., 2006), or in adaptive importance sampling (Bugallo et al., 2016). The integral in (8) 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. We now show that these densities are linked with the gradients of the Rényi divergence.

Proposition 13 *Let $\alpha > 0$. The map $f_\pi^{(\alpha)}$ is of class \mathcal{C}^2 on $\text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$. In particular, for any $\theta \in \text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$,*

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

We can now show that the Bregman gradient update in Algorithm 1 can be seen as a relaxed moment-matching update. Indeed, the moment of the next proposal is a convex combination of two moments: the moment of the geometric average between the target and the current proposal, and the moment of the current proposal. This result is stated under a well-posedness assumption that we investigate in Section 4. We explicit this update in the case of Gaussian proposals.

Proposition 14 *Assume that \mathcal{Q} is minimal and steep, and consider the sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 1. If $\theta_k \in \text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$ and if $\theta_{k+\frac{1}{2}} = \gamma_{\tau_{k+1} f_\pi^{(\alpha)}}^A(\theta_k)$ is well-defined, belonging to $\text{int } \Theta$, then $\theta_{k+\frac{1}{2}}$ satisfies*

$$q_{\theta_{k+\frac{1}{2}}}(\Gamma) = \tau_{k+1} \pi_{\theta_k}^{(\alpha)}(\Gamma) + (1 - \tau_{k+1}) q_{\theta_k}(\Gamma). \quad (9)$$

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*

$$\begin{cases} q_{\theta_{k+\frac{1}{2}}}(x) &= \tau_{k+1} \pi_{\theta_k}^{(\alpha)}(x) + (1 - \tau_{k+1}) q_{\theta_k}(x), \\ q_{\theta_{k+\frac{1}{2}}}(xx^\top) &= \tau_{k+1} \pi_{\theta_k}^{(\alpha)}(xx^\top) + (1 - \tau_{k+1}) q_{\theta_k}(xx^\top). \end{cases} \quad (10)$$

This shows that (6) matches 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 (10) is equivalent to

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

The proximal operator $\text{prox}_{\tau r}^A$ needs to be computed case-by-case depending on the choice of regularizer r . We give hereafter an example that illustrates the applicability of this second step of Algorithm 1. Our example links Eq. (7) with reverse information projections (Dykstra, 1985; Csiszár and Shields, 2004). Explicit expression for this step is provided in Appendix C for a sparsity-inducing regularizer.

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 $C \subset \mathcal{H}$ (Bauschke and Combettes, 2011, Example 12.25). We obtain in this case

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

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 (Csiszár and Shields, 2004, Section 3) for instance.

3.2 Sampling-based implementation

We have shown in Proposition 14 that implementing Algorithm 1 requires the moments of the current proposal as well as the moments of its geometric average with the target distribution. While the moments of the proposals are known, the quantity $\pi_\theta^{(\alpha)}(\Gamma)$ is unavailable. We therefore propose, in Algorithm 2, a sampling-based form for Algorithm 1, where we approximate the moments $\pi_{\theta_k}^{(\alpha)}(\Gamma)$ at every iteration $k \in \mathbb{N}$ using samples from q_{θ_k} .

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 $\theta_0 \in \text{int } \Theta$.

for $k = 0, \dots$ **do**

For every $l \in \llbracket 1, N_{k+1} \rrbracket$, sample $x_l \sim q_{\theta_k}$ and compute the non-linear importance weight $w_l^{(\alpha)}$ with their normalized counterpart $\bar{w}_l^{(\alpha)}$ through

$$w_l^{(\alpha)} = \left(\frac{\tilde{\pi}(x_l)}{q_{\theta_k}(x_l)} \right)^\alpha, \quad \bar{w}_l^{(\alpha)} = \frac{w_l^{(\alpha)}}{\sum_{l=1}^{N_{k+1}} w_l^{(\alpha)}}. \quad (11)$$

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

$$q_{\theta_{k+\frac{1}{2}}}(\Gamma) = \tau_{k+1} \left(\sum_{l=1}^{N_{k+1}} \bar{w}_l^{(\alpha)} \Gamma(x_l) \right) + (1 - \tau_{k+1}) q_{\theta_k}(\Gamma). \quad (12)$$

Compute $\theta_{k+1} = \text{prox}_{\tau r}^A(\theta_{k+\frac{1}{2}})$ as in Eq. (7).

end

Algorithm 2 is written using an approximated moment-matching update instead of a Bregman gradient update. 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 13 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 $\tilde{G}_\pi^{(\alpha)}(\theta_k)$. Following the proof of Proposition 14 in the reversed direction, we can interpret Algorithm 2 as a stochastic Bregman proximal

gradient algorithm (Xiao, 2021). Indeed, $\theta_{k+\frac{1}{2}}$ is updated in Algorithm 2 through

$$\theta_{k+\frac{1}{2}} = \arg \min_{\theta' \in \text{dom } A} \left(f_{\pi}^{(\alpha)}(\theta_k) + \langle \tilde{G}_{\pi}^{(\alpha)}(\theta_k), \theta' - \theta_k \rangle + \frac{1}{\tau_{k+1}} d_A(\theta', \theta_k) \right) \quad (13)$$

which is a stochastic version of the Bregman gradient operator $\gamma_{\tau_{k+1} f_{\pi}^{(\alpha)}}^A$.

3.3 Comparison with existing methods

In Section 3.1, we have proposed Algorithm 1, that is a Bregman proximal gradient algorithm to solve Problem $(P_{\pi}^{(\alpha)})$ and we have shown that it is equivalent to a relaxed moment-matching algorithm with additional proximal step. In Section 3.2, we have provided Algorithm 2, which is a sampling-based approximation of Algorithm 1.

Let us now position these algorithms with respect to existing works. Algorithm 1 is a Bregman proximal gradient algorithm, meaning that it does not rely on the Euclidean geometry, but on another geometry induced by a Bregman divergence (Bauschke et al., 2017; Teboulle, 2018). In our case, we use a Bregman divergence related to the KL divergence (Nielsen and Nock, 2010). Bregman proximal gradient algorithms whose Bregman divergence corresponding to the KL divergence have been used in (Khan et al., 2016; Khan and Lin, 2017). The goal of these works was to minimize the exclusive KL divergence (which is mode-seeking) without any regularization, the proximal splitting being motivated by the assumption that the target is the product of a conjugate and a non-conjugate term. In contrast, here, our method aims at minimizing the sum of a Rényi divergence, whose mode-seeking or mass covering behavior can be tuned, with a possible non-smooth regularizer. The analysis of the schemes in (Khan et al., 2016; Khan and Lin, 2017) lies on Euclidean smoothness assumptions for which we exhibit simple counter-examples in Section 4.1.

Natural gradient methods also exploit the geometry of the approximating distributions and have been used in VI to minimize the exclusive KL divergence in (Honkela et al., 2010; Hoffman et al., 2013; Khan and Nielsen, 2018; Lin et al., 2019). Bregman gradient descent and natural gradient descent methods share close ties in the case of exponential families, as shown by Raskutti and Mukherjee (2015). More explicitly, for exponential families, a Bregman gradient descent step in the variable θ (as it is done in Algorithm 1) is equivalent to a natural gradient descent step in the variable $\nabla A(\theta)$, with the metric tensor being $\nabla^2 A^*$ instead of $\nabla^2 A$, the latter being equal to the Fisher information matrix (Amari and Nagaoka, 2000). The expression of natural gradients when the exponential family is used can be found in (Khan and Nielsen, 2018, Theorem 1). Note also that Bregman gradient descent can be interpreted as a mirror descent algorithm (see Beck and Teboulle (2003) for instance). To our knowledge, there is no counterpart of natural gradient descent for proximal operators, limiting the use of previously mentioned natural gradient methods to smooth objectives. In contrast, our point of view allows for a non-smooth regularizing term.

Recent VI methods have also leveraged the Wasserstein geometry to design efficient algorithms for VI. This is a different point of view from previously discussed methods, that are based on measuring the difference between approximating distributions with the KL divergence. A discussion about the difference between the KL divergence and the Wasserstein distance can be found in (Khan and Zhang, 2022). Let us mention in this

direction (Yao and Yang, 2022; Lambert et al., 2022) which aim at minimizing the exclusive KL divergence over, respectively, mean-field approximations and Gaussian approximations.

Regarding the minimization of divergences other than the exclusive KL divergence in the black-box context, we can mention the α -divergence minimization scheme of Daudel et al. (2023), which is the closest scheme to ours. Indeed, in the case when their algorithm is used over an exponential family, its updates are similar to the relaxed moment-matching updates described in Proposition 14. Although the geometry used in (Daudel et al., 2023) is not explicated, it appears to be related to the geometry induced by the KL divergence (we discuss this matter in depth in Section 4). Apart from this work, most VI works with "alternative" divergences use standard gradient descent, meaning that they use the Euclidean geometry (Hernandez-Lobato et al., 2016; Li and Turner, 2016; Dieng et al., 2017). All of the previously mentioned works allow for wider families than the exponential family, but they do not consider the possibility of adding a regularizer, and their theoretical guarantees are weaker than the ones we provide in Section 4 for our method.

For the sake of illustration, we explicit, in the case of proposals forming an exponential family, the variational Rényi bound (VRB) algorithm, which is an Euclidean algorithm proposed in (Li and Turner, 2016). In the work of Li and Turner (2016), the VRB is an alternative objective akin to the evidence lower bound that does not involve the unknown normalization constant Z_π is constructed from $\theta \mapsto RD_\alpha(q_\theta, \pi)$. Consider in the following $\alpha \in (0, 1)$, and $\theta \in \text{int } \Theta$. 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 (van Erven and Harremoës, 2014, Proposition 2). Therefore, minimizing $\theta \mapsto RD_{1-\alpha}(q_\theta, \pi)$ is equivalent to maximizing

$$\mathcal{L}_\pi^{(\alpha)}(\theta) := \frac{1}{\alpha} \log \left(\int \tilde{\pi}(x)^\alpha q_\theta(x)^{1-\alpha} \nu(dx) \right). \quad (14)$$

Now, following the same computations as in the proof of Proposition 13, we obtain that $\nabla \mathcal{L}_\pi^{(\alpha)} = -\frac{1-\alpha}{\alpha} \nabla f_\pi^{(\alpha)}$. Therefore, the gradient ascent algorithm to maximize $\mathcal{L}_\pi^{(\alpha)}$ on Θ reads as

$$\theta_{k+1} = \theta_k - \tau_{k+1} \nabla f_\pi^{(\alpha)}(\theta_k), \quad (15)$$

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

$$\theta_{k+1} = \theta_k + \tau_{k+1} \left(\sum_{l=1}^{N_{k+1}} \bar{w}_l^{(\alpha)} \Gamma(x_l) - q_{\theta_k}(\Gamma) \right), \quad (16)$$

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

Proposition 14 unveils the moment-matching interpretation of Algorithm 1 and establishes links between our algorithms and the numerous algorithms in statistics resorting to moment-matching updates. The link between moment-matching and inclusive KL minimization is well-known (Bishop, 2006). Indeed, the minimizer of Problem $(P_\pi^{(\alpha)})$ when $\alpha = 1$ and $r \equiv 0$ is the distribution $q_\theta \in \mathcal{Q}$ satisfying

$$q_\theta(\Gamma) = \pi(\Gamma). \quad (17)$$

Updating $q_{\theta_{k+1}}$ with this update has been proposing in the AMIS algorithm of Cornuet et al. (2012) and in the DM-PMC algorithm of Cappé et al. (2008) and is recovered in Algorithm 1 when $\tau_{k+1} = 1$, $\alpha = 1$, and $r \equiv 0$. Hence, Algorithm 1 introduces several additional degrees of freedom to this strict moment-matching approach. Relaxed moment-matching with $\tau_k > 0$ can be found in the covariance learning adaptive importance sampling algorithm of El-Laham et al. (2019). However, all the aforementioned works consider KL-based updates, that is with $\alpha = 1$ and no regularization term (i.e., $r \equiv 0$). The algorithm of Daudel et al. (2023) recovers relaxed moment-matching similar to ours, with $\alpha \neq 1$ but still with $r \equiv 0$.

When $\alpha = 1$, the weights of our 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 Koblents and Míguez (2013), 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 (Ionides, 2008; Vehtari et al., 2015; Korba and Portier, 2022). One can also see the review of Martino et al. (2018). 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 Korba and Portier 2022, Lemma 1). This is confirmed by our numerical experiments in Section 5. Note that, in our setting, this transformation comes naturally from the fact that we minimize a Rényi divergence.

In a different context, moment-matching updates have been used by Grosse et al. (2013) 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 (Neal, 2001; Moral et al., 2006), 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 Bui (2020), or to the *q-paths* of Masrani et al. (2021). Actually, moment-matching and geometric averages both are barycenters between π and q_θ in the sense of the inclusive or exclusive KL divergence (Grosse et al., 2013), indicating updates showcased in Proposition 14 may have a similar interpretation.

4 Convergence analysis

In this section, we analyze the convergence of Algorithms 1 and 2. We first explain in Section 4.1 in which sense the Bregman geometry induced by the KL divergence is well-adapted to handle Problem $(P_\pi^{(\alpha)})$ while the Euclidean geometry may fail. Convergence

results are given in Section 4.2 and are compared with existing results in Section 4.3. The proofs can be found in Appendices A-B.

Before stating our results, we give the assumptions that we use to study the properties of Problem $(P_\pi^{(\alpha)})$ and the convergence of Algorithms 1 and 2.

Assumption 15 *The exponential family \mathcal{Q} and the target π are such that*

- (i) $\text{int } \Theta \neq \emptyset$ and $\text{int } \Theta \subset \text{dom } f_\pi^{(\alpha)}$,
- (ii) \mathcal{Q} is minimal and steep, using the definitions of (Barndorff-Nielsen, 2014, 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 (Barndorff-Nielsen, 2014, Theorem 8.2). Note that when $\alpha \in (0, 1)$, then $\text{dom } f_\pi^{(\alpha)} = \Theta$ so that Assumption 15 (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 $\text{dom } f_\pi^{(\alpha)} = \Theta$, and Assumption 15 (i) holds if $\int \log(\pi(x))\pi(x)\nu(dx)$ and $\pi(\Gamma)$ are finite. However, Assumption 15 (i) may not be satisfied when $\alpha > 1$.

Assumption 16 *For any $\theta \in \text{int dom } A$, $\pi_\theta^{(\alpha)}(\Gamma) \in \text{int dom } A^*$. Equivalently, there exists $\theta^{(\alpha)} \in \text{int } \Theta$ such that $\pi_\theta^{(\alpha)}(\Gamma) = q_{\theta^{(\alpha)}}(\Gamma)$.*

In the case where $\alpha = 1$ and $\mathcal{Q} = \mathcal{G}$, Assumption 16 is equivalent to the target π having finite first and second order moments.

Assumption 17 *The regularizer r is in $\Gamma_0(\mathcal{H})$, is bounded from below, and is such that $\text{int } \Theta \cap \text{dom } r \neq \emptyset$.*

This assumption is standard in the Bregman optimization literature (Bauschke et al., 2003), and allows in particular non-smooth regularizers. For instance, Assumption 17 is satisfied by the ℓ_1 norm often used to enforce sparsity (Hastie et al., 2009, Section 3.4), or by indicator functions of non-empty closed convex sets as in Example 3.

Definition 18 *Under Assumption 17, we introduce for $\alpha > 0$ the set of stationary points of $F_\pi^{(\alpha)}$ as $S_\pi^{(\alpha)} := \{\theta \in \text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}, 0 \in \nabla f_\pi^{(\alpha)}(\theta) + \partial r(\theta)\}$.*

4.1 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 19 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

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

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

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

These properties give indications about the relation between f and its *tangent approximation* at θ' , defined by $\theta \mapsto f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + L d_B(\theta, \theta')$, 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 (Bauschke et al., 2017; Hanzely and Richtárik, 2021). 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 20 Let Assumption 15 be satisfied. The function $f_\pi^{(\alpha)}$, defined in (4), 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 20, 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 21 Let $\alpha > 0$.

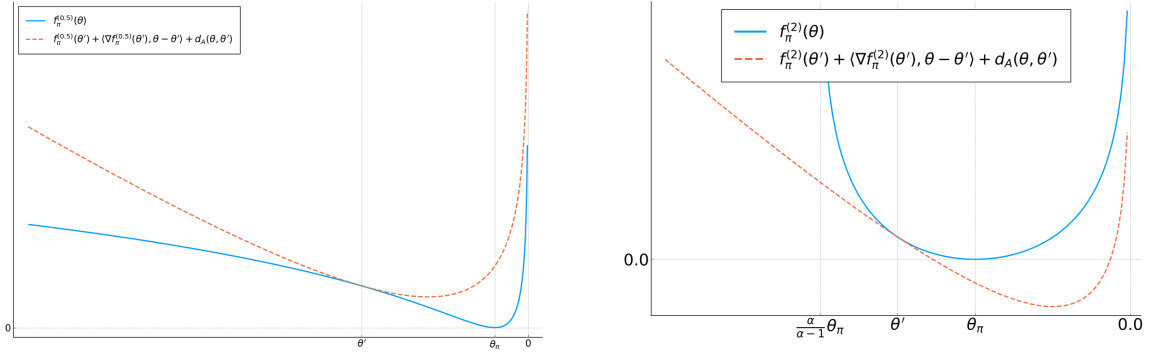
- (i) Under Assumptions 15 and 17, the objective function $F_\pi^{(\alpha)}$ is proper (i.e., with non-empty domain), lower semicontinuous, and bounded from below, that is

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

- (ii) If $\alpha \geq 1$ and Assumptions 15, 16, and 17 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 $\text{int } \Theta$.

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 19, 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 20 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 22 *There exist targets π and exponential families \mathcal{Q} such that the gradient of $f_\pi^{(\alpha)}$ is not Lipschitz on $\text{dom } f_\pi^{(\alpha)}$, for $\alpha > 0$.*

Remark 23 *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*

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

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

The counter-example used in the proof of Proposition 22 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 20, 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 15 (i) fails for $\alpha > 1$ since $\text{dom } f_\pi^{(\alpha)}$ is strictly included in Θ . One could also restrict the search to a smaller set, but the upper bound of $\text{dom } f_\pi^{(\alpha)}$ 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 $\theta_\pi \in \text{int } \Theta$ and lead to further results.

Proposition 24 *Suppose that Assumption 15 is verified and that there exists $\theta_\pi \in \text{int } \Theta$ with $\pi = q_{\theta_\pi}$. Then Assumption 16 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 24 shows that when $r \equiv 0$ and $\pi = q_{\theta_\pi}$ with $\theta_\pi \in \text{int } \Theta$, 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 25 *Suppose that Assumption 15 is satisfied and that $\pi = q_{\theta_\pi}$ for $\theta_\pi \in \text{int } \Theta$. Consider $\theta \in \text{int } \Theta$ in a ball of the form $B(\theta_\pi, v) \subset \text{int dom } \Theta$ for some $v > 0$. 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.,

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

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

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

Proposition 25 (i) shows that, in the neighborhood of its minimizer, function $f_\pi^{(\alpha)}$ has a quadratic behavior. This behavior is mentioned in (Amari and Nagaoka, 2000, Equation (3.11)) and in (van Erven and Harremoës, 2014, Equation (50)) but without precise statements on the needed regularity. We show here that this type of result holds under very mild assumptions. The consequence of this behavior is 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 (Karimi et al., 2016).

4.2 Convergence analysis of Algorithms 1 and 2

We now present our convergence results for Algorithms 1 and 2. We start with Algorithm 1, which is deterministic, by first giving results for values of α in $(0, 1]$, and then stronger results when $\alpha = 1$. We then go to Algorithm 2, which is based on sampling and thus stochastic. We first give our assumption on the noise induced by the sampling procedure and proceed to first give results for $\alpha \in (0, 1]$, followed by more precise results for $\alpha = 1$. Note that 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 first show how Assumptions 15, 16, and 17 ensure the well-posedness of the operators introduced in Definition 11 and used to construct our algorithms.

Proposition 26

- (i) Under Assumptions 15 and 16, if $\tau \in (0, 1]$, the operator $\gamma_{\tau f_\pi^{(\alpha)}}^A$ has domain $\text{int } \Theta$ where it is single-valued, and $\gamma_{\tau f_\pi^{(\alpha)}}^A(\theta) \in \text{int } \Theta$ for every $\theta \in \text{int } \Theta$.
- (ii) Under Assumptions 15 and 17, the domain of $\text{prox}_{\tau r}^A$ is $\text{int } \Theta$. On $\text{int } \Theta$, $\text{prox}_{\tau r}^A$ is single-valued, and $\text{prox}_{\tau r}^A(\theta) \in \text{int } \Theta$ for every $\theta \in \text{int } \Theta$.
- (iii) If Assumptions 15, 16, and 17 are satisfied, and $\tau \in (0, 1]$, $T_{\tau F_\pi^{(\alpha)}}^A = \text{prox}_{\tau r}^A \circ \gamma_{\tau f_\pi^{(\alpha)}}^A$, and a point $\theta \in \text{int } \Theta$ is a fixed point of $T_{\tau F_\pi^{(\alpha)}}^A$ if and only if $\theta \in S_\pi^{(\alpha)}$.

Proposition 26 implies in particular that under Assumptions 15, 16, and 17, the iterates of Algorithm 1 are well-defined and stay in $\text{int } \Theta$ for step-sizes $\{\tau_k\}_{k \in \mathbb{N}}$ belonging to $(0, 1]$. Proposition 26 also establishes that the fixed points of Algorithm 1 are the stationary points of Problem $(P_\pi^{(\alpha)})$.

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

Proposition 27 Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 1 from $\theta_0 \in \text{int } \Theta$, 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 15, 16, and 17, we have the following results.

- (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 $K \in \mathbb{N}$, then $\theta_k = \theta_K$ for all $k \geq K$ and $\theta_K \in S_\pi^{(\alpha)}$.

- (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 $\tau_k \in [\varepsilon, 1]$ for some $\varepsilon > 0$ and there exists a non-empty compact set $C \subset \text{int } \Theta$ containing the iterates, there exists a sequence $\{\rho_k\}_{k \in \mathbb{N}}$ such that $\rho_k \in \nabla f_\pi^{(\alpha)}(\theta_k) + \partial r(\theta_k)$ for every $k \in \mathbb{N}$ and $\rho_k \xrightarrow[k \rightarrow +\infty]{} 0$. If, additionally, r is continuous on C , then the limit points of $\{\theta_k\}_{k \in \mathbb{N}}$ are in $S_\pi^{(\alpha)}$.

The additional assumption used for Proposition 27 (v) is satisfied for instance if $r = \iota_C$, for a compact $C \subset \text{int } \Theta$. 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 $\theta_s \in \text{int } \Theta$, Prop. 27(v) implies that $\theta_k \xrightarrow[k \rightarrow +\infty]{} \theta_s$.

We now refine the result of Proposition 27 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 28 *Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 1 from $\theta_0 \in \text{int } \Theta$, with $\alpha = 1$ and step-sizes $\{\tau_k\}_{k \in \mathbb{N}}$ in $(0, 1]$. Consider the minimizer θ_* established in Proposition 21. Under Assumptions 15, 16, and 17,*

(i) *if $\sum_k \tau_k = +\infty$, the iterates converge to the solution, $\theta_k \xrightarrow[k \rightarrow +\infty]{} \theta_*$,*

(ii) *if $\tau_k \in [\varepsilon, 1]$ for some $\varepsilon > 0$, we have*

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

(iii) *if $\tau_k \in [\varepsilon, 1]$ for some $\varepsilon > 0$, we have*

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

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 $\theta_\pi \in \text{int } \Theta$. In this case, we are able to derive convergence results that are similar to the case $\alpha = 1$.

Proposition 29 *Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 1 from $\theta_0 \in \text{int } \Theta$, with $\alpha \in (0, 1)$ and a sequence of step-sizes $\{\tau_k\}_{k \in \mathbb{N}}$ in $[\varepsilon, 1]$ for some $\varepsilon > 0$. Assume that the iterates stay in a non-empty compact set $C \subset \text{int } \Theta$, that $r \equiv 0$, that there exists $\theta_\pi \in \text{int } \Theta$ such that $\pi = q_{\theta_\pi}$, and that Assumption 15 is satisfied. Then,*

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

(ii) then $RD_\alpha(\pi, q_{\theta_k}) \xrightarrow[k \rightarrow +\infty]{} 0$ and there exist constants $M > 0$ and $\delta \in (0, 1)$ such that

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

We now turn to the study of Algorithm 2, which is a sampling-based counterpart to Algorithm 1. We first present our assumption on the mean square error introduced by our sampling procedure.

Assumption 30 *There exists a function $E_{\pi, \mathcal{Q}}^{(\alpha)} : \text{int } \Theta \rightarrow \mathbb{R}$ that is locally bounded and such that for any $\theta \in \text{int } \Theta$ and any $N \in \mathbb{N} \setminus \{0\}$,*

$$\mathbb{E} \left[\left\| \pi_\theta^{(\alpha)}(\Gamma) - \sum_{l=1}^N \bar{w}_l^{(\alpha)} \Gamma(x_l) \right\|^2 \right] \leq \frac{1}{N} E_{\pi, \mathcal{Q}}^{(\alpha)}(\theta).$$

Such type of control on the mean square error of sampling-based estimators is reminiscent of the importance sampling bounds given by Agapiou et al. (2015). It is possible to show that Assumption 30 is satisfied from more elementary assumptions on relevant moments using the analysis from Doukhan and Lang (2009).

Proposition 31 *Suppose that Assumptions 15, 16, 17, and 30 hold and take $\alpha \in (0, 1]$. Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 2 from $\theta_0 \in \text{int } \Theta$ with step-sizes $\{\tau_k\}_{k \in \mathbb{N}}$ in $(0, 1]$ and sample sizes satisfying $\sum_{k \geq 0} \frac{1}{\sqrt{N_{k+1}}} < +\infty$. If there exists a non-empty compact set $C \subset \text{int } \Theta$ containing the iterates with probability one, we have*

- (i) $\mathbb{E} \left[\sum_{k \geq 0} KL(q_{\theta_{k+1}}, q_{\theta_k}) \right] < +\infty$ and $KL(q_{\theta_{k+1}}, q_{\theta_k}) \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} 0$,
- (ii) if $\tau_k \in [\epsilon, 1]$ for some $\epsilon > 0$, then there exists a sequence $\{\rho_k\}_{k \in \mathbb{N}}$ such that $\rho_k \in \nabla f_\pi^{(\alpha)}(\theta_k) + \partial r(\theta_k)$ for every $k \in \mathbb{N}$ and $\rho_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} 0$.

We now give a second convergence result for Algorithm 2 in the particular case of $\alpha = 1$. This result gives convergence rates in terms of expectation of the KL divergence between the iterates and the optimal proposal. This result involves a intricate interplay between the step sizes and the sample sizes.

Proposition 32 *Suppose that Assumptions 15, 16, 17, and 30 hold and take $\alpha = 1$. Consider the minimizer θ_* established in Proposition 21. Consider a sequence $\{\theta_k\}_{k \in \mathbb{N}}$ generated by Algorithm 2 from $\theta_0 \in \text{int } \Theta$ with step-sizes $\{\tau_k\}_{k \in \mathbb{N}}$ in $(0, 1]$ and sample sizes $\{N_k\}_{k \in \mathbb{N}}$. Suppose that there exists a non-empty compact set $C \subset \text{int } \Theta$ containing the iterates with probability one, then we have the following results.*

- (i) *There exists a constant $M > 0$ such that*

$$\mathbb{E} [KL(q_{\theta_k}, q_{\theta_*})] \leq \left(\prod_{l=0}^k (1 - \tau_{l+1}) \right) KL(q_{\theta_0}, q_{\theta_*}) + \left(\sum_{l=0}^k \frac{\tau_{l+1}}{\sqrt{N_{l+1}}} \prod_{m=l+1}^k (1 - \tau_{m+1}) \right) M.$$

(ii) If we have $\sum_{k \geq 0} \tau_k = +\infty$ and $\sum_{k \geq 0} \frac{1}{\sqrt{N_k}} < +\infty$, then $\theta_k \xrightarrow[k \rightarrow +\infty]{a.s.} \theta_*$.

Remark 33 If $\tau_k \equiv \tau \in (0, 1]$ and $N_k \equiv N \in \mathbb{N} \setminus \{0\}$, then Proposition 32 gives

$$\mathbb{E} [KL(q_{\theta_k}, q_{\theta_*})] \leq (1 - \tau)^{k+1} \left(KL(q_{\theta_0}, q_{\theta_*}) - \frac{M}{\sqrt{N}} \right) + \frac{M}{\sqrt{N}}, \quad (18)$$

meaning that the iterates will get to a neighborhood of the solution whose size is asymptotically controlled by $\frac{M}{\sqrt{N}}$.

4.3 Discussion

We first discuss our assumptions, and why they are weaker compared to existing analyses of similar schemes. Generally, we avoided 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 $\text{dom } A$ is not closed, which contrasts with common assumptions from the literature on optimization schemes based on Bregman divergences (Bauschke et al., 2017; Teboulle, 2018; Bolte et al., 2018; Gao et al., 2020; Hanzely and Richtárik, 2021) or in the statistical literature (Akyildiz and Míguez, 2021; Khan et al., 2016; Li and Turner, 2016; Bungert et al., 2022). Note that the Euclidean smoothness of $KL(q, \pi)$ is proven for instance by Lambert et al. (2022); Kim et al. (2023); Domke et al. (2023) under a log-smoothness assumption on the target for Gaussian or location-scale approximating families. Standard VI works minimizing the exclusive KL divergence benefit from unbiased estimators of the gradients (Kim et al., 2023; Domke et al., 2023). In contrast, Algorithm 2 only yields biased gradient estimates, adding another challenge to the analysis. Although this setting has been studied in the Euclidean case (Tadić and Doucet, 2017; Atchadé et al., 2017; Akyildiz and Míguez, 2021; Dieuleveut et al., 2023), we are only aware of the work of Gruffaz et al. (2024) in a non-Euclidean setting. Namely, Gruffaz et al. (2024) covers expectation-maximization algorithm where the noise comes from using a MCMC algorithm.

Proposition 27 implies a monotonic decrease of $F_{\pi}^{(\alpha)}$ along iterations of Algorithm 1. This kind of result appears in many statistical procedures (Douc et al., 2007; Daudel et al., 2023). 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 27 (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 27 (iv) provides estimates on the number of iterations needed to reach a certain level of stationarity between iterates, while Proposition 27 (v) establishes convergence to the set of stationary points.

We are also able to show the geometric rate of convergence of iterates of Algorithm 1 to the global minimizer of Problem $(P_{\pi}^{(\alpha)})$ when $\alpha = 1$ in Proposition 28 and when $\pi \in \mathcal{Q}$ and $r \equiv 0$ for $\alpha < 1$ in Proposition 29. 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 16). In

comparison, similar rates of convergence are established in the case of the objective function $KL(\cdot, \pi)$ in (Lambert et al., 2022; Yao and Yang, 2022) 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 the geometric convergence of the probability distribution of the samples to the minimizer of $RD_\alpha(\cdot, \pi)$ for $\alpha \geq 1$ that is proven by Vempala and Wibosono (2019) for MCMC under log-smoothness assumption on the target and a weaker version of log-concavity. It is difficult to compare the assumption $\pi \in \mathcal{Q}$ used in Proposition 29 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} .

In the case of Algorithm 2 for $\alpha \in (0, 1]$, Proposition 31 extends the finite-length property of the sequence of iterates to this stochastic setting and establishes the convergence to zero of a sequence of gradient or subgradients. When $\alpha = 1$, we give in Proposition 32 an explicit rate of convergence to the minimizer in terms of the step sizes and sample sizes and exhibit a case where the iterates almost surely converge to the minimizer. When $\alpha \in (0, 1]$, the problem is non-convex with biased stochastic gradients, which is a very challenging setting. Since our analysis mostly leverages standard tools from the study of Bregman proximal gradient algorithms, it can probably be extended to more general settings. In the case $\alpha = 1$, the problem becomes relatively strongly convex, without any log-concavity assumption on the target. Proposition 32 (i) is a KL analogue to the Wasserstein-based (Lambert et al., 2022, Theorem 4), although the latter work keeps the sample size per iteration constant. Proposition 31 (ii) generalizes (Marin et al., 2019, Theorem 3.2 (ii)), which establishes the convergence of a simplified version of the AMIS algorithm of Cornuet et al. (2012), to case where we allow step sizes lower than one and where self-normalized importance sampling is used instead of unnormalized importance sampling. Note that we need to assume that the iterates stay bounded and away from the boundary of the domain of A . Boundedness of stochastic trajectory is often assumed, for instance in the framework of the ODE method (Benaïm, 1999), while the behavior of the iterates of Bregman proximal gradient methods near the boundary of the domain of A (when it is not closed) is usually not addressed (Bauschke et al., 2017; Teboulle, 2018).

Our convergence analysis is restricted to $\alpha \in (0, 1]$, which is also the case in the analysis of Daudel et al. (2023), 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 19 that

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

This is to be compared with (Daudel et al., 2023, Proposition 1), which, in our setting, would read

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

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 Daudel et al. (2023) (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

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

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 (Daudel et al., 2023, 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 and possible bias. Indeed, the convergence result in (Daudel et al., 2023) only shows a monotonic decrease of the objective without regularization in the deterministic case, although in wider variational families.

During the final stage of the review process, we became aware of the work of Bertholom et al. (2024), where the convergence of algorithms minimizing the α -divergence D_α , $\alpha \in (0, 1)$, over exponential families is analyzed. In their work, a deterministic algorithm that is similar to our Algorithm 1 with $r \equiv 0$, is studied but an alternating sampling-based implementation is proposed. More precisely, the relaxed moment-matching algorithm is slightly modified to obtain an algorithm which can be implemented with unbiased estimators, while we stick with the original moment-matching scheme and provide convergence analysis in this case.

5 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. Note that Algorithm 1 is an idealized algorithm and cannot be implemented in general. We also consider the VRB algorithm from Li and Turner (2016), whose implementation for an exponential family is described by Equation (16). It is shown in Section 3.3 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 22) while it is smooth relatively to the Bregman divergence d_A (see Proposition 20). 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 Li and Turner (2016).

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 regression problem where we approximate the posterior distribution of a regression parameter $\beta \in \mathbb{R}^d$. We observe J measurements $y_T^{(j)} \in \mathbb{R}^{d_y}, y_0^{(j)} \in \mathbb{R}^{d_y}$, where $T > 0$ and $j \in \llbracket 1, J \rrbracket$. We set $d_y = 2$ and $d = 6$. We introduce the flow $\Phi_\beta^T : \mathbb{R}^{d_y} \rightarrow \mathbb{R}^{d_y}$ such that, for any $\xi_0 \in \mathbb{R}^{d_y}$, $\Phi_\beta^T(\xi_0) = \xi(T)$ where $\xi(\cdot)$ is the solution of

$$\begin{cases} \dot{\xi}_1 &= \beta_1 \xi_1 + \beta_3 \xi_1^2 + \beta_5 \xi_1 \xi_2, \\ \dot{\xi}_2 &= \beta_2 \xi_2 + \beta_4 \xi_2^2 + \beta_6 \xi_1 \xi_2, \\ \xi(0) &= \xi_0. \end{cases}$$

Then, for every $j \in \llbracket 1, J \rrbracket$, we have $y_T^{(j)} = \Phi_\beta^T(y_0^{(j)}) + n^{(j)}$ where $n^{(j)} \sim \mathcal{N}(0, \sigma^2)$. We thus have that

$$p(y|\beta) = \prod_{j=1}^J \mathcal{N}\left(y_T^{(j)}; \Phi_\beta^T(y_0^{(j)}), \sigma^2\right).$$

This kind of task arises for instance in ecology (Knappe and de Valpine, 2012).

Note that auto-differentiation cannot be used straightforwardly to compute the gradients with respect to β of the above likelihood, due to the need to integrate the dynamics over $[0, T]$. The gradients can still be computed, using for instance the adjoint method (Chen et al., 2018), but this requires special care.

We generate synthetic data from a sparse vector $\bar{\beta}$ being sampled such that

$$\bar{\beta} \sim \prod_{i=1}^d (\rho \delta_0(d\beta_i) + (1 - \rho) \mathcal{L}(\beta_i; 0, \lambda_1) d\beta_i).$$

The above distribution is a spike-and-slab prior, using a Dirac mass at zero (the spike) and a Laplace distribution (the slab). This prior is hard to deal with, in general, as it requires model-specific derivations (Ray and Szabó, 2022). Furthermore, the latter methods minimize the exclusive KL divergence and we are not aware of any work that tries to approximate a spike-and-slab posterior by minimizing a Rényi divergence. We will thus assume the so-called LASSO spike-and-slab prior (Bai et al., 2021), which has a continuous density defined as

$$p_0(\beta) = \prod_{i=1}^d (\rho \mathcal{L}(\beta_i; 0, \lambda_0) + (1 - \rho) \mathcal{L}(\beta_i; 0, \lambda_1)),$$

with $\lambda_0 \ll \lambda_1$. Now, the Dirac mass is replaced by a very spiky Laplace distribution as well.

We compare the RMM, the PRMM, and the VRB algorithm with target $\pi(\beta) \propto p(y|\beta)p_0(\beta)$. All the algorithms are run considering an approximating exponential family \mathcal{Q} , chosen as the family of Gaussian distributions with diagonal covariance matrices. The parametrization of this family is detailed in Appendix C. For the PRMM algorithm, we use the regularizer $r(\theta) = \eta \|\theta_1\|_1$ with $\eta \geq 0$. This can be understood as the Lagrangian relaxation (Hiriart-Urruty and Lemaréchal, 1993) with multiplier $\eta \geq 0$ of the constraint $\sum_{i=1}^d \|\theta_1\|_1 \leq c$, for some $c \geq 0$ 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 idea is to mimic the sparse structure of $\bar{\beta}$ that was simulated from p_0 . The computation

of the corresponding Bregman proximal operator for this choice of regularizer is detailed in Appendix C.

In order to assess the performance of the algorithms, we track the variational Rényi bound, defined Eq. (14), that is estimated at each iteration $k \in \mathbb{N}$ through

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

This quantity can be understood as an approximation of the Rényi divergence that we are trying to minimize (Li and Turner, 2016), and thus as a measure of the proximity between π and its approximation q_{θ_k} . We also consider the F1 score that each algorithm achieves in the prediction of the zeros of the true regression vector $\bar{\beta}$. It is computed at each iteration $k \in \mathbb{N}$, by seeing how the zeros of μ_k match those of $\bar{\beta}$. Additionally, we sample regression vectors β from the final proposal q_{θ_K} , and average their absolute error over a test data set $\{(z_0^{(j)}, z_T^{(j)})\}_{j=1}^{J^{\text{test}}}$. This is done by computing the following mean absolute error (MAE):

$$\text{MAE}^{\text{test}}(\theta) = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \sum_{j=1}^{J^{\text{test}}} \left| z_T^{(j)} - \Phi_{\beta^{(i)}}^T(z_0^{(j)}) \right|, \text{ with } \beta^{(i)} \sim q_\theta, \forall i \in \llbracket 1, N^{\text{test}} \rrbracket. \quad (23)$$

We compute this quantity for the final proposal q_{θ_K} in each run and analyze the distributions of the obtained values, in order to assess the robustness of the methods.

We run the algorithms for $K = 100$ iterations, with $N = 500$ samples per iteration. The PRMM and RMM algorithms have been implemented with constant step size $\tau = 10^{-1}$, while the VRB algorithm uses a constant step size $\tau = 10^{-4}$. These choices correspond to the most favorable step-size rule, for each algorithm, as indicated by our extended experiments in Appendix D. For the PRMM algorithm, we use the regularization parameter $\eta = 0.5$. We have set up the experiment, with $J = 100$, $J^{\text{test}} = 50$, $\rho = 0.3$, $\lambda_0 = 1$, $\lambda_1 = 20$, and $\sigma^2 = 0.5$. We perform $N^{\text{test}} = 50$ tests. We discarded one run of the VRB algorithm for $\alpha = 1$ for which iterates had become singular. We now present and discuss the result of our experiments.

Figure 2 shows the increase of the approximated variational Rényi bound described in Eq. (22). As discussed in Section 3.3, 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 methods decrease the Rényi divergence. However, our methods are able to reach higher values than the VRB method at a faster rate, illustrating the improvement coming by 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, showing using only a spike-and-slab LASSO prior is not enough to get approximations with sparse means. The PRMM algorithm uses a proximity operator to recover the zeros of the sought regression vector. This additional operator leads to a very good recovery of the zeros of the regression vector. Note that the regularizing term is multiplied by a scalar value $\eta > 0$ that controls the strength of the regularization relative to the Rényi divergence term. Setting this value to achieve the best recovery while preserving good

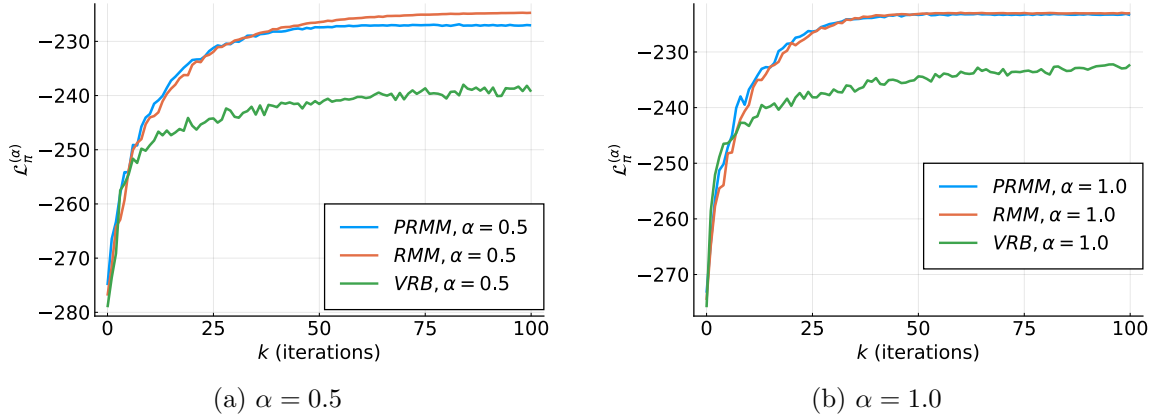


Figure 2: Approximated Rényi bound, averaged over 100 runs with $N = 500$ samples per iteration.

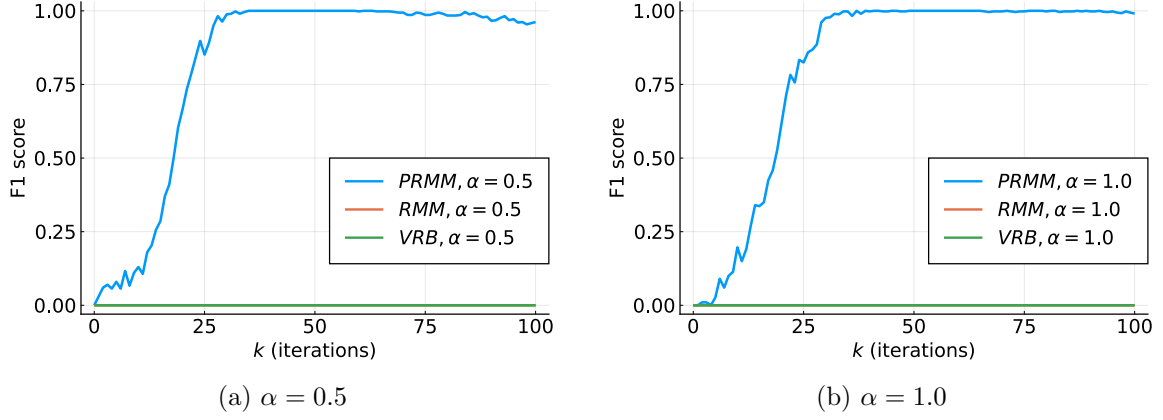


Figure 3: F1 score in the prediction of the zeros of $\bar{\beta}$ by the zeros of $\{\mu_k\}_{k=0}^K$, averaged over 100 runs with $N = 500$ samples per iteration.

approximating properties is difficult a priori, as it is the case for instance in maximum a posteriori estimation.

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 yield reconstruction errors that are less spread and at a lower level than the ones coming from the VRB algorithm. Note also that the VRB algorithm also produces badly-performing outlier values. This shows the higher approximation performance and robustness coming from using a more adapted geometry, as in our proposed method. Errors are more spread for the PRMM algorithm than for the RMM algorithm. This may be due to the sparsity-inducing proximal step, which creates larger eigenvalues for the covariance matrix (see Appendix C for details).

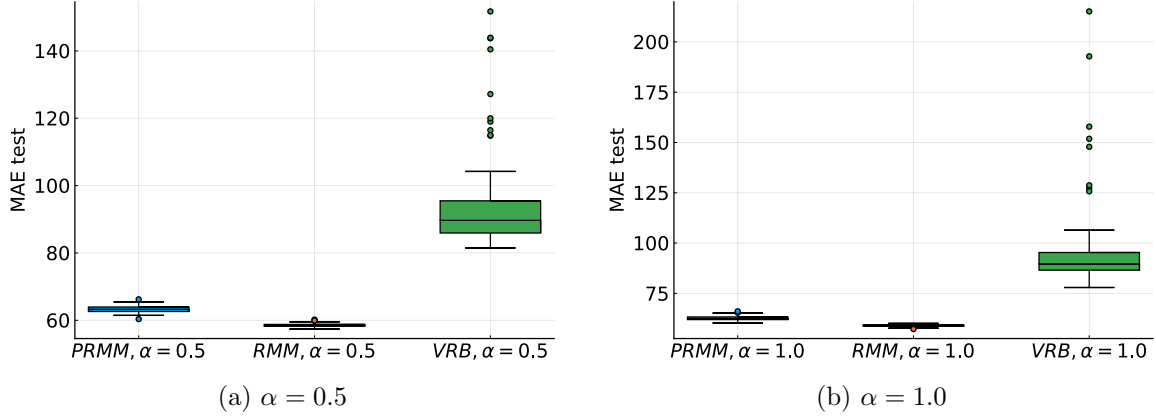


Figure 4: Box plots of the values $\text{MAE}^{\text{test}}(\theta_K)$ obtained over 100 runs with $N = 100$ samples per iteration and $K = 100$ iterations, showing the reconstruction error on the test data.

6 Conclusion and perspectives

We introduced in this work a Bregman proximal gradient algorithm to solve the variational inference problem of minimizing the sum of a Rényi divergence and a regularizing function over an exponential family. We used a Bregman divergence, equivalent to the Kullback-Leibler divergence, linking our algorithm with natural gradient methods. We also showed that our algorithm generalizes several moment-matching algorithms. We provided a black-box implementation for non-conjugate targets along with explicit computations for proximal steps enforcing sparsity of the solutions.

Using this novel Bregman-based perspective, we established strong convergence guarantees for our exact and sampling-based algorithms. For $\alpha \in (0, 1]$ in the deterministic case, 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. When $\alpha = 1$ in the deterministic case, we also established the geometric convergence of the iterates towards the optimal parameters. In the sampling-based case, we established a finite-length property for the sequence of iterates and convergence of subgradients to zero when $\alpha \in (0, 1]$. We provided convergence rates of the iterates to the minimizer in expectation, with rates dependent on the step and sample sizes. We also gave conditions on the step sizes and sample sizes to yield almost sure convergence of the iterates to the minimizer. 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.

We are not aware of any work studying Bregman proximal gradient algorithms with biased estimates of the gradients. As our analysis of our sampling-based algorithm mostly

relies on standard techniques used to study Bregman-based algorithms, it should be possible to extend our results to more general settings. Our analysis also faces open problems in the analysis of Bregman-based algorithms, which are worthy of further investigation. 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 analysis of importance sampling schemes.

Acknowledgments and Disclosure of Funding

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.

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

A.1 Proof of Proposition 13

For the sake of clarity, the expression of $\nabla^2 f_\pi^{(\alpha)}$ was not given in Proposition 13. However, this expression is useful for some of the following proofs. We thus show here that for any $\theta \in \text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$, we have

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

Proof [Proof of Proposition 13] For the case $\alpha = 1$, note that $f_\pi^{(1)}$ can be written as

$$f_\pi^{(1)}(\theta) = \int \log(\pi(x)) \pi(x) \nu(dx) - \langle \theta, \pi(\Gamma) \rangle + A(\theta), \quad \forall \theta \in \Theta \cap \text{dom } f_\pi^{(\alpha)}, \quad (24)$$

where $\Theta = \text{dom } A$, and A defined in Eq. (2). The results come from the properties of A , given in Proposition 9.

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 $\theta \in \text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$ and $x \in \mathcal{X}$.

We can show through standard results that at any $\theta \in \text{int } \Theta$, 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 $\theta \in \Theta \cap \text{dom } f_\pi^{(\alpha)}$, and $f_\pi^{(\alpha)} = A + \frac{1}{\alpha-1} \log \circ \tilde{h}$, these results with those of Proposition 9 about A give the following. On $\text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$, 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 $\theta \in \text{int } \Theta \cap \text{dom } f_\pi^{(\alpha)}$. For $i \in \llbracket 1, n \rrbracket$, 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

$$\frac{\partial f_\pi^{(\alpha)}}{\partial \theta_i}(\theta) = \frac{\partial A}{\partial \theta_i}(\theta) - \frac{\int \Gamma_i(x) \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)}{\int \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)}. \quad (25)$$

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

Before computing the second order partial derivatives, we introduce another intermediate quantity. Denote $\tilde{g}_i : \theta \mapsto \int \Gamma_i(x) \pi(x)^\alpha \exp(\langle \theta, \Gamma(x) \rangle)^{1-\alpha} \nu(dx)$ for $i \in \llbracket 1, n \rrbracket$. In fact, $\tilde{g}_i(\theta) = \frac{1}{1-\alpha} \frac{\partial \tilde{h}}{\partial \theta_i}(\theta)$, and from Eq. (25), 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 $j \in \llbracket 1, n \rrbracket$

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

Using those intermediate results, we obtain for $i, j \in \llbracket 1, n \rrbracket$ that

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

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

A.2 Proof of Proposition 14

Proof [Proof of Proposition 14] From the definition of the operator $\gamma_{\tau_{k+1} f_\pi^{(\alpha)}}^A$, we write the optimality conditions that $\theta_{k+\frac{1}{2}}$ must satisfied (we assume that it is uniquely defined and in $\text{int dom } \Theta$):

$$0 = \nabla f_\pi^{(\alpha)}(\theta_k) + \frac{1}{\tau_{k+1}} (\nabla A(\theta_{k+\frac{1}{2}}) - \nabla A(\theta_k))$$

Re-arranging the terms, we obtain that

$$\nabla A(\theta_{k+\frac{1}{2}}) = \nabla A(\theta_k) - \tau_{k+1} \nabla f_\pi^{(\alpha)}(\theta_k).$$

Using the characterization of ∇A from Proposition 9 and the expression of $\nabla f_\pi^{(\alpha)}$ from Proposition 13, we obtain that the above is equivalent to $q_{\theta_{k+\frac{1}{2}}}(\Gamma) = \tau_{k+1} \pi_{\theta_k}^{(\alpha)}(\Gamma) + (1 - \tau_{k+1}) q_{\theta_k}(\Gamma)$, showing the result. \blacksquare

A.3 Proof of Proposition 20

Proof [Proof of Proposition 20] We prove relative smoothness and relative strong convexity by using the alternative characterizations given in (Hanzely and Richtárik, 2021, Proposition 2.2) and (Hanzely and Richtárik, 2021, Proposition 2.3). $f_\pi^{(\alpha)}$ and A are twice differentiable on $\text{int } \Theta$, so thanks to these results, $f_\pi^{(\alpha)}$ is L -relatively smooth with respect to A if and only if $\nabla^2 f_\pi^{(\alpha)} \preceq L \nabla^2 A$, on $\text{int } \Theta$, and it is ρ -relatively strongly convex with respect to A if and only if $\rho \nabla^2 A \preceq \nabla^2 f_\pi^{(\alpha)}$ on $\text{int } \Theta$.

We first cover the case $\alpha = 1$. In this case, we have that for every $\theta \in \text{int } \Theta$, $\nabla^2 f_\pi^{(1)}(\theta) = \nabla^2 A(\theta)$ from Proposition 13 (see Section A.1). Therefore, the functions $f_\pi^{(1)} - A$ and $A - f_\pi^{(1)}$ have null Hessian on $\text{int } \Theta$, showing that they are convex, hence the result.

Now, consider $\alpha \neq 1$, then, under Assumption 15, we recall from Proposition 13 (see Section A.1) that

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

Consider $\theta \in \text{int } \Theta$, 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 $\theta \in \text{int } \Theta$, $\nabla^2 (f_\pi^{(\alpha)} - A)(\theta)$ 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 (Hanzely and Richtárik, 2021, Proposition 2.2) and (Hanzely and Richtárik, 2021, Proposition 2.3). \blacksquare

A.4 Proof of Proposition 21

Proof [Proof of Proposition 21] Consider $\alpha > 0$.

(i) $F_\pi^{(\alpha)}$ is proper because $f_\pi^{(\alpha)}$ is non-negative from Proposition 3, takes finite values for some $\theta \in \Theta$ by Assumption 15, and because r is proper by Assumption 17. The fact that the infimum of $(F_\pi^{(\alpha)})$ 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 17.

We now prove the lower semicontinuity. When $\alpha = 1$, we recall from Eq. (24) that

$$f_\pi^{(1)}(\theta) = H(\pi) - \langle \theta, \pi(\Gamma) \rangle + A(\theta), \forall \theta \in \Theta, \quad (26)$$

where $H(\pi) = \int \log(\pi(x))\pi(x)\nu(dx)$. Because A is lower semicontinuous on Θ from Proposition 9, 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 (Carothers, 2000, 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. (26). Because of Assumption 16, $\pi(\Gamma) \in \text{int dom } A^*$. Thanks to (Bauschke and Borwein, 1997, Fact 2.11) and Proposition 9, this ensures that $f_\pi^{(1)}$ is coercive. Because of Assumption 15 which ensures the well-posedness of $f_\pi^{(\alpha)}$, we have from (van Erven and Harremoës, 2014, Theorem 3) that

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

This ensures that $f_\pi^{(\alpha)}$ is coercive for $\alpha > 1$. The regularizer r is bounded from below thanks to Assumption 17, 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 $\theta_* \in \text{dom } \Theta$ such that $F_\pi^{(\alpha)}(\theta_*) = \vartheta_\pi^{(\alpha)}$. We now use the optimality conditions that θ_* satisfies to show that $\theta_* \in \text{int } \Theta$. In particular, we have from (Bauschke and Combettes, 2011, Theorem 16.2) that

$$0 \in \partial F_\pi^{(\alpha)}(\theta_*). \quad (27)$$

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 (24), Assumption 17 and the convexity and properness of $\theta \mapsto -\langle \theta, \pi(\Gamma) \rangle$, A and r , and (Bauschke and Combettes, 2011, 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 8, so Eq. (27) implies that $0 \in \emptyset$. This shows that $\theta_* \in \text{int } \Theta$.

Finally, since A is strictly convex on $\text{int } \Theta$ (Proposition 9), so is $F_\pi^{(1)}$, so such θ_* is unique. ■

A.5 Proof of Proposition 22

Proof [Proof of Proposition 22] 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 13 (see Section A.1), $(f_\pi^{(\alpha)})''$ is independent of the choice of the target π , and is equal to

$$(f_\pi^{(\alpha)})''(\theta) = A''(\theta) = \frac{1}{2\theta^2}. \quad (28)$$

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 $\text{dom } f_\pi^{(\alpha)} = (0, \frac{\alpha}{\alpha-1}\theta_\pi)$ for $\alpha > 1$. In the following, we consider $\theta \in \text{dom } f_\pi^{(\alpha)}$. Then we compute using the result of Proposition 13 (see Section A.1) 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

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

and we note that $A(\theta) = \log(\sqrt{-\frac{\pi}{\theta}})$. 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

$$(f_\pi^{(\alpha)})''(\theta) = \frac{1}{2\theta^2} + \frac{\alpha - 1}{2((\alpha-1)\theta - \alpha\theta_\pi)^2}. \quad (29)$$

Equations (28) and (29) 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 $\alpha > 1$. \blacksquare

A.6 Proof of Proposition 24

Proof [Proof of Proposition 24]

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 9) and $\theta_\pi, \theta \in \text{int } \Theta$, $\alpha\theta_\pi + (1-\alpha)\theta \in \text{int } \Theta$. Therefore, $\pi_\theta^{(\alpha)}(\Gamma) = \nabla A(\alpha\theta_\pi + (1-\alpha)\theta) \in \text{int dom } A^*$, since A is Legendre and $\alpha\theta_\pi + (1-\alpha)\theta \in \text{int dom } A$, showing that Assumption 16 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., $\theta \in \text{int } \Theta$ such that $\nabla f_\pi^{(\alpha)}(\theta) = 0$. This implies that

$$q_\theta(\Gamma) = \pi_\theta^{(\alpha)}(\Gamma), \quad (30)$$

due to the characterization given in Proposition 13. Under our assumptions, Eq. (30) 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 if $\theta = \theta_\pi$, showing the existence and unicity of the stationary point of $f_\pi^{(\alpha)}$. ■

A.7 Proof of Proposition 25

Proof [Proof of Proposition 25]

Under our Assumptions on π , we can compute

$$f_\pi^{(\alpha)}(\theta) = \frac{1}{1-\alpha} (\alpha A(\theta_\pi) + (1-\alpha)A(\theta) - A(\alpha\theta_\pi + (1-\alpha)\theta)), \quad \forall \theta \in \text{int } \Theta \quad (31)$$

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

(i) We can compute

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

Using Eq. (31), these two equalities imply in particular that

$$f_\pi^{(\alpha)}(\theta) = \frac{\alpha}{2} \|\theta - \theta_\pi\|_{\nabla^2 A(\theta_\pi)}^2 + o(v^2). \quad (32)$$

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

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

which yields with Eq. (32) that

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

We thus have that

$$f_\pi^{(\alpha)}(\theta) = f_\pi^{(\alpha)}(\theta') + \langle \nabla f_\pi^{(\alpha)}(\theta'), \theta - \theta' \rangle + \frac{\alpha}{2} \|\theta - \theta'\|_{\nabla^2 A(\theta_\pi)}^2 + o(v^2). \quad (33)$$

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

$$\nabla f_\pi(\theta') + \alpha\nabla^2 A(\theta_\pi)(\theta - \theta') = 0. \quad (34)$$

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

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

Now take $\eta = \nabla A(\theta_\pi)$ which belongs to $\text{int dom } A^*$ since $\theta_\pi \in \text{int dom } A$. We obtain that $\nabla^2 A^*(\nabla A(\theta_\pi)) \nabla^2 A(\theta_\pi) = I_d$. This shows that the optimality condition of Eq. (34) 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. (33) can be minorized to obtain:

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

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

Appendix B. Convergence analysis of Algorithms 1 and 2

B.1 Proof of Proposition 26

Proof [Proof of Proposition 26]

(i) We start by considering the quantity

$$\nabla A(\theta) - \tau \nabla f_{\pi}^{(\alpha)}(\theta) = \tau \pi_{\theta}^{(\alpha)}(\Gamma) + (1 - \tau) \nabla A(\theta)$$

From Assumption 15 and Proposition 9, A is a Legendre function and thus benefits from the results of Proposition 8. In particular, $\nabla A(\theta) \in \text{int dom } A^*$. Similarly, $\pi_{\theta}^{(\alpha)}(\Gamma) \in \text{int dom } A^*$ from Assumption 16. Since $\tau \in (0, 1]$ and $\text{int dom } A^*$ is convex, we have

$$\nabla A(\theta) - \tau \nabla f_{\pi}^{(\alpha)}(\theta) \in \text{int dom } A^*.$$

Due to A having the Legendre property, we have from Proposition 8 that there exists a unique $\theta_{\gamma} \in \text{int } \Theta$ such that

$$\nabla A(\theta_{\gamma}) = \nabla A(\theta) - \tau \nabla f_{\pi}^{(\alpha)}(\theta).$$

We have thus shown that θ_{γ} uniquely solves the necessary optimality conditions involved in the definition of $\gamma_{\tau f_{\pi}^{(\alpha)}}^A$, meaning that $\gamma_{\tau f_{\pi}^{(\alpha)}}^A(\theta)$ is well-defined, as it is equal to θ_{γ} which is uniquely defined, and that $\gamma_{\tau f_{\pi}^{(\alpha)}}^A(\theta) \in \text{int } \Theta$, since $\theta_{\gamma} \in \text{int } \Theta$.

(ii) We conclude about the proximal operator with (Bauschke et al., 2003, Proposition 3.21 (vi)), which ensures that $\text{dom prox}_{\tau r}^A = \text{int } \Theta$, with (Bauschke et al., 2003, Proposition 3.23 (v)) which ensures that $\text{ran prox}_{\tau_{k+1}r}^A \subset \text{int dom } A$, and with (Bauschke et al., 2003, Proposition 3.22 (ii)(d)), showing that $\text{prox}_{\tau r}^A$ is single-valued.

(iii) The third point comes from (Gao et al., 2020, Lemma 3). ■

In order to prove Propositions 27 and 28, we start with a *sufficient decrease lemma* that reads as follows.

Lemma 34 *Under Assumptions 15, 16, and 17, for $\tau > 0$ and $\alpha \in (0, 1]$, we have that for every $\theta \in \text{int } \Theta$,*

$$\tau \left(F_{\pi}^{(\alpha)}(T_{\tau F_{\pi}^{(\alpha)}}^A(\theta)) - F_{\pi}^{(\alpha)}(\theta) \right) \leq -d_A(\theta, T_{\tau F_{\pi}^{(\alpha)}}^A(\theta)) + (\tau - 1)d_A(T_{\tau F_{\pi}^{(\alpha)}}^A(\theta), \theta). \quad (35)$$

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

$$\begin{aligned} \tau \left(F_{\pi}^{(1)}(T_{\tau F_{\pi}^{(1)}}^A(\theta)) - F_{\pi}^{(1)}(\theta') \right) &\leq (1 - \tau)d_A(\theta', \theta) - (1 - \tau)d_A(T_{\tau F_{\pi}^{(1)}}^A(\theta), \theta) \\ &\quad - d_A(\theta', T_{\tau F_{\pi}^{(1)}}^A(\theta)), \forall \theta' \in \text{int } \Theta. \end{aligned} \quad (36)$$

Proof Using (Teboulle, 2018, Lemma 4.1), which still holds in our finite-dimensional Hilbert setting, we get that

$$\begin{aligned} \tau \left(F_{\pi}^{(\alpha)}(T_{\tau F_{\pi}^{(\alpha)}}^A(\theta)) - F_{\pi}^{(\alpha)}(\theta') \right) &\leq d_A(\theta', \theta) - (1 - \tau)d_A(T_{\tau F_{\pi}^{(1)}}^A(\theta), \theta) \\ &\quad - d_A(\theta', T_{\tau F_{\pi}^{(1)}}^A(\theta)) - \tau d_{f_{\pi}^{(\alpha)}}(\theta', \theta), \forall \theta' \in \text{int } \Theta, \end{aligned}$$

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 (35) comes by evaluating the above at $\theta' = \theta$. To get Eq. (36), the strong convexity of $f_\pi^{(1)}$ relatively to A yields

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

showing the result. ■

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

Lemma 35 *Consider two sequences $\{\theta_k\}_{k \in \mathbb{N}}$ and $\{\theta'_k\}_{k \in \mathbb{N}}$ and assume that there exists a compact set $C \subset \text{int } \Theta$ such that $\theta_k, \theta'_k \in C$ for every $k \in \mathbb{N}$. In this case, $d_A(\theta_k, \theta'_k) \xrightarrow[k \rightarrow +\infty]{} 0$ implies $\|\theta_k - \theta'_k\| \xrightarrow[k \rightarrow +\infty]{} 0$.*

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

A is proper, strictly convex, and continuous on $\text{conv } C \subset \text{int } \Theta$, therefore, A is *uniformly convex*, following the definition of (Bauschke and Combettes, 2011, Definition 10.5) on $\text{conv } C$ (Bauschke and Combettes, 2011, Proposition 10.15). This means that there exists an increasing function $\psi : \mathbb{R}_+ \rightarrow [0, +\infty]$ that vanishes only at 0, such that for every $\theta, \theta' \in \text{conv } C$,

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

Because A is convex on $\text{conv } C$, we prove following the proof of (Butnariu and Iusem, 2000, Eq. (1.32)) that for every $\theta, \theta' \in \text{conv } C$,

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

Suppose now by contradiction that $d_A(\theta_k, \theta'_k) \xrightarrow[k \rightarrow +\infty]{} 0$ 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.2 Proof of Proposition 27

Proof [Proof of Proposition 27]

We first give an intermediate result that will be used several times. Using Lemma 34, we can get for any $k \in \mathbb{N}$ that

$$F_\pi^{(\alpha)}(\theta_{k+1}) - F_\pi^{(\alpha)}(\theta_k) \leq -\frac{1}{\tau_{k+1}}d_A(\theta_k, \theta_{k+1}) - \left(\frac{1}{\tau_{k+1}} - 1\right)d_A(\theta_{k+1}, \theta_k). \quad (37)$$

(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. (37) 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 34 and $\tau_{K+1} \leq 1$, $d_A(\theta_K, \theta_{K+1}) \leq 0$. By Proposition 8, this shows that $\theta_{K+1} = \theta_K$ and hence, that θ_K is a fixed point of Algorithm 1. From Proposition 26, it is a stationary point of $F_\pi^{(\alpha)}$.

(iii) By summing Eq. (37) over $k \in \llbracket 0, K \rrbracket$, 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 $k \in \llbracket 0, K-1 \rrbracket$, $d_A(\theta_k, \theta_{k+1}) > \varepsilon$. Consider Eq. (B.2) 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 (Penot, 2013, Chapter 6) and the Fréchet subdifferential ∂_F (Penot, 2013, Chapter 4). Our working space \mathcal{H} is a finite-dimensional Hilbert space, which is included in the setting of Penot (2013).

Set $k \in \mathbb{N}$. Under Assumptions 15, 16, and 17, since $\theta_0 \in \text{int } \Theta$, Proposition 26 applies and thus $\theta_{k+1} = T_{\tau_{k+1} F_\pi^{(\alpha)}}^A(\theta_k)$. This implies that there exists $u_{k+1} \in \partial r(\theta_{k+1})$ such that

$$\frac{1}{\tau_{k+1}} (\nabla A(\theta_{k+1}) - \nabla A(\theta_k)) + \nabla f_\pi^{(\alpha)}(\theta_k) + u_{k+1} = 0. \quad (38)$$

Denote $\rho_{k+1} = \nabla f_\pi^{(\alpha)}(\theta_{k+1}) + u_{k+1} \in \nabla f_\pi^{(\alpha)}(\theta_{k+1}) + \partial r(\theta_{k+1})$. We then organize Equation (38) as $\tau_{k+1}(\rho_{k+1} + \nabla f_\pi^{(\alpha)}(\theta_k) - \nabla f_\pi^{(\alpha)}(\theta_{k+1})) = \nabla A(\theta_k) - \nabla A(\theta_{k+1})$. Using the triangle inequality and the additional assumption on the step sizes, we obtain that

$$\|\rho_{k+1}\| \leq \|\nabla f_\pi^{(\alpha)}(\theta_{k+1}) - \nabla f_\pi^{(\alpha)}(\theta_k)\| + \frac{1}{\epsilon} \|\nabla A(\theta_{k+1}) - \nabla A(\theta_k)\|. \quad (39)$$

By assumption, θ_{k+1} and θ_k belong to C , a compact set included in $\text{int } \Theta$. Since $\nabla^2 f_\pi^{(\alpha)}$ is continuous on C (by Proposition 13) 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

$$\|\varrho_{k+1}\| \leq s \|\theta_{k+1} - \theta_k\|. \quad (40)$$

Now, we deduce from (iii) that $d_A(\theta_{k+1}, \theta_k) \xrightarrow[k \rightarrow +\infty]{} 0$. Using Lemma 35, this yields $\|\theta_{k+1} - \theta_k\| \xrightarrow[k \rightarrow +\infty]{} 0$, showing that the sequence $\{\varrho_k\}_{k \in \mathbb{N}}$ is such that

$$\varrho_k \in \partial_F F_\pi^{(\alpha)}(\theta_k), \forall k \in \mathbb{N}, \text{ and } \varrho_k \xrightarrow[k \rightarrow +\infty]{} 0. \quad (41)$$

We now turn to the second part of the result. Consider any limit point θ_{\lim} of $\{\theta_k\}_{k \in \mathbb{N}}$. This means that there exists a strictly increasing function $\varphi : \mathbb{N} \rightarrow \mathbb{N}$ such that $\theta_{\varphi(k)} \xrightarrow[k \rightarrow +\infty]{} \theta_{\lim}$ and $\theta_{\lim} \in C$ since C is compact. Further, we have that $\rho_k \in \partial_F F_\pi^{(\alpha)}(\theta_k)$ according to (Penot, 2013, Corollary 4.35) and that the regularizing term r is continuous on C by assumption. Therefore, we have

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

By definition of the limiting subdifferential $\partial_L F_\pi^{(\alpha)}$ (Penot, 2013, Definition 6.1), this shows that $0 \in \partial_L F_\pi^{(\alpha)}(\theta_{\lim})$. Hence θ_{\lim} is a stationary point of $F_\pi^{(\alpha)}$ using the characterization of $\partial_L F_\pi^{(\alpha)}$ given in (Penot, 2013, Proposition 6.17). \blacksquare

B.3 Proof of Proposition 28

Proof [Proof of Proposition 28] We first give a useful inequality to prove our results. Consider iteration k of Algorithm 1, and evaluate Eq. (36) from Lemma 34 at $\theta' = \theta_*$, yielding

$$\begin{aligned} \tau_{k+1} \left(F_\pi^{(\alpha)}(\theta_{k+1}) - F_\pi^{(\alpha)}(\theta_*) \right) &\leq (1 - \tau_{k+1}) d_A(\theta_*, \theta_k) \\ &\quad - (1 - \tau_{k+1}) d_A(\theta_{k+1}, \theta_k) - d_A(\theta_*, \theta_{k+1}). \end{aligned} \quad (42)$$

(i) From Equation (42), we obtain, using the non-negativity of the Bregman divergence and the fact that $\tau_{k+1} \in (0, 1)$, that $\tau_{k+1} \left(F_\pi^{(\alpha)}(\theta_{k+1}) - F_\pi^{(\alpha)}(\theta_*) \right) \leq d_A(\theta_k, \theta_*) - d_A(\theta_{k+1}, \theta_*)$. Summing this inequality over $k \in \llbracket 0, K \rrbracket$, we get

$$\sum_{k=0}^K \tau_{k+1} \left(F_\pi^{(\alpha)}(\theta_{k+1}) - F_\pi^{(\alpha)}(\theta_*) \right) \leq -d_A(\theta_*, \theta_{K+1}) + d_A(\theta_*, \theta_0) \leq d_A(\theta_*, \theta_0). \quad (43)$$

The bound on the right-hand side of Equation (43) does not depend on K , and given our assumption on the sum of the step sizes, we deduce that $F_\pi^{(\alpha)}(\theta_k) \xrightarrow[k \rightarrow +\infty]{} F_\pi^{(\alpha)}(\theta_*)$.

Using Proposition 27 (i), we obtain that for every $k \in \mathbb{N}$, $F_\pi^{(1)}(\theta_k) \leq F_\pi^{(1)}(\theta_0)$, 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 from Proposition 21, and it is lower semicontinuous from Proposition 9, 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 $\theta_{\varphi(k)} \xrightarrow[k \rightarrow +\infty]{} \theta_{\text{lim}}$. $F_\pi^{(1)}$ is lower semicontinuous, so

$$\liminf_{k \rightarrow +\infty} F_\pi^{(1)}(\theta_{\varphi(k)}) \geq F_\pi^{(1)}(\theta_{\text{lim}}).$$

However, because of the previous point, $\liminf F_\pi^{(1)}(\theta_{\varphi(k)}) = F_\pi^{(1)}(\theta_*)$, so we obtain that $F_\pi^{(1)}(\theta_{\text{lim}}) = F_\pi^{(1)}(\theta_*)$. The minimizer of $F_\pi^{(1)}$ is unique from Proposition 21, showing that $\theta_{\text{lim}} = \theta_*$.

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.

(ii) Since $\tau_{k+1} \in [\epsilon, 1]$, $F_\pi^{(1)}(\theta_{k+1}) \geq F_\pi^{(1)}(\theta_*)$, and d_A takes non-negative values (from Proposition 8), Eq. (42) gives

$$d_A(\theta_*, \theta_{k+1}) \leq (1 - \tau_{k+1})d_A(\theta_*, \theta_k), \quad (44)$$

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

(iii) Since $\tau_{k+1} \in [\epsilon, 1]$ and d_A takes non-negative values, we get from Eq. (42) that

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

With Eq. (44) and the condition on τ_{k+1} , we obtain

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

from which we conclude using point (ii) and Proposition 21. ■

B.4 Proof of Proposition 29

Proof [Proof of Proposition 29]

We first prove that we can apply the results of Proposition 27. Assumption 16 holds because of the result of Proposition 24. Since $r \equiv 0$, Assumption 17 holds. Therefore, all the hypotheses of Proposition 27 hold, showing the result.

(i) Because of Proposition 27, 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 24). 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 20, we have for any $k \in \mathbb{N}$ that

$$f_\pi^{(\alpha)}(\theta_{k+1}) \leq f_\pi^{(\alpha)}(\theta_k) + \langle \nabla f_\pi^{(\alpha)}(\theta_k), \theta_{k+1} - \theta_k \rangle + d_A(\theta_{k+1}, \theta_k). \quad (45)$$

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. (45) reads

$$f_\pi^{(\alpha)}(\theta_{k+1}) \leq f_\pi^{(\alpha)}(\theta_k) - \frac{1}{\tau_{k+1}} \langle \nabla A(\theta_{k+1}) - \nabla A(\theta_k), \theta_{k+1} - \theta_k \rangle + d_A(\theta_{k+1}, \theta_k). \quad (46)$$

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

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

Now, we consider some $v > 0$. Since $\theta_k \rightarrow \theta_\pi$, there exists $K \in \mathbb{N}$ such that for any $k \geq K$, $\theta_k \in B(\theta_\pi, v)$ and $\nabla A(\theta_k) \in B(\nabla A(\theta_\pi), v)$ (recall that ∇A is continuous on $\text{int } \Theta$). Thus, we can write following the same steps as in the proof of Proposition 25 that for any $k \geq K$,

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

From there, we obtain that

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

where the last inequality comes from Proposition 25. With our previous point, this yields

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

for any constant $\delta \in (0, 1)$. By choosing δ or v 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 $M = (1 - \alpha \delta \epsilon)^{-K}$. ■

B.5 Proof of Proposition 31

Proof [Proof of Proposition 31]

(i) At every iteration $k \in \mathbb{N}$, we have that $\nabla A(\theta_{k+1}) = \nabla A(\theta_k) - \tau_{k+1}(\nabla f_\pi^{(\alpha)}(\theta_k) + n_{k+1} + u_{k+1})$ where $u_{k+1} \in \partial r(\theta_{k+1})$ and

$$n_{k+1} = \sum_{n=1}^{N_{k+1}} \bar{w}_n^{(\alpha)} \Gamma(x_n) - \pi_{\theta_k}^{(\alpha)}(\Gamma).$$

Therefore, we have that

$$\tau_{k+1} \langle \nabla f_\pi^{(\alpha)}(\theta_k) + n_{k+1} + u_{k+1}, \theta_{k+1} - \theta_k \rangle = \langle \nabla A(\theta_k) - \nabla A(\theta_{k+1}), \theta_{k+1} - \theta_k \rangle. \quad (47)$$

Using the three-points equality (Teboulle, 2018, Lemma 2.2), we obtain that

$$d_A(\theta_k, \theta_{k+1}) = \tau_{k+1} \langle \nabla f_\pi^{(\alpha)}(\theta_k) + n_{k+1} + u_{k+1}, \theta_k - \theta_{k+1} \rangle - d_A(\theta_{k+1}, \theta_k). \quad (48)$$

We now bound the terms appearing in the right-hand side of Equation (48).

First, r is convex from Assumption 17 and $u_{k+1} \in \partial r(\theta_{k+1})$, so we have $\tau_{k+1} \langle u_{k+1}, \theta_k - \theta_{k+1} \rangle \leq \tau_{k+1} (r(\theta_k) - r(\theta_{k+1}))$.

Second, $f_\pi^{(\alpha)}$ is 1-relatively smooth from Proposition 20, ensuring that $\langle \nabla f_\pi^{(\alpha)}(\theta_k), \theta_k - \theta_{k+1} \rangle \leq d_A(\theta_{k+1}, \theta_k) - f_\pi^{(\alpha)}(\theta_{k+1}) + f_\pi^{(\alpha)}(\theta_k)$.

We thus get from Equation (48) and the two previously stated facts that

$$d_A(\theta_k, \theta_{k+1}) \leq -(1 - \tau_{k+1}) d_A(\theta_{k+1}, \theta_k) + \tau_{k+1} (F_\pi^{(\alpha)}(\theta_k) - F_\pi^{(\alpha)}(\theta_{k+1})) + \tau_{k+1} \langle n_{k+1}, \theta_k - \theta_{k+1} \rangle. \quad (49)$$

Now, since $\tau_{k+1} \in (0, 1]$ and the Bregman divergences take non-negative values, $-(1 - \tau_{k+1}) d_A(\theta_{k+1}, \theta_k) \leq 0$. Using the Cauchy-Schwarz inequality and leveraging that $\tau_{k+1} \leq 1$, we further obtain

$$d_A(\theta_k, \theta_{k+1}) \leq F_\pi^{(\alpha)}(\theta_k) - F_\pi^{(\alpha)}(\theta_{k+1}) + \|n_{k+1}\| \|\theta_k - \theta_{k+1}\|. \quad (50)$$

Remark that the iterates staying almost surely in the compact C ensures that $\|\theta_k - \theta_{k+1}\| \leq \sup_{\theta_1, \theta_2 \in C} \|\theta_1 - \theta_2\| := \text{diam } C < +\infty$.

We now introduce for any $k \in \mathbb{N}$ the filtration \mathcal{F}_{k+1} , which is the σ -algebra defined by $\mathcal{F}_{k+1} = \sigma(\theta_0, \{x_n^1\}_{n=1}^{N_1}, \dots, \theta_k, \{x_n^{k+1}\}_{n=1}^{N_{k+1}})$. Taking expectation with respect to F_{k+1} in Equation (50) yields

$$\mathbb{E}[d_A(\theta_k, \theta_{k+1}) | \mathcal{F}_{k+1}] \leq F_\pi^{(\alpha)}(\theta_k) - \mathbb{E}[F_\pi^{(\alpha)}(\theta_{k+1}) | \mathcal{F}_{k+1}] + (\text{diam } C) \mathbb{E}[\|n_{k+1}\| | \mathcal{F}_{k+1}]. \quad (51)$$

Using Jensen's inequality (using that the square root is concave), we obtain the bound $\mathbb{E}[\|n_{k+1}\| | \mathcal{F}_{k+1}] \leq \sqrt{\mathbb{E}[\|n_{k+1}\|^2 | \mathcal{F}_{k+1}]}$. Because of Assumption 30 on the sampling procedure and the fact that the iterates are bounded, we finally obtain, by taking expectation, that

$$\mathbb{E}[d_A(\theta_k, \theta_{k+1})] \leq \mathbb{E}[F_\pi^{(\alpha)}(\theta_k)] - \mathbb{E}[F_\pi^{(\alpha)}(\theta_{k+1})] + \frac{M}{\sqrt{N_{k+1}}}, \quad (52)$$

with $M = \sqrt{\sup_C E_{\pi, \mathcal{Q}}^{(\alpha)}} \text{diam } C \in (0, +\infty)$. Now, summing the above for $k \in \llbracket 0, K \rrbracket$ yields

$$\sum_{k=0}^K \mathbb{E}[d_A(\theta_k, \theta_{k+1})] \leq \mathbb{E}[F_\pi^{(\alpha)}(\theta_0)] - \inf_C F_\pi^{(\alpha)} + M \sum_{k \geq 0} \frac{1}{\sqrt{N_{k+1}}} \quad (53)$$

using the assumption on the sample sizes as well. This ensures that $\mathbb{E} \left[\sum_{k \geq 0} d_A(\theta_k, \theta_{k+1}) \right] < +\infty$. In particular, $\mathbb{E}[d_A(\theta_k, \theta_{k+1})] \xrightarrow[k \rightarrow +\infty]{} 0$, so we obtain the convergence in probability from Markov's inequality.

(ii) Since the iterates are supposed to stay in $C \subset \text{int } \Theta$, we can write, as in the proof of Proposition 27 (v), that there exists $s > 0$ satisfying

$$\|\rho_{k+1}\| \leq s\|\theta_{k+1} - \theta_k\| + \|n_{k+1}\| \quad (54)$$

for any $k \in \mathbb{N}$. Taking expectation with respect to the filtration \mathcal{F}_{k+1} and leveraging Assumption 30 as in the proof of (i) then gives

$$\mathbb{E}[\|\rho_{k+1}\|] \leq s\mathbb{E}[\|\theta_{k+1} - \theta_k\|] + \frac{\sqrt{\sup_C E_{\pi, \mathcal{Q}}^{(\alpha)}}}{\sqrt{N_{k+1}}}. \quad (55)$$

From (i), and using Lemma 35, we get that $\mathbb{E}[\|\theta_{k+1} - \theta_k\|] \xrightarrow[k \rightarrow +\infty]{} 0$. Due to the assumption on the step sizes, we also have that $\frac{1}{\sqrt{N_{k+1}}} \xrightarrow[k \rightarrow +\infty]{} 0$. Finally, Assumption 30 on the sampling procedure ensures that $\sup_C E_{\pi, \mathcal{Q}}^{(\alpha)} < +\infty$. These facts imply that $\mathbb{E}[\|\rho_{k+1}\|] \xrightarrow[k \rightarrow +\infty]{} 0$. Using Markov's inequality, we finally obtain that $\{\rho_k\}_{k \in \mathbb{N}}$ converges in probability to 0, with $\rho_k \in f_{\pi}^{(\alpha)}(\theta_k) + \partial r(\theta_k)$ for every $k \in \mathbb{N}$. ■

B.6 Proof of Proposition 32

Proof [Proof of Proposition 32]

(i) At every iteration $k \in \mathbb{N}$, we have that $\nabla A(\theta_{k+1}) = \nabla A(\theta_k) - \tau_{k+1}(\nabla f_{\pi}^{(\alpha)}(\theta_k) + n_{k+1} + u_{k+1})$ where $u_{k+1} \in \partial r(\theta_{k+1})$ and

$$n_{k+1} = \sum_{n=1}^{N_{k+1}} \bar{w}_n^{(\alpha)} \Gamma(x_n) - \pi_{\theta_k}^{(\alpha)}(\Gamma).$$

Therefore, we have for any $\theta \in \text{int } \Theta$ that

$$\tau_{k+1} \langle \nabla f_{\pi}^{(\alpha)}(\theta_k) + n_{k+1} + u_{k+1}, \theta_{k+1} - \theta \rangle = \langle \nabla A(\theta_k) - \nabla A(\theta_{k+1}), \theta_{k+1} - \theta \rangle. \quad (56)$$

Using the three-points equality (Teboulle, 2018, Lemma 2.2), we obtain that

$$d_A(\theta, \theta_{k+1}) = d_A(\theta, \theta_k) + \tau_{k+1} \langle \nabla f_{\pi}^{(\alpha)}(\theta_k) + n_{k+1} + u_{k+1}, \theta - \theta_{k+1} \rangle - d_A(\theta_{k+1}, \theta_k). \quad (57)$$

We now bound the terms appearing in the right-hand side of Equation (57).

First, because of Assumption 17 on r and because $u_{k+1} \in \partial r(\theta_{k+1})$, we have that $\tau_{k+1} \langle u_{k+1}, \theta - \theta_{k+1} \rangle \leq \tau_{k+1}(r(\theta) - r(\theta_{k+1}))$.

Second, we write $\langle \nabla f_{\pi}^{(\alpha)}(\theta_k), \theta - \theta_{k+1} \rangle = \langle \nabla f_{\pi}^{(\alpha)}(\theta_k), \theta - \theta_k \rangle - \langle \nabla f_{\pi}^{(\alpha)}(\theta_k), \theta_{k+1} - \theta_k \rangle$. On the one hand, $f_{\pi}^{(1)}$ is 1-relatively strongly convex from Proposition 20, so

$$\langle \nabla f_{\pi}^{(\alpha)}(\theta_k), \theta - \theta_k \rangle \leq f_{\pi}^{(\alpha)}(\theta) - f_{\pi}^{(\alpha)}(\theta_k) - d_A(\theta, \theta_k).$$

On the other hand, $f_{\pi}^{(1)}$ is also 1-relatively smooth from Proposition 20, allowing to write

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

We thus get that

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

Gathering these results concerning the terms involving r and the terms involving $f_\pi^{(\alpha)}$, we thus obtain from Equation (57) that

$$d_A(\theta, \theta_{k+1}) \leq (1 - \tau_{k+1}) d_A(\theta, \theta_k) + \tau_{k+1} (F_\pi^{(\alpha)}(\theta) - F_\pi^{(\alpha)}(\theta_{k+1})) + \tau_{k+1} \langle n_{k+1}, \theta - \theta_{k+1} \rangle - (1 - \tau_{k+1}) d_A(\theta_{k+1}, \theta_k). \quad (58)$$

We now evaluate Equation (58) at $\theta = \theta_*$, giving that $F_\pi^{(\alpha)}(\theta_*) - F_\pi^{(\alpha)}(\theta_{k+1}) \leq 0$. We also notice that the assumption on the step sizes give $(1 - \tau_{k+1}) d_A(\theta_{k+1}, \theta_k) \geq 0$. These two remarks give a simplified version of Equation (58) that reads as

$$d_A(\theta_*, \theta_{k+1}) \leq (1 - \tau_{k+1}) d_A(\theta_*, \theta_k) + \tau_{k+1} \langle n_{k+1}, \theta_* - \theta_{k+1} \rangle. \quad (59)$$

Using the Cauchy-Schwarz inequality, we further obtain

$$d_A(\theta_*, \theta_{k+1}) \leq (1 - \tau_{k+1}) d_A(\theta_*, \theta_k) + \tau_{k+1} \|n_{k+1}\| \|\theta_* - \theta_{k+1}\|. \quad (60)$$

Because the iterates remain inside the compact set C almost surely, we have with probability one that

$$d_A(\theta_*, \theta_{k+1}) \leq (1 - \tau_{k+1}) d_A(\theta_*, \theta_k) + \tau_{k+1} \sup_C \|\theta_* - \cdot\| \|n_{k+1}\|. \quad (61)$$

We will now handle the noise. Consider \mathcal{F}_{k+1} the filtration defined in the proof of Proposition 31. Then, taking expectation yields

$$\mathbb{E}[d_A(\theta_*, \theta_{k+1}) | \mathcal{F}_{k+1}] \leq (1 - \tau_{k+1}) d_A(\theta_*, \theta_k) + \tau_{k+1} \sup_C \|\theta_* - \cdot\| \mathbb{E}[\|n_{k+1}\| | \mathcal{F}_{k+1}]. \quad (62)$$

With Jensen's inequality, we obtain the bound $\mathbb{E}[\|n_{k+1}\| | \mathcal{F}_{k+1}] \leq \sqrt{\mathbb{E}[\|n_{k+1}\|^2 | \mathcal{F}_{k+1}]}$. This allows us to apply Assumption 30 to obtain that $\mathbb{E}[\|n_{k+1}\| | \mathcal{F}_{k+1}] \leq \sqrt{\frac{E_{\pi, Q}^{(\alpha)}(\theta_{k+1})}{N_{k+1}}}$. Since $E_{\pi, Q}^{(\alpha)}$ is locally bounded on $\text{int } \Theta$ and the iterates stay in the compact C with probability one, we finally obtain that

$$\mathbb{E}[d_A(\theta_*, \theta_{k+1}) | \mathcal{F}_{k+1}] \leq (1 - \tau_{k+1}) d_A(\theta_*, \theta_k) + \frac{\tau_{k+1}}{\sqrt{N_{k+1}}} \sup_C \|\theta_* - \cdot\| \sqrt{\sup_C E_{\pi, Q}^{(\alpha)}} \quad (63)$$

with probability one. We define $M = \sup_C \|\theta_* - \cdot\| \sqrt{\sup_C E_{\pi, Q}^{(\alpha)}} \in (0, +\infty)$. We then get the result by taking expectation in the above and applying this inequality for every iteration.

(ii) Using that $\tau_{k+1} \leq 1$, we rewrite Eq. (63) as

$$\mathbb{E}[d_A(\theta_*, \theta_{k+1}) | \mathcal{F}_{k+1}] \leq d_A(\theta_*, \theta_k) - \tau_{k+1} d_A(\theta_*, \theta_k) + \frac{M}{\sqrt{N_{k+1}}}. \quad (64)$$

Under our assumptions on the sample sizes, we can use (Robbins and Siegmund, 1971, Theorem 1) to obtain that $\{d_A(\theta_*, \theta_k)\}_{k \in \mathbb{N}}$ converges almost surely to a non-negative random variable and that $\sum_{k \geq 0} \tau_{k+1} d_A(\theta_*, \theta_k) < +\infty$ almost surely. Due to the step sizes not being summable, this implies that $d_A(\theta_*, \theta_k) \xrightarrow[k \rightarrow +\infty]{a.s.} 0$. We then get the result from Lemma 35. ■

Appendix C. Computation of a Bregman proximal operator

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

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

$$r(\theta) = \sum_{i=1}^d \eta_i |(\theta_1)_i|, \quad (65)$$

where $\eta_i \geq 0$ for $i \in \llbracket 1, d \rrbracket$.

Since $\sigma_i^2 > 0$ for all $i \in \llbracket 1, d \rrbracket$, 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 $\check{\theta} = \text{prox}_{r,r}^A(\theta)$.

Lemma 36 *Consider the Gaussian family defined above. Consider q_θ in this family, with $\theta \in \text{int } \Theta$ and whose mean and covariance are respectively μ and $Q \text{diag}(\sigma_1^2, \dots, \sigma_d^2) Q^\top$. If we consider the regularizing function defined in Eq. (65), then $\check{\theta} = \text{prox}_{r,r}^A(\theta)$ is such that the mean $\check{\mu}$ and covariance $Q \text{diag}(\check{\sigma}_1^2, \dots, \check{\sigma}_d^2) Q^\top$ of $q_{\check{\theta}}$ satisfy for any $i \in \llbracket 1, d \rrbracket$*

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

$$\check{\sigma}_i^2 = (\sigma_i)^2 + ((Q^\top \mu)_i^2 - (Q^\top \check{\mu})_i^2).$$

Consider $i \in \llbracket 1, d \rrbracket$. In the particular case where $\eta_i = 0$, then $\mu_i^* = \mu_i$ and $\check{\sigma}_i^2 = (\sigma_i)^2$. We can also remark that we always have $\check{\sigma}_i^2 \geq \sigma_i^2$, with equality if and only if $(Q^\top \mu)_i = 0$. Therefore, the operator $\text{prox}_{r,r}^A$ 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 $i \in \llbracket 1, d \rrbracket$. This is justified by (Bauschke and Combettes, 2011, 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 $i \in \llbracket 1, d \rrbracket$, we have

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

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

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

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

from which we can obtain the result. ■

Appendix 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

$$\tilde{\pi}(x) = \exp\left(-\frac{1}{2}(x - \bar{\mu})^\top \bar{\Sigma}_\kappa^{-1}(x - \bar{\mu})\right), \forall x \in \mathbb{R}^d. \quad (66)$$

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 (Moré and Toraldo, 1989, 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 .

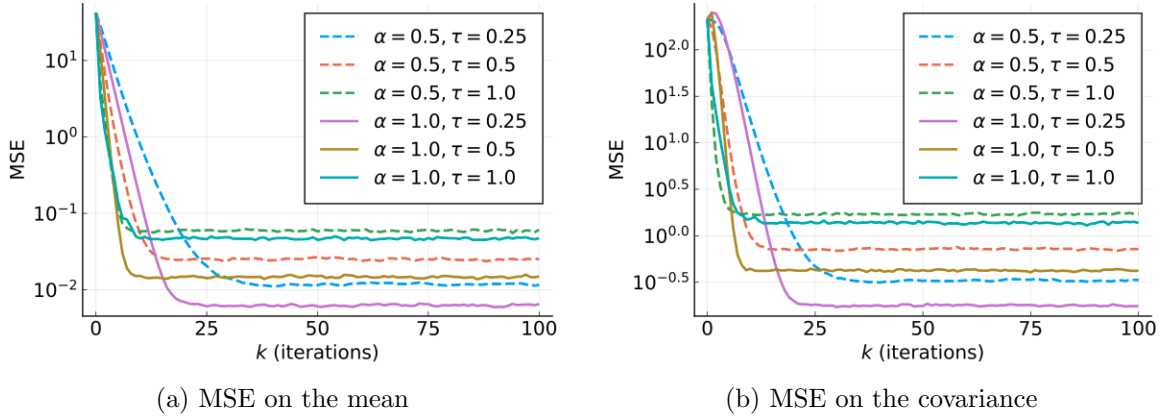
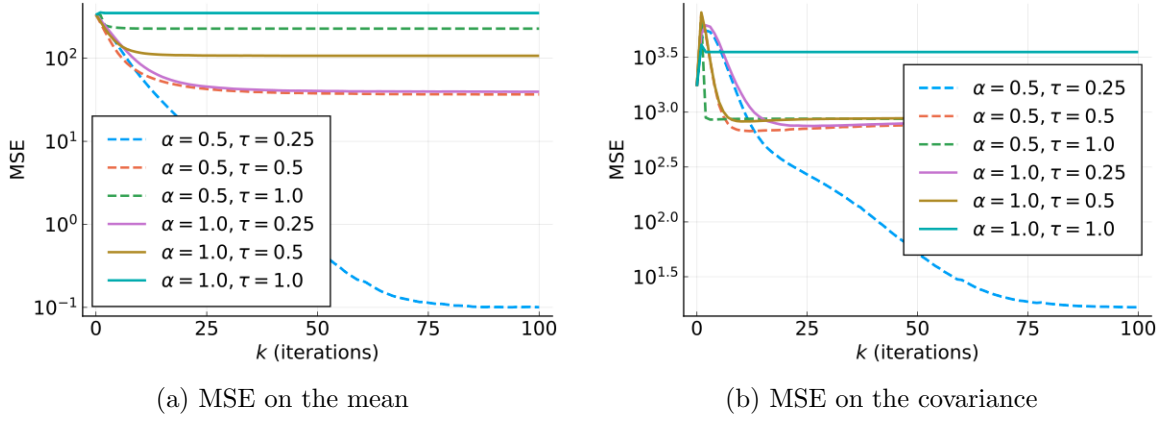
The influence of τ can be understood through the theory on stochastic Bregman gradient descent with fixed step-size. In particular, (Hanzely and Richtárik, 2021, Theorem 5.3) states that such methods converge to a neighborhood of the optimum, whose size decreases with τ . On the other hand, low values 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 (Koblents and Míguez, 2013).

In order to highlight this compromise between speed and robustness, we use the RMM algorithm to approximate the target described in Eq. (66) 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 Bengsston et al. (2008). 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.

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


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

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

approximation errors. As expected, the convergence is slower and the final MSE values are higher than in lower dimensions.

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 27 and 28) but also shows that there is no equivalent guarantees for the VRB algorithm (see Proposition 22). 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. (66), 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.

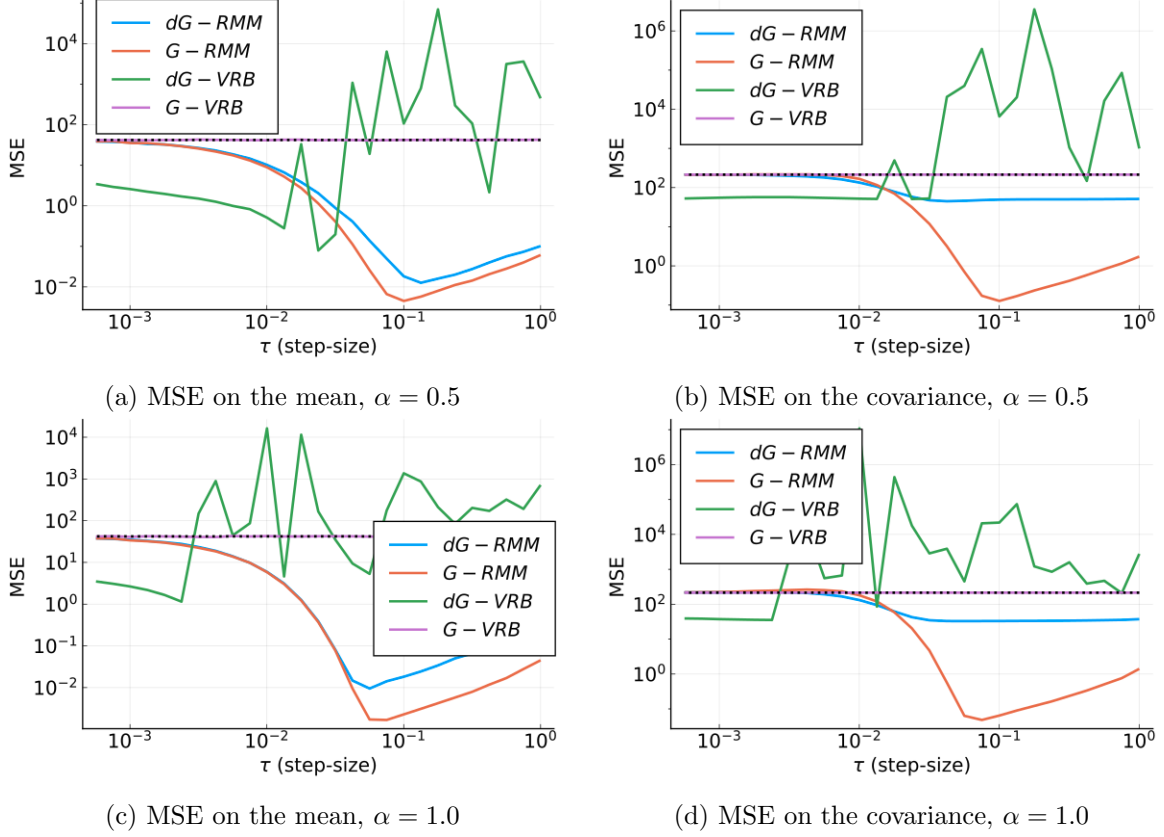


Figure 7: MSE in the estimation of $\bar{\mu}$ and $\bar{\Sigma}_\kappa$ ($d = 5$) after 100 iterations, against values of τ . For each value of τ , 10^3 runs with 500 samples per iteration are conducted. The dotted black lines represent the MSE at initialization. The prefix dG refer to the family of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

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 4.

References

- S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart. Importance sampling: Computational complexity and intrinsic dimension. *Statistical Science*, 32, 11 2015.
- O. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(12), 2021.
- S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
- S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.
- Y. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.
- R. Bai, V. Ročková, and E. I. George. *Spike-and-Slab meets LASSO: A review of the Spike-and-Slab LASSO*. Chapman and Hall/CRC, 2021.
- O. Banerjee, L. El 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.
- O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, Ltd, 2014.
- H. Bauschke and J.M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4:27–67, 1997.
- H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.
- H. Bauschke, J. Borwein, and P. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.
- 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.
- A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, 1999.
- 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.

- F. Bertholom, R. Douc, and F. Roueff. Asymptotics of alpha-divergence variational inference with exponential families. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- 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.
- 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.
- L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, 1986.
- 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.
- T. Bui. Connecting the thermodynamic variational objective and annealed importance sampling. Technical report, 2020.
- 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.
- D. Butnariu and A. Iusem. *Totally convex functions*, chapter 1. Kluwer Academic Publisher, 2000.
- T. Campbell and B. Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- O. Cappé, R. Douc, A. Guillin, J.M. Marin, and C.P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.
- N. L. Carothers. *Real Analysis*. Cambridge University Press, 2000.
- R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- P.L. Combettes and J.-C. Pesquet. *Proximal Splitting Methods in Signal Processing*, page 185–212. Springer-Verlag, New York, 2010.
- 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.
- I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial*. Now Foundations and Trends, 2004.
- 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.

- K. Daudel, R. Douc, and F. Roueff. Monotonic Alpha-divergence minimization. *Journal of Machine Learning Research*, 62(24):1–76, 2023.
- 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.
- A. Dieuleveut, G. Fort, E. Moulines, and H.-T. Wai. Stochastic approximation beyond gradient for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 71:3117–3148, 2023.
- J. Domke, G. Garrigos, and R. Gower. Provable convergence guarantees for black-box variational inference. <https://arxiv.org/abs/2306.03638>, 2023.
- R. Douc, A. Guillin, J.M. Marin, and C.P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Annals of Statistics*, 35:420–448, 2007.
- P. Doukhan and G. Lang. Evaluation for moments of a ratio with application to regression estimation. *Bernoulli*, 15(4):1259–1286, 2009.
- 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.
- R.L. Dykstra. An iterative procedure for obtaining I -projections onto the intersection of convex sets. *The Annals of Probability*, 13(3):975–984, 1985.
- 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.
- 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.
- 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.
- S. Gruffaz, K. Kim, A. Durmus, and J. Gardner. Stochastic approximation with biased MCMC for expectation maximization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2332–2340, 2024.
- F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *Computational Optimization and Applications*, 79(3):717–766, 2021.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.
- J. Hensman, M. Rattray, and N.D. Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, 2012.

- 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.
- J.-B. Hiriart-Urruty and C. Lemaréchal. *Abstract Duality for Practitioners*, pages 137–193. Springer, 1993.
- M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.
- 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.
- 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.
- E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.
- G. Ji, D. Sujono, and E. B. Sudderth. Marginalized stochastic natural gradients for black-box variational inference. In *International Conference on Machine Learning (ICML)*, volume 139, pages 4870–4881, 2021.
- 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.
- G. Khan and J. Zhang. When information geometry meets optimal transport. *Information Geometry*, 5:47–78, 2022.
- 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.
- 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.
- M. E. Khan and D. Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *International Symposium on Information Theory and its Applications (ISITA)*, page 31635, 2018.
- K. Kim, J. Oh, K. Wu, Y. Ma, and J. Gardner. On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.
- J. Knappe and P. de Valpine. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology*, 93(2):256–263, 2012.

- 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.
- 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.
- 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.
- M. Ashok Kumar and I. Sason. Projection theorems for the Rényi divergence on α -convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935, 2016.
- M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.
- Y. Li and R. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.
- W. Lin, M.E. 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.
- C. Margossian and L. Saul. The shrinkage-delinkage trade-off: An analysis of factorized Gaussian approximations for variational inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.
- J.-M. Marin, P. Pudlo, and M. Sedki. Consistency of adaptive importance sampling and recycling schemes. *Bernoulli*, 25(3):1977–1998, 2019.
- 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.
- L. Martino, V. Elvira, J. Míguez, A. Artés-Rodríguez, and P.M. Djurić. A comparison of clipping strategies for importance sampling. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 558–562, 2018.
- V. Masrani, R. Brekelmans, T. Bui, F. Nielsen, A. Galstyan, G. Ver 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.
- T. Minka. Divergence measures and message passing. Technical report, 2005.
- P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

- J. J. Moré and G. Toraldo. Algorithms for bound constrained quadratic programming problems. *Numerische Mathematik*, 55(4):377–400, 1989.
- M.C. 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.
- R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.
- F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *IEEE International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.
- J.-P. Penot. *Calculus without Derivatives*. Springer, 2013.
- 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.
- G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE transactions on Information Theory*, 61(3):1451–1457, 2015.
- K. Ray and B. Szabó. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117:1270–1281, 2022.
- H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and applications. In *Optimizing Methods in Statistics*, pages 233–257, 1971.
- 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.
- 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.
- V. B. Tadić and A. Doucet. Asymptotic bias of stochastic gradient search. *Annals of Applied Probability*, 6(27):3255–3304, 2017.
- M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.
- 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.
- T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions of Information Theory*, 60(7):3797–3820, 2014.
- A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. <https://arxiv.org/abs/1507.02646>, 2015.
- 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.

- 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.
- R. Yao and Y. Yang. Mean field variational inference via Wasserstein gradient flow. <https://arxiv.org/abs/2207.08074>, 2022.
- 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.
- Y. Zheng, A. Fraysse, and T. Rodet. Efficient unsupervised variational Bayesian image reconstruction using a sparse gradient prior. *Neurocomputing*, 359:449–465, 2019.