Sample Average Approximation for Alpha-Divergence Variational Inference with Exponential Convergence Guarantees

François Bertholom
SAMOVAR, Télécom Sud-Paris

François Roueff
LTCI, Télécom Paris

Randal Douc
SAMOVAR, Télécom Sud-Paris

Abstract

We study the problem of approximating an unnormalized target distribution using probability densities from an exponential family. Specifically, we establish convergence guarantees for a monotonic alpha-divergence minimization algorithm, which decreases the alpha-divergence at each iteration. To illustrate our theoretical results, we propose an implementable Sample Average Approximation algorithm that solves a discrete approximation of the original problem. Through a detailed analysis of the loss landscape and the algorithm's dynamics, we provide practical design guidelines which suffice to ensure its convergence.

1 INTRODUCTION

Motivations. Bayesian inference fundamentally relies on accurately computing or sampling from the posterior distribution, a task that quickly becomes challenging as model complexity grows. In many practical scenarios, the posterior is only known up to an intractable normalization constant, rendering exact inference infeasible. Variational Inference (VI) [Blei et al., 2017, Wainwright and Jordan, 2008] is a powerful and widely-adopted class of approximation methods, which simplifies inference by finding a suitable posterior approximation within a tractable family of distributions. Traditional VI methods use the reverse Kullback-Leibler (KL) divergence to assess the quality of the approximation. Despite its popularity, the reverse KL divergence has known limitations, such as posterior variance underestimation [Minka, 2005, Margossian and Saul, 2023, Margossian et al., 2025]. This has driven recent interest in alternative divergence choices, particularly the class of alpha-divergences [Amari, 2016], which offers a flexible framework to help alleviate these issues. In particular, it allows practitioners to tune between mode-seeking and mass-covering behaviors [Daudel and Douc, 2021, Dieng et al., 2017, Geffner and Domke, 2020, Geffner and Domke, 2021, Hernandez-Lobato et al., 2016, Li and Turner, 2016, Wang et al., 2018]. This class encompasses a broad spectrum of divergence measures, including the forward and reverse KL divergences as special cases, and the approach is also connected to tempered posterior approximation [Alquier and Ridgway, 2020, Masrani et al., 2019, Yang et al., 2020].

General setting. Consider a latent variable probabilistic model with an observed variable x, a hidden variable y, and a known joint density $p_\theta(x, y)$. We are interested in approximating the posterior density $p_\theta(y \mid x)$ for fixed values of x and θ. For simplicity, we let $p(y) = p_\theta(x, y)$, so that $p_\theta(y \mid x) \propto p(y)$. The corresponding normalizing constant is called the evidence, and is often intractable. To circumvent this, we rely on a variational family $\mathcal{Q} = \{q_\eta \ : \ \eta \in E\}$, where E is the parameter space. For a given $\alpha \in (0, 1)$, let ℓ_α be the real-valued function defined for $\eta \in E$ by

$$\ell_\alpha(\eta) = \int q_\eta(y)^\alpha p(y)^{1-\alpha} \, \nu(\mathrm{d}y). \tag{1}$$

Above, ν is a measure whose properties will be clarified in Section 2.1. Minimizing the alpha-divergence between the variational approximation and the posterior density is equivalent to solving $\arg\max \ell_\alpha$ [Daudel and Douc, 2021, Appendix A.1]. In this paper, we rigorously investigate a recently proposed iterative algorithm [Daudel et al., 2023] which solves this problem and exhibits the compelling property of decreasing the alpha-divergence at each iteration. The Monotonic Algorithm (MA) introduces a map $(\gamma, \eta) \mapsto \mathcal{M}_\gamma(\eta)$ and constructs a sequence $\eta_{k+1} = \mathcal{M}_{\gamma_k}(\eta_k)$, where $(\gamma_k)_{k \geq 0}$ are step-sizes ensuring that the iterates remain confined to the legal parameter space. The mapping $\mathcal{M}_\gamma$ is generally intractable, but it admits

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

a very simple form when the variational family $\mathcal{Q}$ is an exponential model. However, this simple form involves an expectation that is generally not available in closed form, and thus can only be approximated. Convergence guarantees for this algorithm have been established by [Bertholom et al., 2024], who show that when the problematic expectations are tractable, the algorithm converges exponentially fast. However, when approximating them through Monte-Carlo estimation in a Stochastic Approximation scheme (SA, [Robbins and Monro, 1951]), the reliance of the procedure on biased gradient estimators hinders the theoretical analysis. Moreover, generic methods to obtain exact convergence rates often rely on Lyapunov functions [Dieuleveut et al., 2023], requiring assumptions that generally do not hold for alpha-divergence minimization with exponential families [Guilmeau et al., 2025].

Sample Average Approximation. The stochasticity introduced at each new step of an SA scheme creates an unbridgeable gap with the deterministic MA, which would converge much faster. This motivates the introduction of a framework in which the random samples used to compute the intractable expectations in the SA version of the MA are frozen at the start of the optimization process, and reused throughout. A precise change of measure allows us to cast this as an instanciation of the exact MA on a new discrete alpha-divergence minimization task. This can be seen as a special case of Sample Average Approximation (SAA, [Kim et al., 2014]). In the broad context of black-box VI [Ranganath et al., 2014], recent works have shown promising results using SAA-based methods [Burroni et al., 2024, Giordano et al., 2024, Zimmermann et al., 2024]. While experimental results do not show a significant improvement over using an SA scheme for the MA in terms of raw performance, our proposed SAA algorithm exhibits faster convergence. More importantly, it provides a clear illustration of our theoretical framework's applicability.

Theoretical Guarantees. Our main contibution is the theoretical study of the deterministic MA. [Bertholom et al., 2024] showed that when $\mathcal{Q}$ is an exponential family of densities, if a) the iterates remain bounded, b) the step-sizes do not vanish, and c) the critical points of ℓ_α satisfy a certain condition; then the iterates converge to a local maximizer of ℓ_α at an asymptotically exponential rate. Leveraging novel insights into the properties of $\mathcal{M}_\gamma$ and developing a refined understanding of the loss landscape allows us to establish similar convergence guarantees in a much broader setting. Specifically, we prove that the assumptions a) to c), which constrain the behavior of the optimization path, can be obtained as consequences of simple design choices. By providing precise guidance regarding the initialization and the step-size selection strategy, we are able to remove a) and b), while weakening c) to a mild algorithm-independent condition. This is a significant improvement, since assumptions about a particular behavior during the optimization process cannot be checked prior to running the algorithm. Moreover, our analysis reveals a natural distinction between continuous and discrete exponential models, which is particularly interesting in the context of our illustrative SAA framework.

2 FRAMEWORK

2.1 General setting

To assist readers unfamiliar with some topological and measure-theoretic concepts, we briefly recall a few definitions in Appendix A.

Problem setup. Throughout this paper, we let Y be a metric space that is separable, Hausdorff, and σ-locally compact, and let $\mathcal{Y}$ denote its Borel σ-field. We consider a non-null measure ν defined on $(\mathsf{Y}, \mathcal{Y})$ and finite on any compact subset of Y, and a measurable function $p : \mathsf{Y} \to \mathbb{R}_+$ satisfying

$$\nu(\{p = 0\}) = 0 \quad \text{and} \quad \int p(y)\, \nu(\mathrm{d}y) < +\infty. \tag{2}$$

For all $s \in [0, 1]$, we let μ_s denote the measure on $(\mathsf{Y}, \mathcal{Y})$ whose density with respect to ν is p^s, i.e.,

$$\mu_s(\mathrm{d}y) = p^s(y)\, \nu(\mathrm{d}y). \tag{3}$$

Let $d \geq 1$, and let $(\mathsf{H}, \langle \cdot, \cdot \rangle)$ be a d-dimensional real inner product space with the induced norm $\|\cdot\|$. Consider a parameter space $E \subseteq \mathsf{H}$ and a family of probability densities $\mathcal{Q} = \{q_\eta : \eta \in E\}$ with respect to ν. By Hölder's inequality and (2), the integral in (1) is well-defined and, for all $\eta \in E$,

$$0 < \ell_\alpha(\eta) \leq \left(\int p \, \mathrm{d}\nu \right)^{1-\alpha} < \infty. \tag{4}$$

Examples. Prominent examples that fit this setup and our topological assumptions include Euclidean spaces $\mathbb{R}^m$ equipped with Lebesgue measures, spherical measures on the n-sphere (used in directional statistics), and counting measures on discrete sets.

2.2 Exponential families

Exponential families play a central role in statistical modeling and machine learning (see, e.g., [Wainwright and Jordan, 2008]). In this section, we introduce the exponential family setting, relate it to our previous assumptions, and highlight several illustrative examples.

Key definitions. Given the dominating measure ν and the inner-product space $(\mathsf{H}, \langle\cdot,\cdot\rangle)$, we say that $\mathcal{Q} = \{q_\eta \,:\, \eta \in E\}$ is an exponential family if there exists a measurable mapping $S : \mathsf{Y} \to \mathsf{H}$ such that

$$q_\eta(y) = \exp\big[\langle \eta, S(y)\rangle - A(\eta)\big], \tag{5}$$

where A is the log-partition function. We let E denote the set for which a density of this form is feasible, i.e., $E = \big\{\eta \in \mathbb{R}^d \,:\, \int \exp\big[\langle \eta, S(y)\rangle\big]\nu(\mathrm{d}y) < +\infty\big\}$, which is a convex subset of H. For any $\eta \in E$, $A(\eta)$ is the constant that normalizes q_η into a probability density.

Throughout this paper, we assume that the exponential family $\mathcal{Q}$ is *regular* and *minimal*. The former means that the set E is open. Then, $\|S\|$ admits a finite exponential moment under the probability $q_\eta \,\mathrm{d}\nu$ for all $\eta \in E$; the log-partition function is infinitely differentiable on E, and its derivatives are related to the cumulants of S. In particular, the gradient and Hessian of A write

$$\nabla A(\eta) = \mathbb{E}_{q_\eta}[S] \quad \text{and} \quad \mathbf{H}_A(\eta) = \mathrm{Cov}_{q_\eta}(S). \tag{6}$$

The minimality assumption holds if no hyperplane in H contains the image of S ν-almost surely. Formally, this means that for all $v \in \mathsf{H} \setminus \{0\}$ and $M \in \mathbb{R}$, we have $\nu(\{\langle v, S\rangle = M\}^c) > 0$. By Lemma 5, a regular exponential model is minimal if and only if $\mathrm{Cov}_{q_\eta}(S) \succ 0$ for some $\eta \in E$ (then, the covariance matrices are positive-definite for all $\eta \in E$). By (6), openness and convexity of E, the mapping ∇A is a $\mathcal{C}^\infty$-diffeomorphism from E onto $F := \nabla A(E)$. Then, for any $\eta \in E$, $\nabla A(\eta)$ can be seen as an alternative parameterization of the family, sometimes referred to as the *mean parameterization*. Hence, F denotes the set of these parameters, and it is an open convex subset of H by [Rockafellar, 1970, Theorem 26.5].

Additional assumptions. We formulate two additional assumptions on the exponential family $\mathcal{Q}$.

(A1) The function S is continuous from Y to H.

(A2) For all $(u, M) \in \mathbb{U} \times \mathbb{R}$, $\nu(\{\langle u, S\rangle = M\}) = 0$.

Above, $\mathbb{U}$ denotes the set of all directions that are asymptotically attained for points in E, i.e.,

$$\mathbb{U} = \bigcap_{r>0} \mathrm{cl}\left\{\frac{\eta}{\|\eta\|} \,:\, \eta \in E\,,\, \|\eta\| > r\right\}. \tag{7}$$

(A1) is mild and easy to check (for many usual exponential families, S is logarithmic or polynomial). In contrast, (A2) is not trivial to check and may not be satisfied when E is unbounded. That said, if E is bounded, the set $\mathbb{U}$ is empty and (A2) trivially holds. It is important to note that the condition $\nu(\{\langle u, S\rangle = M\}) = 0$ is stronger than minimality, which is characterized by $\nu(\{\langle u, S\rangle = M\}^c) > 0$. For instance, it does not hold if the measure ν has one or more atoms, as seen in the Bernoulli model and its extension, the Ising model, even though these are regular minimal exponential models. Despite this restriction, (A2) still applies to many widely used exponential families, e.g., Gaussian, Gamma, chi-squared, von Mises-Fisher, etc.

2.3 Alpha-divergence minimization algorithm

Preliminary definitions and properties. Our study revolves around an algorithm introduced in [Daudel et al., 2023], which consists in successively solving maximization problems to ensure a decrease in the alpha-divergence at each iteration (see Theorem 2). When $\mathcal{Q}$ is a minimal regular exponential model as assumed here, the procedure reduces to a moment-matching algorithm. For $\eta \in E$ and $s \in [0,1]$, we let $\pi_{s,\eta}$ denote the probability density with respect to the dominating measure ν defined by

$$\pi_{s,\eta} \propto q_\eta^s \, p^{1-s}.$$

The normalizing constant depends on the value of s. Note that when $s = \alpha$, it equals $\ell_\alpha(\eta)$ as defined in (1); when $s = 1$, it equals 1; and when $s = 0$, it is the normalizing constant of p, also known as the model evidence. We introduce special notation for the first two moments of S under $\pi_{s,\eta}$:

$$\mathbf{m}_s(\eta) = \mathbb{E}_{\pi_{s,\eta}}[S] \quad \text{and} \quad \mathbf{v}_s(\eta) = \mathrm{Cov}_{\pi_{s,\eta}}(S). \tag{8}$$

In particular, using the fundamental properties of minimal regular exponential families recalled above, $\mathbf{m}_1 = \nabla A$ and $\mathbf{v}_1$ are infinitely differentiable on E, and $\mathbf{v}_1(\eta) \succ 0$ for all $\eta \in E$. Likewise, using (2), the functions ℓ_α, $\mathbf{m}_\alpha$ and $\mathbf{v}_\alpha$ are infinitely differentiable on E, and $\mathbf{v}_\alpha(\eta) \succ 0$ by Lemma 4 in Appendix D.1.

The Monotonic Algorithm (MA). The MA is an iterative procedure of the form

$$\eta_{k+1} = \mathcal{M}_{\gamma_k}(\eta_k), \qquad k = 0, 1, 2, \ldots, \tag{9}$$

where, for all $\gamma \in [0,1]$, the mapping $\mathcal{M}_\gamma$ is defined by

$$\nabla A \circ \mathcal{M}_\gamma(\eta) = \gamma \mathbf{m}_\alpha(\eta) + (1-\gamma)\mathbf{m}_1(\eta). \tag{10}$$

The right-hand side in (10) is a convex combination between the current mean parameter $\mathbf{m}_1(\eta)$ and the mean of S under the (normalized) geometric average density $\pi_{\alpha,\eta}$ that involves the target p. The mapping $\mathcal{M}_\gamma$ drives the current mean parameter towards the mean of S under $\pi_{\alpha,\eta}$, hence the moment-matching

denomination. Additionally, note that γ can be interpreted as a step-size that controls how far the updated parameter is from the current parameter.

To complete the definition of the algorithm, there only remains to set the initial parameter $\eta_0 \in E$ and the sequence $(\gamma_k)_{k\in\mathbb{N}} \in (0,1]^{\mathbb{N}}$. However, there is a substantial constraint: for every $k \geq 0$, η_k must remain within the definition set of $\mathcal{M}_{\gamma_k}$. Indeed, the mapping $\mathcal{M}_\gamma$ may not be defined on the whole set E, since it may occur that $\mathbf{m}_\alpha(\eta)$ does not lie in F for some $\eta \in E$. Then, $\mathcal{M}_1(\eta)$ is not properly defined. That said, one can always find a sufficiently small positive γ such that $\mathcal{M}_\gamma(\eta)$ is well-defined, as explained in Appendix D.4. To study this constraint and choose the sequence $(\gamma_k)_{k\geq 0}$ accordingly, we let $\mathcal{D}_\gamma$ denote the definition set of $\mathcal{M}_\gamma$, that is,

$$\mathcal{D}_\gamma = \{\eta \in E \ : \ \gamma \mathbf{m}_\alpha(\eta) + (1-\gamma)\mathbf{m}_1(\eta) \in F\}\,. \quad (11)$$

3 DISCRETIZED ALGORITHM

In most cases, the function $\mathbf{m}_\alpha$ in (8) is intractable. To create a tangible instance of the MA, we now develop a discretized version that can be seen as an instantiation of the Sample Average Approximation (SAA) paradigm. This serves two purposes: first, it provides a practical, implementable algorithm. Second, it creates a setting that falls squarely within our main theoretical convergence result (Section 4). By introducing a carefully chosen variational family, we can shift to a discrete setting in which the expectation $\mathbf{m}_\alpha(\eta)$ becomes tractable.

Discretized family. Assume that the variational family $\mathcal{Q} = \{q_\eta : \eta \in E\}$ satisfies (A2), but the MA updates are intractable. Let $\eta_\diamond \in E$ and let $y_1, \ldots, y_n \sim q_{\eta_\diamond}$ be n independent samples with $n \geq d$. As recalled in Section 2.2, the family $\mathcal{Q}$ is entirely characterized by the inner-product space H, the sufficient statistic S, and the base measure ν. We propose to replace ν by the discrete measure

$$\widehat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \frac{\delta_{\{y_i\}}}{q_{\eta_\diamond}(y_i)}.$$

This defines a new exponential family $\widehat{\mathcal{Q}}_n$ whose log-partition function $\widehat{A}_n$ is finite for all $\eta \in \mathsf{H}$,

$$\widehat{A}_n(\eta) = \ln\left(\frac{1}{n}\sum_{i=1}^n \frac{\exp(\langle \eta, S(y_i)\rangle)}{q_{\eta_\diamond}(y_i)}\right).$$

By introducing this family, whose properties are explored in Appendix B.5, we have rendered our framework applicable to a case where intractable integrals used in the MA are evaluated by self-normalized importance sampling. Indeed, the objective in (1) and the associated geometric mean in (8), both defined with ν replaced by $\widehat{\nu}_n$, are rigorously equal to the importance sampling estimators

$$\widehat{\ell}_{\alpha,n}(\eta) = \frac{1}{n}\sum_{i=1}^n \exp\Big(\ln(w_i) - \alpha\,\widehat{A}_n(\eta)\Big)$$
$$\widehat{\mathbf{m}}_{\alpha,n}(\eta) = \frac{\sum_{i=1}^n S(y_i)\,w_i}{\sum_{i=1}^n w_i}, \quad (12)$$

where, for $i \in [\![1,n]\!]$, we define the log-weight $\ln(w_i)$ by $\ln(w_i) = \alpha\langle \eta, S(y_i)\rangle + (1-\alpha)\ln(p(y_i)) - \ln(q_{\eta_\diamond}(y_i))$.

SAA algorithm. To practically implement the MA, we present Algorithm 1 below. For the sake of clarity, this version is slightly simplified by taking $\gamma_k = 1$ for all $k \in \mathbb{N}$. A more comprehensive discussion can be found in Appendix E.

Algorithm 1 SAA procedure

Initialize $\eta_\diamond \sim \mathbb{P}$ and $\eta \leftarrow \eta_\diamond$
Draw i.i.d. samples $y_1, \ldots, y_n \sim q_{\eta_\diamond}$
for $k = 1$ **to** T **do**
  Compute $\mathbf{m} \leftarrow \widehat{\mathbf{m}}_{\alpha,n}(\eta)$ with (12)
  Solve $\eta_{\text{new}} = \arg\max_{\eta'}\Big[\langle \eta', \mathbf{m}\rangle - \widehat{A}_n(\eta')\Big]$
  Update $\eta \leftarrow \eta_{\text{new}}$
end for
return η

The computation of $\widehat{\mathbf{m}}_{\alpha,n}(\eta)$ in the main loop of Algorithm 1 is straightforward as long as η is known. Therefore, given η_k, we need a way to find η_{k+1} such that $\nabla\widehat{A}_n(\eta_{k+1}) = \widehat{\mathbf{m}}_{\alpha,n}(\eta_k)$ (this comes directly from (10) with $\gamma = 1$). Inverting $\nabla\widehat{A}_n$ analytically isn't possible in general, thus we exploit the conjugate dual of the log-partition $\widehat{A}_n$ and compute $\arg\max_{\eta'}\Big[\langle \eta', \mathbf{m}\rangle - \widehat{A}_n(\eta')\Big]$ using a numerical solver. If applicable, Newton's method in Algorithm 2 is a sensible choice to take advantage of the strict concavity of the objective (see Section 5.3 for a discussion regarding solvers).

Algorithm 2 Newton's method for η_{new}

Require: Target mean $\mathbf{m}$, initial parameter η.
Ensure: η_{new} such that $\nabla\widehat{A}_n(\eta_{\text{new}}) \simeq \mathbf{m}$.
Initialize $\eta' \leftarrow \eta$
while $\|\nabla\widehat{A}_n(\eta') - \mathbf{m}\| >$ tolerance **do**
  Compute $\mathbf{g} \leftarrow \nabla\widehat{A}_n(\eta') - \mathbf{m}$
  Compute $\mathbf{H} \leftarrow \mathbf{H}_{\widehat{A}_n}(\eta')$
  Update $\eta' \leftarrow \eta' - \mathbf{H}^{-1}\mathbf{g}$
end while
return $\eta_{\text{new}} \leftarrow \eta'$

We will see that the discretized algorithm converges exponentially fast, but to a minimizer of the approximate objective $\hat{\ell}_\alpha$, not the true objective ℓ_α. Point-wise, each $\hat{\ell}_\alpha(\eta)$ is an unbiased estimator of $\ell_\alpha(\eta)$.

4 CONVERGENCE GUARANTEES

Having introduced the SAA method, we now establish the convergence guarantees for the underlying deterministic MA. Our SAA method, by design, creates a discrete setting where the conditions for these guarantees can be met.

4.1 Assumptions

The MA was introduced by [Daudel et al., 2023] and its asymptotic behavior studied by [Bertholom et al., 2024] with the assumptions:

a) The iterates $(\eta_k)_{k\geq 0}$ remain in a compact set.

b) Critical points of the objective function in this compact are strict local maximizers.

c) The sequence $(\gamma_k)_{k\geq 0}$ stays away from 0.

The construction of the compact set in *a*) is not specified. Furthermore, any critical point in this hypothesized set must be a local maximizer of ℓ_α, due to *b*). Lastly, *c*) sounds reasonable given *a*) and *b*), because any critical point of ℓ_α belongs to the set $\mathcal{D}_\gamma$ defined in (11), and we can expect the loss landscape to be favorable once close to a minimizer. However, this remains an assumption and it is unclear why the algorithm could avoid diverging whith positive step-sizes.

We propose to replace this set of assumptions by two design guidelines, and a mild assumption about the function ℓ_α. The first design guidelines involves a probability $\mathbb{P}$ on H endowed with its Borel σ-field, $\mathcal{H}$.

(D1) We draw η_0 according to $\mathbb{P}$, which satisfies $\mathbb{P}(E) = 1$ and $\mathbb{P}(\mathsf{M}) = 0$ for every manifold M whose dimension is strictly less than d.

(D2) Let $\gamma_0 \in (0,1)$ and define, for all $\eta \in E$, $\beta(\eta) = \max\left\{\gamma_0^n \ : \ \eta \in \mathcal{D}_{\gamma_0^n},\, n \in \mathbb{N}\right\}$. The step-size sequence $(\gamma_k)_{k\geq 0}$ is constructed according to one of the following two definitions:

- (*i*) At each step $k \geq 0$, set $\gamma_k = \gamma_0$.
- (*ii*) Set $\gamma_0 := \beta(\eta_0)$ and at each step $k \geq 1$, set $\gamma_k = \gamma_{k-1} \wedge \beta(\eta_k)$.

In (D2), the first option is of course the simplest. The second scheme is guaranteed to yield the optimal asymptotic rate of convergence. This will become clear with the statement of our main theorem.

We now state our only assumption about ℓ_α.

(A3) The Hessian matrix of ℓ_α is non-degenerate at critical points of ℓ_α, i.e., ℓ_α is a Morse function.

For any $\gamma \in (0,1]$, the set of fixed points of the iterative mapping $\mathcal{M}_\gamma$ can be characterized from (10). In fact, this set does not depend on γ. [Bertholom et al., 2024] further showed that these fixed points were exactly the critical points of ℓ_α:

$$\begin{aligned}\mathrm{Fix}(\mathcal{M}) &= \{\eta \in E \ : \ \mathbf{m}_\alpha(\eta) = \mathbf{m}_1(\eta)\} \\ &= \{\eta \in E \ : \ \mathcal{M}_\gamma(\eta) = \eta\}\,, \quad \gamma \in (0,1].\end{aligned}$$

Indeed, it not hard to show that the gradient of ℓ_α is $\nabla \ell_\alpha(\eta) = \alpha\, \ell_\alpha(\eta)(\mathbf{m}_\alpha(\eta) - \mathbf{m}_1(\eta))$.

We let $\mathrm{Sp}(\cdot)$ denote the spectrum. From the discussion in Appendix C, (A3) is equivalent to any of the following two assertions:

(*i*) For all $\eta \in \mathrm{Fix}(\mathcal{M})$, the Jacobian matrix $\mathbf{J}_{\mathcal{M}_\gamma}(\eta)$ does not have 1 among its eigenvalues.

(*ii*) For all $\eta \in \mathrm{Fix}(\mathcal{M})$, $1 \notin \mathrm{Sp}\big(\alpha\, \mathbf{v}_1(\eta)^{-1}\, \mathbf{v}_\alpha(\eta)\big)$.

It is noteworthy that Assumption (A3) holds in well-specified models. Indeed, if p belongs to the regular minimal exponential family $\mathcal{Q}$, the means are uniquely defined and there is a single point $\eta_\star$ such that $\mathbf{m}_\alpha(\eta_\star) = \mathbf{m}_1(\eta_\star)$. It is the only maximizer of ℓ_α.

In general, for Assumption (A3) to not hold, the covariance matrices $\alpha \mathbf{v}_1(\eta_\star)$ and $\mathbf{v}_\alpha(\eta_\star)$ must align in a very specific way (see Appendix C). This is extremely unlikely when p stems from a complicated, real-world distribution.

To conclude this section, we introduce a constant that depends on the family $\mathcal{Q}$ and the target p:

$$\Pi_{\mathcal{Q},p} = \sup_{\substack{u\in\mathbb{U}\\ M\in\mathbb{R}}} \mu_1(\{y \in \mathsf{Y} \ : \ \langle u, S(y)\rangle = M\})\,, \qquad (13)$$

where μ_1 is defined in (3). By (2), ν and μ_1 are equivalent measures, hence (A2) is equivalent to having $\Pi_{\mathcal{Q},p} = 0$.

4.2 Main theorem

We now state our main theorem, which guarantees the convergence of (9) to a local maximizer of the function ℓ_α at an asymptotically geometric rate. The rate will be expressed through the following quantity:

$$\rho(\eta) = \alpha\, \varrho\big(\mathbf{v}_1(\eta)^{-1}\, \mathbf{v}_\alpha(\eta)\big)\,, \qquad (14)$$

where $\varrho(\cdot)$ denote the spectral radius application.

Theorem 1. *Let Assumptions* (A1) *and* (A3) *hold and suppose* (D1) *with*

$$\mathbb{P}\Big(\Big\{\ell_\alpha > \Pi_{\mathcal{Q},p}^{1-\alpha}\Big\}\Big) = 1. \tag{15}$$

For all $k \geq 0$, compute γ_k as in (D2)*, and update $\eta_{k+1} = \mathcal{M}_{\gamma_k}(\eta_k)$. Then, $(\eta_k)_{k\geq 0}$ converges almost surely to a parameter $\eta_\star \in \mathrm{Fix}(\mathcal{M})$ which is a local maximizer of the function ℓ_α. Additionally, the following assertions hold with probability* 1*.*

(i) $\rho(\eta_\star) < 1$.

(ii) There exists $T \in \mathbb{N}$ such that, for all $k \geq T$, $\gamma_k = \gamma_T \in (0, 1]$.

(iii) For any $\rho \in (\gamma_T\, \rho(\eta_\star) + (1 - \gamma_T), 1)$, we have $\|\eta_k - \eta_\star\| = \mathcal{O}(\rho^k)$ as $k \to +\infty$.

(iv) Under (D2)*-(i), we have $\gamma_T = 1$.*

We have the following important remarks about (15):

- Under Assumption (A2), we have $\Pi_{\mathcal{Q},p} = 0$, so by (4), the condition in (15) always holds as a consequence of $\mathbb{P}(E) = 1$ in (D1).
- Assume the existence of a probability $\mathbb{P}'$ such that $\mathbb{P}'\Big(\Big\{\ell_\alpha > \Pi_{\mathcal{Q},p}^{1-\alpha}\Big\}\Big) > 0$. If we define $\mathbb{P}$ by conditionning $\mathbb{P}'$ on the event $\Big\{\ell_\alpha > \Pi_{\mathcal{Q},p}^{1-\alpha}\Big\}$, then (15) holds. This conditioning also ensures that the properties required in (D1) are inherited by $\mathbb{P}$ from $\mathbb{P}'$. Furthermore, one can simulate from $\mathbb{P}$ using a simple rejection algorithm, provided independent draws can be generated from $\mathbb{P}'$.

4.3 Proof outline and key insights

Key arguments in our proof rely on the monotonicity of the algorithm: each iteration improves the objective function. Precisely, the following theorem is established in [Daudel et al., 2023, Corollary 1].

Theorem 2. *For all $\alpha \in (0, 1)$, $\gamma \in (0, 1]$, and $\eta \in \mathcal{D}_\gamma$, we have $\ell_\alpha \circ \mathcal{M}_\gamma(\eta) \geq \ell_\alpha(\eta)$ with equality if and only if $\eta = \mathcal{M}_\gamma(\eta)$.*

This property is pivotal in the proof of [Bertholom et al., 2024, Theorem 1], from which we derive the ensuing result.

Theorem 3. *Let $\underline{\gamma} \in (0, 1]$ and let $\mathsf{K} \subset E$ be a compact set. Suppose that $(\eta_k) \in \mathsf{K}^{\mathbb{N}}$ and $(\gamma_k) \in [\underline{\gamma}, 1]^{\mathbb{N}}$ satisfy* (9)*. Then, the following three assertions hold.*

(i) All the limit points of (η_k) belong to $\mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}$.

(ii) Moreover, if $\mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}$ is a finite set, then (η_k) converges to some $\eta_\star \in \mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}$.

(iii) Further assume that $\eta_\star$ is a strict local maximizer of ℓ_α and let $\gamma_\star := \liminf_{k\to\infty} \gamma_k$. Then, for all $\rho \in (\gamma_\star\, \rho(\eta_\star) + (1 - \gamma_\star), 1)$, the asymptotic rate of convergence is $\|\eta_k - \eta_\star\| = \mathcal{O}(\rho^k)$ as $k \to +\infty$.

The proof of Theorem 3 can be found in Appendix D.6, where we also complete the steps that lead from Theorem 3 to Theorem 1. While Theorem 3 builds upon results from [Bertholom et al., 2024], the leap to Theorem 1 requires addressing two particularly challenging questions. They reflect the key structural hurdles of the problem, and encapsulate the main theoretical difficulties.

(Q1) Does the sequence of iterates produced by the algorithm remain in a compact set K as assumed in Theorem 3?

(Q2) Why does random initialization allow the algorithm to almost surely avoid local minima or saddle points of the objective ℓ_α, as assumed in Assertion *(iii)* of Theorem 3?

Question (Q2) is solved by taking advantage of the properties of the MA under (A3) and applying the stable manifold theorem, a key result from dynamical systems theory. The core idea behind this reasoning is borrowed from [Lee et al., 2016], with details specific to our study (where, in contrast to [Lee et al., 2016], the step-size is possibly non-constant) available in Appendix D.5.

We now proceed to explain the development that allowed us to solve Question (Q1). To obtain a compact set K as in Theorem 3, we again rely on the monotonicity property granted by Theorem 2. Indeed, if we define for any $\varkappa > 0$ the level set

$$\mathsf{K}_\varkappa = \{\eta \in E \;:\; \ell_\alpha(\eta) \geq \varkappa\}, \tag{16}$$

then by Theorem 2, the sequence $(\eta_k)_{k\geq 0}$ remains in $\mathsf{K}_{\ell_\alpha(\eta_0)}$. The set $\mathsf{K}_{\ell_\alpha(\eta_0)}$ being closed, it is a plausible candidate to be the compact set K required by Theorem 3. However, all level sets are not necessarily bounded. This is visually illustrated by the well-specified Bernoulli model developed in Appendix B.2.

The following theorem gathers the results of Lemmas 7 and 8, found in Appendix D.2. It investigates the behavior of the objective ℓ_α as η moves away from the compact sets included in E.

Theorem 4. *Recall that $\Pi_{\mathcal{Q},p}$ is the constant in* (13)*.*

(i) For all $\overline{\eta} \in \mathrm{bd}(E)$, $\lim_{\eta\to\overline{\eta}} \ell_\alpha(\eta) = 0$.

(*ii*) (A1) *implies that* $\limsup_{\substack{\|\eta\|\to+\infty \\ \eta\in E}} \ell_\alpha(\eta) \le \Pi_{\mathcal{Q},p}^{1-\alpha}$.

In the theorem, $\mathrm{bd}(E)$ denotes the boundary of E. Combining Theorem 4 with the continuity of ℓ_α on E ensures the ν-almost sure compactness of $\mathsf{K}_{\ell_\alpha(\eta_0)}$ under the condition in (15).

Having shown that the iterates remain in a compact subset of E, we can design step-size sequences that do not vanish, as assumed in Theorem 3. The rationale behind (D2) is motivated by the following key insight: the set F of the means that are achievable by the family $\mathcal{Q}$ equals the interior of the convex set

$$\tilde{F} := \left\{ \mathbb{E}_P[S] \ : \ P \ll \nu \text{ and } \int \|S\| \, \mathrm{d}P < \infty \right\}.$$

For more details, see Appendix D.4.

Finally, we show that under (A3), the fixed points are isolated (see Corollary 1 in Appendix D.3). An essential consequence of this fact is that $\mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}$ is a finite set as assumed in Theorem 3-(*ii*).

4.4 Condition (15) for the discretized version

The condition in (15) only needs to be checked when we are working with a discrete family, since it is automatically verified when Assumption (A2) holds. While the constant $\Pi_{\mathcal{Q},p}$ may not always be easy to compute, we show that it has a simple form in the case of the discretized family. Let $\eta_\diamond \in E$ and let $y_1, \dots, y_n, \dots \sim q_{\eta_\diamond}$ be n independent samples on the same probability space $(\Omega, \mathcal{A}, \mathbb{Q})$. We consider the exponential family $\widehat{\mathcal{Q}}_n$ defined by the triple $(\mathsf{H}, S, \widehat{\nu}_n)$, as in Section 3.

Proposition 1. *Let* (i) *denote the index of the* i*-th order statistic for the* n*-sample* $(p(y_i)/q_{\eta_\diamond}(y_i))_{1\le i\le n}$*, i.e., such that* $p(y_{(1)})/q_{\eta_\diamond}(y_{(1)}) \le \dots \le p(y_{(n)})/q_{\eta_\diamond}(y_{(n)})$*. If the original family* $\mathcal{Q}$ *satisfies* (A2)*, then the following three assertions hold.*

(*i*) $\widehat{\mathcal{Q}}_n$ *defines a regular exponential model.*

(*ii*) *If* $n \ge d+1$*, then* $\widehat{\mathcal{Q}}_n$ *is* $\mathbb{Q}$*-almost surely minimal.*

(*iii*) *The constant* $\Pi_{\widehat{\mathcal{Q}}_n,p}$ *has the expression*

$$\Pi_{\widehat{\mathcal{Q}}_n,p} = \frac{1}{n} \sum_{i=n-d+1}^{n} \frac{p(y_{(i)})}{q_{\eta_\diamond}(y_{(i)})}. \tag{17}$$

From this expression, it is not hard to find conditions that ensure that $\Pi_{\widehat{\mathcal{Q}}_n,p}$ converges $\mathbb{Q}$-almost surely to zero as n goes to infinity, meaning that (15) is more likely to be satisfied with a larger n, for a given $\eta \in E$. More details are provided in Appendix B.5.

5 EXPERIMENTS

5.1 Toy example with von-Mises Fisher distributions

The von Mises-Fisher (vMF) distribution (see Appendix B.4) has recently gained a lot of attention in the machine learning community for representing spherical data [Gopal and Yang, 2014, Huynh et al., 2024, Taghia et al., 2013]. It structurally inherits various desirable properties that make it compelling to VI practictioners, as exemplified by Hyperspherical Variational Auto-Encoders [Davidson et al., 2018]. Despite its appeal, significant limitations arise from the intractable normalization constant of the vMF density, and solving related issues is an active line of research [Nonnenmacher and Sahani, 2024, You et al., 2025]. Our SAA method advantageously bypasses this intractability. It operates directly with the natural parameter $\eta \in \mathbb{R}^d$ and samples drawn from some $q_{\eta_\diamond}$, thereby avoiding the need to compute the normalization constant or the related mean parameter.

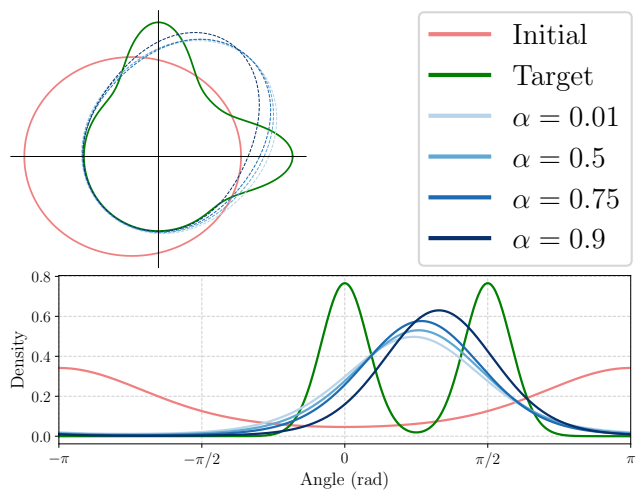

Figure 1: Approximation of a bimodal von Mises-Fisher mixture (green) using the SAA method, starting from a near-uniform initial distribution (red). Right: probability densities plotted against the angle in radians. Left: corresponding visualization of the distributions on the circle (arbitrary scale).

5.2 Bayesian Logistic Regression

We evaluate our method on the California Housing dataset (20 640 samples, 9 features) using an 80%-20% train-test split. A Bayesian Logistic Regression model classifies house prices relative to the training set median, with features standardized and full-batch optimization. We implement Algorithm 1 with Newton's method (Algorithm 2), including backtracking line search for

enhanced stability. By choosing multivariate Gaussian densities with diagonal covariance matrices as the variational family, we optimize over an 18-dimensional parameter space. We set $\alpha = 0.5$, $n = 50\,000$, and randomly initialize the mean and coefficients of the covariance matrix diagonal for each of the 100 runs we performed. Figure 2 shows boxplots from these 100 runs, illustrating the evolution of $\hat{\ell}_{\alpha,n}$ (in blue, primary y-axis), train accuracy, and test accuracy (respectively in red and green, secondary y-axis) against iterations, including the inner optimization loop steps. In terms of task-related performance, the method allows to consistently reach decent levels of accuracy, with best-case scenarios reaching over 80% accuracy.

Ultimately, the confidence in our method's output depends entirely on how close the approximate loss is to ℓ_α. Indeed, the SAA framework trades computational complexity for precision. As it only approximates the original problem, it can only yield an approximation of the desired solution. However, it is much faster at finding this solution. The quality of the approximation is determined by two factors: the sample size and the sample quality: good samples cover the modes of p, while also covering low-mass regions to capture its complexity. These choices are discussed in Appendix E.1.

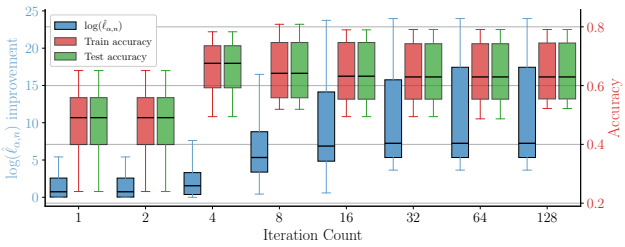

Figure 2: SAA algorithm with Newton (Algorithms 1 and 2). Iteration count includes Newton steps.

We also compare our SAA scheme to existing methods. We generate a single synthetic dataset of 3000 datapoints, each consisting of 100 features and one binary response. We consider a Bayesian Logistic Regression model with a fully-factorized Gaussian variational family and optimize it in four different ways: SA-AW is AdamW on $\log \ell_\alpha$; SA-MA is the MA with an SA scheme; SAA-N is our SAA algorithm with Newton's method and SAA-A is our SAA method using Adam. Each method has a time budget of 10 seconds to perform as many runs as possible, with a new random initialization for each run. Once the budget is exhausted, the best run is retained. This entire experiment is repeated 10 times. The results are summarized in Table 1.

SA-AW serves as a consistent baseline. While SA-MA can achieve superior performance (the absolute best test accuracy reported was over 92% with SA-MA, compared to 89% with SA-AW), it appears highly sensitive to initial conditions and has high variance. In the worst case scenario, for SA-MA, 36 runs were not enough to find a proper initialization. In contrast, both SAA algorithms are remarkably fast, always completing over 200 runs within the time budget. This allows SAA-N and SAA-A to achieve test accuracies comparable to SA-AW in potentially less time, and with considerably fewer samples. This also suggests a powerful synergy: SAA's rapid exploration can find promising initializations, which are refined using SA-MA to achieve superior performance. This is helpful to solve the difficult but necessary problem of exploring the parameter space in multi-modal settings, where the SA version of the algorithm is quite sensitive to initialization. Because of its faster convergence, we can afford to run multiple SAA trajectories from random initializations to identify promising points that can be improved by a more precise method.

Due to its deterministic nature and rapid early-stage improvement, running the SAA for a small number of iterations is an efficient strategy for identifying a strong starting point, thereby mitigating the initialization-sensitivity of other methods.

Table 1: Performance and computational cost comparison on Bayesian Logistic Regression (synthetic dataset, 3000 data points, 100 features and $\alpha = 0.5$).

	Computational Cost		
Alg.	Avg. Runs	Avg. Samples	
SA-AW	47.6 ± 1.93	7×10^5	
SA-MA	36.0 ± 2.4	8×10^4	
SAA-N	227.4 ± 0.5	2000	
SAA-A	223.6 ± 0.5	2000	
	Performance		
Alg.	Accuracy (%)	Log-likelihood	$\|\hat{\mu} - \mu_\star\|_2^2$
SA-AW	85.13 ± 1.06	-0.36 ± 0.01	4.22 ± 0.44
SA-MA	71.11 ± 16.21	-0.58 ± 0.31	8.46 ± 5.00
SAA-N	85.50 ± 1.75	-0.36 ± 0.01	4.95 ± 0.30
SAA-A	84.17 ± 1.22	-0.37 ± 0.02	4.89 ± 0.28

5.3 Practical considerations

Certain computational challenges arise when applying the proposed methods in high-dimensional settings or with large datasets. Firstly, in high dimensions, the computation and inversion of the Hessian matrix $\mathbf{H}$ in Algorithm 2 can become prohibitively expensive. It may then be preferable to use first-order optimization methods to determine η_{new} in Algorithm 1. The tradeoff is that these methods need more iterations to converge than Newton's algorithm. Secondly, large datasets introduce intractability when comput-

ing quantities that involve the function p, i.e., $\widehat{\mathbf{m}}_\alpha$ in Algorithm 1, the gradient $\mathbf{g}$ and Hessian $\mathbf{H}$ in Algorithm 2. While mini-batching is a common technique to accelerate such computations, its integration into our framework requires careful consideration. Doing it in a practically viable and theoretically justified way is an interesting direction for future research.

As for the initialization, it is drawn from a distribution with support over the entire parameter space. While we assume (15) in the theoretical analysis, it only provides a sufficient condition for the algorithm to converge. In practice, it is often not satisfied, and empirical observations suggest that a less stringent condition could be formulated instead.

While our theoretical analysis assumes the ability to recover the natural parameter from the mean exactly, the SAA algorithm approximates it numerically by minimizing a tractable, strictly convex function, with tractable first and second derivatives. Extending the analysis to explicitly account for this approximation would require controlling the perturbation induced by the numerical solver on the mapping $\mathcal{M}_\gamma$, hence a general result seems hard to obtain. Moreover, our suggested choice of the step-size schedule would need to be redefined to accomodate the solver's properties. The primary technical difficulty to address this question would certainly be to prove that the algorithm converges to a neighborhood of a fixed point, where $\mathcal{M}_\gamma$ is a contraction. A more comprehensive analysis of the dynamics at play is left for future work.

References

[Alquier and Ridgway, 2020] Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497.

[Amari, 2016] Amari, S. (2016). *Information geometry and its applications*, volume 194. Springer.

[Bertholom et al., 2024] Bertholom, F., Douc, R., and Roueff, F. (2024). Asymptotics of alpha-divergence variational inference algorithms with exponential families. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

[Burroni et al., 2024] Burroni, J., Domke, J., and Sheldon, D. (2024). Sample average approximation for black-box variational inference. In *The 40th Conference on Uncertainty in Artificial Intelligence*.

[Daudel and Douc, 2021] Daudel, K. and Douc, R. (2021). Mixture weights optimisation for alpha-divergence variational inference. *Advances in Neural Information Processing Systems*, 34:4397–4408.

[Daudel et al., 2023] Daudel, K., Douc, R., and Roueff, F. (2023). Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76.

[Davidson et al., 2018] Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. (2018). Hyperspherical Variational Auto-encoders. *arXiv preprint arXiv:1804.00891*.

[Dieng et al., 2017] Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. (2017). Variational inference via chi-upper bound minimization. *Advances in Neural Information Processing Systems*, 30.

[Dieuleveut et al., 2023] Dieuleveut, A., Fort, G., Moulines, E., and Wai, H.-T. (2023). Stochastic approximation beyond gradient for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 71:3117–3148.

[Geffner and Domke, 2020] Geffner, T. and Domke, J. (2020). On the difficulty of unbiased alpha divergence minimization. *arXiv preprint arXiv:2010.09541*.

[Geffner and Domke, 2021] Geffner, T. and Domke, J. (2021). Empirical evaluation of biased methods for alpha divergence minimization. *arXiv preprint arXiv:2105.06587*.

[Giordano et al., 2024] Giordano, R., Ingram, M., and Broderick, T. (2024). Black-box variational inference with a deterministic objective: Faster, more accurate, and even more black box. *Journal of Machine Learning Research*, 25(18):1–39.

[Gopal and Yang, 2014] Gopal, S. and Yang, Y. (2014). von Mises-Fisher Clustering Models. In *International Conference on Machine Learning*, pages 154–162. PMLR.

[Guilmeau et al., 2025] Guilmeau, T., Chouzenoux, E., and Elvira, V. (2025). Regularized rényi divergence minimization through bregman proximal gradient algorithms. *Journal of Machine Learning Research*, 26(157):1–56.

[Hernandez-Lobato et al., 2016] Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. (2016). Black-box alpha divergence minimization. In *International conference on machine learning*, pages 1511–1520. PMLR.

[Huynh et al., 2024] Huynh, T., An, G., Kim, M., Jeon, Y.-S., and Lee, J. (2024). A variational quantum algorithm by Bayesian Inference with von Mises-Fisher distribution. *arXiv preprint arXiv:2410.03130*.

[Kim et al., 2014] Kim, S., Pasupathy, R., and Henderson, S. G. (2014). A guide to sample average approximation. *Handbook of simulation optimization*, pages 207–243.

[Lee et al., 2016] Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA. PMLR.

[Li and Turner, 2016] Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. *Advances in neural information processing systems*, 29.

[Margossian et al., 2025] Margossian, C. C., Pillaud-Vivien, L., and Saul, L. K. (2025). Variational inference for uncertainty quantification: An analysis of trade-offs. *Journal of Machine Learning Research*, 26(202):1–41.

[Margossian and Saul, 2023] Margossian, C. C. and Saul, L. K. (2023). The shrinkage-delinkage trade-off: An analysis of factorized gaussian approximations for variational inference. In *Uncertainty in Artificial Intelligence*, pages 1358–1367. PMLR.

[Masrani et al., 2019] Masrani, V., Le, T. A., and Wood, F. (2019). The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 32.

[Minka, 2005] Minka, T. (2005). Divergence measures and message passing.

[Nonnenmacher and Sahani, 2024] Nonnenmacher, M. and Sahani, M. (2024). A solution for the mean parametrization of the von Mises-Fisher distribution. *arXiv preprint arXiv:2404.07358*.

[Ranganath et al., 2014] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[Rockafellar, 1970] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton.

[Shub, 2013] Shub, M. (2013). *Global stability of dynamical systems*. Springer Science & Business Media.

[Taghia et al., 2013] Taghia, J., Ma, Z., and Leijon, A. (2013). On von-Mises Fisher Mixture Model in Text-Independent Speaker Identification. In *INTERSPEECH*, pages 2499–2503.

[Wainwright and Jordan, 2008] Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305.

[Wang et al., 2018] Wang, D., Liu, H., and Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. *Advances in Neural Information Processing Systems*, 31.

[Yang et al., 2020] Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905.

[You et al., 2025] You, K., Shung, D., and Giuffrè, M. (2025). Learning over von Mises-Fisher Distributions via a Wasserstein-like Geometry. *arXiv preprint arXiv:2504.14164*.

[Zimmermann et al., 2024] Zimmermann, H., Naesseth, C. A., and van de Meent, J.-W. (2024). VISA: Variational Inference with Sequential Sample-Average Approximations. *arXiv preprint arXiv:2403.09429*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] Complete proofs are provided in the appendix.
   (c) Clear explanations of any assumptions. [Yes] The assumptions are discussed in the main text.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes] We use the California housing dataset.

   (b) The license information of the assets, if applicable. [No] We release the assets under the same license as the paper.

   (c) New assets either in the supplemental material or as a URL, if applicable. [No] We release the assets under the same license as the paper.

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [No] We do not foresee any potential misuse of our work in the near future.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable] We do not use crowdsourcing.

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Sample Average Approximation for Alpha-Divergence Variational Inference with Exponential Convergence Guarantees: Supplementary Materials

A SOME TOPOLOGICAL AND MEASURE-THEORETIC DEFINITIONS

We say that a topological space is:

(*i*) *separable* if it contains a countable dense subset;

(*ii*) *Hausdorff* if distinct points can always be separated by disjoint neighborhoods;

(*iii*) *locally compact* if each point has a compact neighborhood;

(*iv*) *σ-compact* if it can be covered by a countable sequence of compact subsets.

If $\mathcal{O}$ denotes the topology of Y, the *support* of a measure ν is given by:

$$\mathrm{Supp}(\nu) = \left(\bigcup \{U \in \mathcal{O} : \nu(U) = 0\}\right)^c.$$

When the space Y is separable, we have $\nu(\mathrm{Supp}(\nu)^c) = 0$.

B EXAMPLES OF COMMON EXPONENTIAL FAMILIES

In the following examples of exponential models, we specify the sufficient statistic S, the natural parameter η and the dominating measure ν.

B.1 Multivariate Gaussian

Set $\mathsf{H} = \mathbb{R}^d \times \mathbb{R}^{d\times d}$ and consider the inner product defined for the vectors a_1, a_2 and matrices A_1, A_2 by $\langle (a_1, A_1), (a_2, A_2)\rangle = a_1^\top a_2 + \mathrm{tr}(A_1^\top A_2)$, where tr denotes the trace operator. We obtain a multivariate Gaussian density with mean μ and invertible covariance matrix Σ by taking ν as the Lebesgue measure on $\mathbb{R}^d$, $S : y \mapsto (y, yy^\top)$ and $\eta = (\Sigma^{-1}\mu, -\Sigma^{-1}/2) \in E := \mathbb{R}^d \times \mathcal{S}_d^{<0}$, where $\mathcal{S}_d^{<0}$ denotes the set of negative-definite matrices. This defines a regular exponential family, and Assumption (A1) is trivially verified. To prove minimality, one can use Lemma 2, which also establishes the stronger condition (A2). Specifically, letting $\eta \in E$ and $M \in \mathbb{R}$, and applying the result of Lemma 2 to the continuously differentiable function $y \mapsto \langle S(y), \eta\rangle - M$ yields the desired outcome.

B.2 Bernoulli

Let $\mathsf{H} = \mathbb{R}$, $(\mathsf{Y}, \mathcal{Y}) = (\{0, 1\}, \mathcal{P}(\{0, 1\}))$ and let ν be the counting measure restricted to $\{0, 1\}$. The Bernoulli distribution with parameter $\theta \in (0, 1)$ admits a density with respect to ν, obtained by letting $\eta = \ln(\theta/(1-\theta))$ and $S : y \mapsto y$ in (5). This implies $E = \mathbb{R}$, which means that the family is regular. In this model, (A1) is verified (as a consequence of $0 < \theta < 1$), but not (A2). Indeed, we have $E = \mathbb{R}$ and $\mathbb{U} = \{-1, 1\}$. Notice that $\nu(\{S = 1\}) = \nu(\{S = 0\}) = 1$, thus (A2) does not hold.

We now study the well-specified Bernoulli model to develop a better understanding of the behavior of ℓ_α when $\|\eta\| \to +\infty$ in discrete models. Let $\theta_\star \in (0, 1)$ be an unknown parameter and consider a target $p\colon y \mapsto \theta_\star^y(1-\theta_\star)^{1-y}$, approximated by a Bernoulli variational family. For all $\eta \in \mathbb{R}$, we let $\theta_\eta = e^\eta/(1 + e^\eta)$. The objective ℓ_α admits the following closed form:

$$\ell_\alpha(\eta) = \theta_\eta^\alpha \theta_\star^{1-\alpha} + (1-\theta_\eta)^\alpha (1-\theta_\star)^{1-\alpha}.$$

Figure 3 shows ℓ_α for different values of α. We also plot $\Pi_{\mathcal{Q},p}^{1-\alpha} = \max(\theta_\star, 1-\theta_\star)^{1-\alpha}$. Notice that we have, for instance, $\mathsf{K}_{0.7} \supset \mathbb{R}_+$ in all three cases shown in Figure 3, hence this is one example of a case where the sets $(\mathsf{K}_\varkappa)_{\kappa>0}$ defined in (24) are not always compact. This motivates the study leading to Theorem 4-(ii), which is perhaps the most enlightening result of this work regarding the behavior of alpha-divergence variational inference, while also being the most technically challenging.

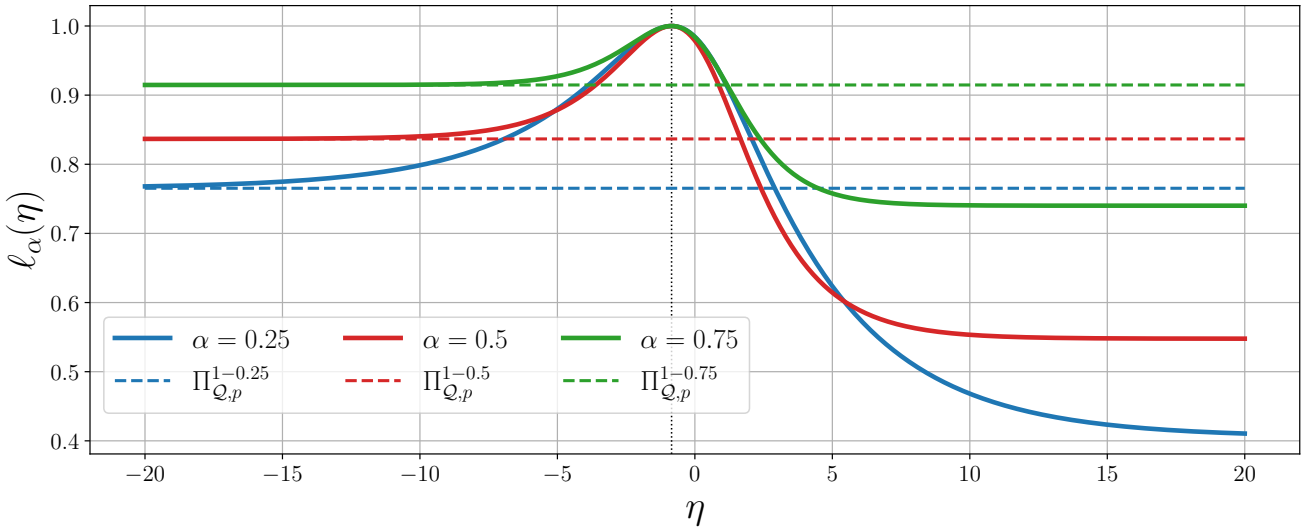

Figure 3: Objective $\ell_\alpha(\eta)$ and thresholds $\Pi_{\mathcal{Q},p}^{1-\alpha}$ for various values of α in the well-specified Bernoulli model. Here, $\theta_\star = 0.3$, corresponding to $\eta_\star = \ln(0.3/0.7)$. The true parameter $\eta_\star$ is indicated by a vertical black dotted line intersecting the x-axis at $\eta_\star$.

B.3 Ising Model

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, that is, $\mathcal{E}$ is a subset of all possible pairs of elements of the finite set $\mathcal{V}$. Set $d = \mathrm{Card}(\mathcal{V}) + \mathrm{Card}(\mathcal{E})$. The Ising model defines a family of probability distributions where each vertex $i \in \mathcal{V}$ is associated to a random variable valued in $\{-1, 1\}$. It is constructed by letting ν be the counting measure on $\{-1,1\}^{\mathrm{Card}(\mathcal{V})}$, S the function defined on this space by $S : y \mapsto ((y_i)_{i\in\mathcal{V}}, (y_i y_j)_{\{i,j\}\in\mathcal{E}})$, and $E = \mathsf{H} := \mathbb{R}^d$ endowed with the Euclidean inner product. Like in the Bernoulli example, the family is minimal but doesn't verify the stronger condition (A2).

B.4 von-Mises Fisher

Let ν denote the uniform probability measure on the unit sphere. The von Mises-Fisher distribution admits a density f with respect to ν, characterized by $f(y;\eta) \propto \exp(\langle \eta, y\rangle)$, where $\eta \in \mathbb{R}^d$. Clearly, this defines a regular exponential family with a continuous sufficient statistic S. To prove its minimality, as well as (A2), we need to show that the intersection of a hyperplane with the unit sphere in $\mathbb{R}^d$ has ν-measure zero. Intuitively, this happens because the intersection is a $(d-2)$-dimensional manifold (e.g., in $\mathbb{R}^3$, the intersection is a circle), and ν assigns zero to it.

Formally, let $M \in \mathbb{R}$ and consider the hyperplane $\{y \in \mathbb{R}^d \ : \ \langle y, \eta\rangle = M\}$. For an appropriate choice of a rotation σ and a translation τ, we have

$$\tau \circ \sigma(\{y \in \mathbb{R}^d \ : \ \langle y, \eta\rangle = M\}) = \{y = (y_1, \dots, y_d) \in \mathbb{R}^d \ : \ y_d = c\}.$$

Without loss of generality, we consider the hyperplane $\mathcal{H}_c = \{y = (y_1, \dots, y_d) \in \mathbb{R}^d \ : \ y_d = c\}$ and its intersection with the unit sphere. The unit hypersphere $\mathbb{S}^{d-1} = \{y \in \mathbb{R}^d \ : \ \|y\|^2 = 1\}$ corresponds to the set of vectors $y = (y_1, \dots, y_d) \in \mathbb{R}^d$ that satisfy $y_1^2 + \cdots + y_d^2 = 1$. Thus, we have $\mathbb{S}^{d-1} \cap \mathcal{H}_c = \{y = (y_1, \dots, y_d) \in \mathbb{R}^d \ : \ y_1^2 + \cdots + y_{d-1}^2 = 1 - c^2\}$. Now, there are three cases:

- if $1 - c^2 < 0$, there is no real solution and the intersection is empty;

- if $1 - c^2 = 0$, the hyperplane is tangent to the unit hypersphere and the intersection is a single point;
- if $1 - c^2 > 0$, the intersection describes a sphere of radius $\sqrt{1 - c^2}$ and centered at the origin in $\mathbb{R}^{d-1}$.

The measure ν is defined as the normalized $(d-1)$-dimensional Hausdorff measure on the metric space $(\mathbb{S}^{d-1}, r_{d-1})$, where r_{d-1} is the geodesic distance on $\mathbb{R}^{d-1}$. Since $\mathbb{S}^{d-1} \cap \mathcal{H}_c$ is at most a $(d-2)$-dimensional smooth manifold, the properties of Hausdorff measures imply $\nu(\mathbb{S}^{d-1} \cap \mathcal{H}_c) = 0$.

B.5 Discretized family

We now study the discretized family introduced in Section 3. Recall that $\mathcal{Q} = \{q_\eta \ : \ \eta \in E\}$ denotes an exponential family that satisfies (A2). Let $\eta_\diamond \in E$ and let $y_1, \dots, y_n \sim q_{\eta_\diamond}$ be n independent samples on the same probability space $(\Omega, \mathcal{A}, \mathbb{Q})$. We consider the exponential family $\widehat{\mathcal{Q}}_n$ defined by the triple $(\mathsf{H}, S, \widehat{\nu}_n)$, where $\widehat{\nu}_n$ is the discrete measure defined by

$$\widehat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{\{y_i\}}}{q_{\eta_\diamond}(y_i)}.$$

Proof of Proposition 1. (i) The log-partition function $\widehat{A}_n$ is finite for all $\eta \in \mathsf{H}$, having the expression

$$\widehat{A}_n(\eta) = \ln\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\exp(\langle \eta, S(y_i) \rangle)}{q_{\eta_\diamond}(y_i)} \right).$$

(ii) Let us start by noting that since the base family $\mathcal{Q}$ satisfies Assumption (A2), the random vector S $\mathbb{Q}$-almost surely will not fall into any proper (i.e., non-empty and not equal to H) affine subspace $\{u \in \mathsf{H} \ : \ \langle u, v \rangle = M\}$.

We now prove by induction that for $i \in [\![1, d+1]\!]$, the points $S(y_1), \dots, S(y_i)$ are $\mathbb{Q}$-almost surely affinely independent, i.e., the affine hull $\mathsf{A}_i := \mathrm{aff}((S(y_1), \dots, S(y_i)))$ has dimension $i - 1$ $\mathbb{Q}$-almost surely. The statement holds for the base case $i = 1$, since $S(y_1)$ is a single vector in H. For the induction step, assume that the statement holds for some $i \in [\![1, d]\!]$. Then, the affine hull A_i has dimension $i - 1 < d$, which means that A_i is $\mathbb{Q}$-almost surely a proper affine subspace of H. Hence, under Assumption (A2), the vector $S(y_{i+1})$ $\mathbb{Q}$-almost surely does not lie in A_i. As we took $i \leq d$, this implies that the points $S(y_1), \dots, S(y_{i+1})$ are affinely independent.

Since the $d+1$ points $S(y_1), \dots, S(y_{d+1})$ are $\mathbb{Q}$-almost surely affinely independent, their affine hull A_{d+1} is a d-dimensional affine subspace of H, i.e., H itself. A fortiori, the affine hull A_n equals H $\mathbb{Q}$-almost surely. In other words, the points $S(y_1), \dots, S(y_n)$ $\mathbb{Q}$-almost surely cannot all lie in a single affine hyperplane of H, that is, there exist no pair $(v, M) \in (\mathsf{H} \setminus \{0\}) \times \mathbb{R}$ such that $S(y_i) \in \{u \in \mathsf{H} \ : \ \langle u, v \rangle = M\}$ for all $i \in [\![1, n]\!]$ $\mathbb{Q}$-almost surely. This is equivalent to saying that $\widehat{\nu}_n(\{\langle v, S \rangle = M\}^c) > 0$, i.e., the family $\widehat{\mathcal{Q}}_n$ is $\mathbb{Q}$-almost surely minimal.

(iii) In the case of the family $\widehat{\mathcal{Q}}_n$, since the parameter space is H, the set $\mathbb{U}$ defined in (7) is the unit sphere of H. Finding the pair $(u, M) \in \mathbb{U} \times \mathbb{R}$ that realizes the supremum in (13) means designing a hyperplane to maximize

$$\widehat{\mu}_{1,n}(\{y \in \mathsf{Y} \ : \ \langle u, S(y) \rangle = M\}) = \frac{1}{n} \sum_{i=1}^{n} \frac{p(y_i)}{q_{\eta_\diamond}(y_i)} \mathbb{1}(\langle u, S(y_i) \rangle = M).$$

By a similar argument to (ii), one can show that there are at most d positive terms in this sum. Indeed, for a subset of indices $J \subset [\![1, n]\!]$ such that $\mathrm{Card}(J) \leq d+1$, the affine hull $\mathsf{A}_J = \mathrm{aff}((S(y_j))_{j \in J})$ has dimension $\mathrm{Card}(J) - 1$ $\mathbb{Q}$-almost surely. In particular, if $\mathrm{Card}(J) = d$, then A_J is an affine hyperplane of H. Then, by (A2), any point in $(S(y_i))_{i \notin J}$ belongs to A_J with $\mathbb{Q}$-probability 0.

Therefore, computing $\Pi_{\widehat{\mathcal{Q}}_n, p}$ reduces to selecting the d indices $(n-d+1), \dots, (n)$ that represent the d highest-order statistics for the sample $(p(y_i)/q_{\eta_\diamond}(y_i))_{1 \leq i \leq n}$. Finally, we obtain the expression

$$\Pi_{\widehat{\mathcal{Q}}_n, p} = \frac{1}{n} \sum_{i=n-d+1}^{n} \frac{p(y_{(i)})}{q_{\eta_\diamond}(y_{(i)})}.$$

□

Note that if there exists $\varepsilon > 0$ such that $\mathbb{E}_{q_{\eta_\diamond}}[(p/q_{\eta_\diamond})^{1+\varepsilon}] < +\infty$, then we have

$$\lim_{n\to+\infty} \Pi_{\widehat{\mathcal{Q}}_n,p} = 0, \qquad \mathbb{Q}\text{-almost surely.}$$

Indeed, if we let $w_i = p(y_i)/q_{\eta_\diamond}(y_i)$ for all $i \in [\![1,n]\!]$, then the strong law of large numbers ensures

$$\frac{\max_{1\le i\le n} w_i}{n} = \frac{\left(\max_{1\le i\le n} w_i^{1+\varepsilon}\right)^{1/(1+\varepsilon)}}{n} \le \frac{1}{n^{1-1/(1+\varepsilon)}} \left(\frac{1}{n}\sum_{i=1}^n w_i^{1+\varepsilon}\right)^{1/(1+\varepsilon)} \xrightarrow[n\to+\infty]{\mathbb{Q}-\text{a.s.}} 0.$$

To conlcude, remark that $\Pi_{\widehat{\mathcal{Q}}_n,p} \le (d/n)\cdot \max_{1\le i\le n} w_i$.

C LOCAL PROPERTIES OF THE OBJECTIVE AND THE ITERATION MAPPING

We place ourselves in the context of Section 2.3

The set of fixed points of the iterative mapping $\mathcal{M}_\gamma$, i.e., parameters satisfying $\mathcal{M}_\gamma(\eta) = \eta$, can be characterized from (10). In [Bertholom et al., 2024], the authors show that this set is independent of γ and that it writes

$$\mathrm{Fix}(\mathcal{M}) := \{\eta \in E \ : \ \mathbf{m}_\alpha(\eta) = \mathbf{m}_1(\eta)\} = \{\eta \in E \ : \ \mathcal{M}_\gamma(\eta) = \eta\}, \quad \gamma \in (0,1]. \tag{18}$$

Letting $\varrho(\cdot)$ denote the spectral radius application, they define

$$\rho(\eta) = \alpha\, \varrho\big(\mathbf{v}_1(\eta)^{-1}\, \mathbf{v}_\alpha(\eta)\big)\,.$$

We reproduce in Lemma 9 of Appendix D an important result of [Bertholom et al., 2024] on the stability properties of the fixed points of $\mathcal{M}_\gamma$. Let us briefly explain how they are related to the behavior of the objective ℓ_α around these points. First, we compute the gradient of the objective function ℓ_α:

$$\nabla \ell_\alpha(\eta) = \alpha\, \ell_\alpha(\eta)\Big(\mathbf{m}_\alpha(\eta) - \mathbf{m}_1(\eta)\Big). \tag{19}$$

By (18) and (19), $\mathrm{Fix}(\mathcal{M})$ is exactly the set of critical points of ℓ_α. To get deeper insight into the nature of these critical points, we examine the second-order properties of ℓ_α via its Hessian matrix. At any $\eta \in \mathrm{Fix}(\mathcal{M})$, it writes

$$\mathbf{H}_{\ell_\alpha}(\eta) = \alpha\, \ell_\alpha(\eta)\, \Big(\alpha \mathbf{v}_\alpha(\eta) - \mathbf{v}_1(\eta)\Big). \tag{20}$$

Comparatively, the Jacobian matrix of the iteration operator $\mathcal{M}_\gamma$ at $\eta \in \mathrm{Fix}(\mathcal{M})$ is given by

$$\mathbf{J}_{\mathcal{M}_\gamma}(\eta) = \gamma\alpha \mathbf{v}_1(\eta)^{-1}\mathbf{v}_\alpha(\eta) + (1-\gamma)\mathrm{Id}. \tag{21}$$

An essential consequence of these expressions is that if $\eta \in \mathrm{Fix}(\mathcal{M})$ verifies $\rho(\eta) < 1$, then it is a strict local maximizer of ℓ_α and a locally stable fixed point of $\mathcal{M}_\gamma$. Indeed, the eigenvalues of $\mathbf{H}_{\ell_\alpha}(\eta)$ are of the same signs as those of its congruent matrix $\mathbf{v}_1(\eta)^{-1/2}\mathbf{H}_{\ell_\alpha}(\eta)\mathbf{v}_1(\eta)^{-1/2}/(\alpha\ell_\alpha(\eta))$, which also writes $\alpha\mathbf{v}_1(\eta)^{-1/2}\mathbf{v}_\alpha(\eta)\mathbf{v}_1(\eta)^{-1/2} - \mathrm{Id}$. By remarking that $\alpha\mathbf{v}_1(\eta)^{-1/2}\mathbf{v}_\alpha(\eta)\mathbf{v}_1(\eta)^{-1/2}$ has the same eigenvalues as $\alpha\mathbf{v}_1(\eta)^{-1}\,\mathbf{v}_\alpha(\eta)$, we see that $\mathbf{J}_{\mathcal{M}_\gamma}(\eta)$ has spectral radius below 1 if and only if $\rho(\eta) < 1$. Likewise, $\mathbf{H}_{\ell_\alpha}(\eta)$ has only negative eigenvalues if and only if $\rho(\eta) < 1$.

D PROOFS

D.1 Preliminary technical lemmas

In what follows, we let $\mathcal{B}(\mathbf{y}, r)$ denote the open ball of center $\mathbf{y}$ and radius r, and Leb_d the Lebesgue measure on $\mathbb{R}^d$. The first two lemmas are useful for proving (A2) in the multivariate Gaussian case. Lemma 3 is a technical lemma used in the proof of the main theorem, and Lemmas 4 and 5 are used to study the covariance matrices $\mathbf{v}_1$ and $\mathbf{v}_\alpha$. Lastly, Lemma 6 facilitates the proof of Theorem 4.

Lemma 1. *Let* A *be a Borel set of* $\mathbb{R}^d$*. Suppose that for all* $\mathbf{y} \in \mathrm{A}$*, there exists* $r > 0$ *such that* $\mathrm{Leb}_d(\mathrm{A} \cap \mathcal{B}(\mathbf{y}, r)) = 0$*. Then,* $\mathrm{Leb}_d(\mathrm{A}) = 0$*.*

Proof. Since Leb_d is a regular measure, it suffices to show that for every compact set $\mathsf{K} \subset \mathrm{A}$, we have $\mathrm{Leb}_d(\mathsf{K}) = 0$. We can thus assume without loss of generality that A is compact. In this case, the given assumption implies that A can be covered by finitely many balls whose intersections with A are all Lebesgue-negligible. Hence A is also Lebesgue-negligible. □

Lemma 2. *Let* $d \geq 2$ *and let* f *be a continuously differentiable function from an open set of* $\mathbb{R}^d$ *to* $\mathbb{R}$*. Let* $\mathrm{Z} = f^{-1}(\{0\})$ *and* $\mathrm{C} = (\nabla f)^{-1}(\{0\})$ *denote the zero set and the critical set of* f*. Then we have* $\mathrm{Leb}_d(\mathrm{Z} \setminus \mathrm{C}) = 0$*. In particular, if* $\mathrm{Leb}_d(\mathrm{C}) = 0$*, then* $\mathrm{Leb}_d(\mathrm{Z}) = 0$*.*

Proof. By Lemma 1, to prove that $\mathrm{Leb}_d(\mathrm{Z} \setminus \mathrm{C}) = 0$, it is sufficient to show that for all $\mathbf{y} \in \mathrm{Z} \setminus \mathrm{C}$, there exists $r > 0$ such that $\mathrm{Leb}_d(\mathrm{Z} \cap \mathcal{B}(\mathbf{y}, r)) = 0$.

We let ∂_i denote the partial derivative with respect to the i-th component. If $\mathbf{y} = (y_1, \ldots, y_d)$ is in $\mathbb{R}^d$, we let y_{-i} denote the $(d-1)$-component vector that contains every element in $\mathbf{y}$ except y_i, that is, $\mathbf{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_d)$.

Let $\mathbf{y} = (y_1, \ldots, y_d)$ be an element in $\mathrm{Z} \setminus \mathrm{C}$. Then, there exists $i \in [\![1, d]\!]$ such that $\partial_i f(\mathbf{y}) \neq 0$. By the implicit function theorem, one can find a positive number $r > 0$ and a continuously differentiable function g from $\mathcal{B}(\mathbf{y}_{-i}, r)$ to $\mathbb{R}$ such that for all $\mathbf{y}' \in \mathcal{B}(\mathbf{y}, r)$, we have $f(\mathbf{y}') = 0$ if and only if $g(\mathbf{y}'_{-i}) = y'_i$. Thus, we may express $\mathrm{Leb}_d(\mathrm{Z} \cap \mathcal{B}(\mathbf{y}, r))$ as

$$\mathrm{Leb}_d(\mathrm{Z} \cap \mathcal{B}(\mathbf{y}, r)) = \int_{\mathcal{B}(\mathbf{y}, r)} \mathbb{1}_{\{f(\mathbf{y}')=0\}} \mathrm{Leb}_d(\mathrm{d}\mathbf{y}') = \int_{\mathcal{B}(\mathbf{y}, r)} \mathbb{1}_{\{g(\mathbf{y}'_{-i})=\mathbf{y}'_i\}} \mathrm{Leb}_d(\mathrm{d}\mathbf{y}').$$

By Fubini's theorem, this integral can be computed by integrating with respect to y'_i first. Since the one-dimensional Lebesgue measure of a singleton is null, we get $\mathrm{Leb}_d(\mathrm{Z} \cap \mathcal{B}(\mathbf{y}, r)) = 0$. □

Lemma 3. *Let* $\tilde{\nu}$ *be a positive measure on the Borel sets of a separable topological space* Y*, and* B *an open set of* Y*. If* $\tilde{\nu}(B \cap \mathrm{Supp}(\tilde{\nu})) = 0$*, then* $B \cap \mathrm{Supp}(\tilde{\nu}) = \emptyset$*.*

Proof. Define $F = \mathrm{Supp}(\tilde{\nu}) \cap B^c$. Its complement verifies

$$\tilde{\nu}(F^c) = \tilde{\nu}(B \cup \mathrm{Supp}(\tilde{\nu})^c) \leq \tilde{\nu}(B \cap \mathrm{Supp}(\tilde{\nu})) + \tilde{\nu}(\mathrm{Supp}(\tilde{\nu})^c).$$

When the space Y is separable, we have $\tilde{\nu}(\mathrm{Supp}(\tilde{\nu})^c) = 0$ (see the comments at the end of Section 2.1). Moreover, $\tilde{\nu}(B \cap \mathrm{Supp}(\tilde{\nu})) = 0$ by assumption. Therefore, F is a closed subset of $\mathrm{Supp}(\tilde{\nu})$, and its complement set is open with $\tilde{\nu}$-measure zero. By definition of the support of a measure, we deduce that $F = \mathrm{Supp}(\tilde{\nu})$.

This implies $\mathrm{Supp}(\tilde{\nu}) \subset B^c$, i.e., $B \cap \mathrm{Supp}(\tilde{\nu}) = \emptyset$. □

Recall that two measures defined on the same measurable space are said to be equivalent if they share the same negligible sets.

The following two useful lemmas rely on the fact that if $\mathbb{E}_P[\|S\|^2] < +\infty$, then

$$\begin{aligned}
\mathrm{Ker}(\mathrm{Cov}_P(S)) &= \{u \in \mathsf{H} \ : \ \mathrm{Cov}_P(S)\, u = 0\} \\
&= \{u \in \mathsf{H} \ : \ \langle u, \mathrm{Cov}_P(S)\, u \rangle = \mathrm{Var}_P(\langle u, S \rangle) = 0\} \\
&= \{u \in \mathsf{H} \ : \ \exists M \in \mathbb{R}, \ \{\langle u, S \rangle = M\}, \quad P\text{-a.s.}\}\,. \qquad (22)
\end{aligned}$$

Lemma 4. *Let P and Q be two equivalent probability distributions on $(\mathsf{Y}, \mathcal{Y})$. Let S be a measurable function from $(\mathsf{Y}, \mathcal{Y})$ to the Hilbert space H endowed with its Borel σ-field, and suppose that $\mathbb{E}_P[\|S\|^2] < +\infty$ and $\mathbb{E}_Q[\|S\|^2] < +\infty$. Then we have $\mathrm{Ker}(\mathrm{Cov}_P(S)) = \mathrm{Ker}(\mathrm{Cov}_Q(S))$.*

Proof. The proof follows directly by (22) and by noting that, since P and Q are equivalent, we have $\{\langle u, S\rangle = M\}$ P-a.s if and only if $\{\langle u, S\rangle = M\}$ Q-a.s. □

Lemma 5. *Let P be a probability distribution and ν a measure on $(\mathsf{Y}, \mathcal{Y})$. Let S be a measurable function from $(\mathsf{Y}, \mathcal{Y})$ to the Hilbert space H endowed with its Borel σ-field, and suppose that $\mathbb{E}_P[\|S\|^2] < +\infty$. Suppose that P and ν are equivalent. Then $\mathrm{Cov}_P(S)$ is positive definite if and only if for all $v \in \mathsf{H} \setminus \{0\}$ and $M \in \mathbb{R}$, we have $\nu(\{\langle v, S\rangle = M\}^c) > 0$.*

Proof. Using again (22), $\mathrm{Cov}_P(S)$ is positive definite if and only if, for all $v \in \mathsf{H} \setminus \{0\}$ and $M \in \mathbb{R}$, we have $P(\{\langle v, S\rangle = M\}^c) > 0$. The proof follows from the fact that P and ν are equivalent. □

Lemma 6. *Suppose that E is an open convex subset of H. Then, for all $\eta \in E$ and $\overline{\eta} \in \mathrm{bd}(E)$, we have $[\overline{\eta}, \eta] \cap E = (\overline{\eta}, \eta]$.*

Proof. Let $\eta \in E$, $\overline{\eta} \in \mathrm{bd}(E)$ and $s \in (0, 1)$. It suffices to show that $\eta_\star := s\overline{\eta} + (1-s)\eta$ is in E. Since $\overline{\eta} \in \mathrm{bd}(E)$ and E is open, we can find $\overline{\eta}' \in E$ arbitrarily close to $\overline{\eta}$. Letting η' be defined by the equation $\eta_\star = s\overline{\eta}' + (1-s)\eta'$, we can make it arbitrarily close to η with a proper choice of $\overline{\eta}'$. Since E is open and $\eta \in E$, we can make both $\overline{\eta}'$ and η' belong to E. Finally, our construction implies $\eta_\star \in [\overline{\eta}', \eta']$, and by convexity of E we get $\eta_\star \in E$. □

D.2 Behavior of the α-divergence away from compact sets

We start by studying the case where the parameter gets close to the boundary of E.

Lemma 7. *Let $\mathcal{Q} = \{q_\eta \ : \ \eta \in E\}$ be a regular exponential family. Let $\overline{\eta} \in \mathrm{bd}(E)$ and $(\eta_k) \in E^{\mathbb{N}}$ converging to $\overline{\eta}$. Then we have*

$$\lim_{k\to\infty} \ell_\alpha(\eta_k) = 0. \tag{23}$$

Proof. The proof is divided into two steps.

Step 1 We prove that for any $\eta \in E$, $\lim\limits_{k\to+\infty} \int q_{\eta_k}^\alpha q_\eta^{1-\alpha} \, \mathrm{d}\nu = 0$.

Step 2 We deduce that (23) holds.

Proof of Step 1. Let $\eta \in E$. Recall that E is open and convex, by definition of a regular exponential family, hence $\overline{\eta} \notin E$. Let $(\eta'_k)_{k\geq 0}$ be the sequence whose general term is given by the convex combination $\eta'_k = \alpha\eta_k + (1-\alpha)\eta \in E$, and recall that $\int q_{\eta'_k} \, \mathrm{d}\nu = 1$. Elementary computations show that

$$\int q_{\eta_n}^\alpha q_\eta^{1-\alpha} \, \mathrm{d}\nu = \exp(A(\eta'_n) - \alpha\, A(\eta_n) - (1-\alpha)A(\eta)).$$

Using Fatou's lemma, since the sequence $(\eta_k)_{k\geq 0}$ converges to $\overline{\eta}$, we have

$$\liminf_{k\to+\infty} \exp(A(\eta_k)) = \liminf_{k\to+\infty} \int \exp(\langle \eta_k, S\rangle) \, \mathrm{d}\nu \geq \int \exp(\langle \overline{\eta}, S\rangle) \, \mathrm{d}\nu,$$

which is infinite since $\overline{\eta} \notin E$. On the other hand, $A(\eta)$ is finite. Moreover, the sequence $(A(\eta'_k))_{k\geq 0}$ converges in $\mathbb{R}$. Indeed, $(\eta'_k)_{k\geq 0}$ converges to $\alpha\overline{\eta} + (1-\alpha)\eta$, which belongs to E by Lemma 6. By gathering these three observations, we obtain the claim of **Step 1**.

Proof of Step 2. Let $\eta \in E$. For any positive integer N, since $\alpha \in (0, 1)$, Hölder's inequality yields

$$\begin{aligned}
\ell_\alpha(\eta_k) &= \int_{\{p\leq N\, q_\eta\}} q_{\eta_k}^\alpha p^{1-\alpha} \, \mathrm{d}\nu + \int_{\{p> N\, q_\eta\}} q_{\eta_k}^\alpha p^{1-\alpha} \, \mathrm{d}\nu \\
&\leq N^{1-\alpha} \int_{\{p\leq N\, q_\eta\}} q_{\eta_k}^\alpha q_\eta^{1-\alpha} \, \mathrm{d}\nu + \left(\int_{\{p> N\, q_\eta\}} q_{\eta_k} \, \mathrm{d}\nu\right)^\alpha \left(\int_{\{p> N\, q_\eta\}} p \, \mathrm{d}\nu\right)^{1-\alpha}.
\end{aligned}$$

Using the previous step and the fact that q_{η_k} is a probability density w.r.t. ν, we get that, for any positive integer N,

$$\limsup_{k\to\infty} \ell_\alpha(\eta_k) \leq \left(\int_{\{p > N q_\eta\}} p \, \mathrm{d}\nu \right)^{1-\alpha}.$$

By the dominated convergence theorem and the second condition in (2), and since $q_\eta > 0$ by definition, this upper bound tends to zero as $N \to \infty$, which compmletes the proof. □

We now examine the case where the norm of the parameter goes to infinity.

Lemma 8. *Let $\mathcal{Q} = \{q_\eta \ : \ \eta \in E\}$ be an exponential family that satisfies Assumption* (A1)*, and define $\Pi_{\mathcal{Q},p}$ as in* (13)*. We have*

$$\limsup_{\substack{\|\eta\|\to+\infty \\ \eta\in E}} \ell_\alpha(\eta) \leq \Pi_{\mathcal{Q},p}^{1-\alpha}.$$

Proof. For all $\eta \in E$, let us write $\eta = ru$ with $r \in \mathbb{R}^+$ and $u \in \mathbb{U}$, where $\mathbb{U}$ is defined in (7), so that

$$A(\eta) = \log \left[\int_{\mathsf{Y}} \exp(r\langle S(y), u\rangle) \, \nu(\mathrm{d}y) \right],$$

and $\ell_\alpha(\eta)$ writes

$$\ell_\alpha(\eta) = \int_{\mathsf{Y}} \left[\frac{\exp(r\langle S(y), u\rangle)}{\int_{\mathsf{Y}} \exp(r\langle S(y'), u\rangle)\nu(\mathrm{d}y')} \right]^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y).$$

Let $\varepsilon > 0$. By (2), since Y is σ-compact, there exists a compact set $\mathsf{K} \subset \mathsf{Y}$ whose complement K^c verifies

$$\left(\int_{\mathsf{K}^c} p \, \mathrm{d}\nu \right)^{1-\alpha} \leq \varepsilon. \tag{24}$$

Using Hölder's inequality and (24), we get the following upper-bound on $\ell_\alpha(\eta)$:

$$\begin{aligned} \ell_\alpha(\eta) &= \int_{\mathsf{K}} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) + \int_{\mathsf{K}^c} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \\ &\leq \int_{\mathsf{K}} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) + \varepsilon. \end{aligned} \tag{25}$$

Now it is sufficient to prove that for any compact subset K of Y, we have

$$\limsup_{\substack{\|\eta\|\to+\infty \\ \eta\in E}} \int_{\mathsf{K}} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \leq \Pi_{\mathcal{Q},p}^{1-\alpha}. \tag{26}$$

Let us consider the finite (by assumption on ν in Section 2.1) measure $\tilde{\nu} = \nu\big|_{\mathsf{K}}$ defined for all $A \in \mathcal{Y}$ by $\tilde{\nu}(A) = \nu(A \cap \mathsf{K})$. We set $\mathsf{K}_0 = \mathrm{Supp}(\tilde{\nu})$, which is compact as a closed subset of K.

Define, for all $u \in \mathbb{S}^{d-1}$, $M(u) := \sup\{\langle S(y), u\rangle \ : \ y \in \mathsf{K}_0\}$. The function $(u, y) \mapsto \langle S(y), u\rangle$ is continuous on $\mathbb{S}^{d-1} \times \mathsf{K}_0$ since S is continuous under Assumption (A1). Hence, for all $u \in \mathbb{S}^{d-1}$ the supremum $M(u)$ is attained on the compact set K_0, and the function $u \mapsto M(u)$ is continuous on $\mathbb{S}^{d-1}$.

Lastly, given $u \in \mathbb{U}$ and $\delta > 0$, we introduce the set $\mathsf{D}_\delta(u) := \{y \in \mathsf{K}_0 \ : \ \langle S(y), u\rangle > M(u) - \delta\}$.

Recall that μ_1 and $\mu_{1-\alpha}$ are the measures defined in (3). The remainder of the proof is divided into five steps:

Step 1 For all $u \in \mathbb{S}^{d-1}$ and $\delta > 0$, we have $\nu(\mathsf{D}_\delta(u)) > 0$.

Step 2 For all $u \in \mathbb{U}$, we have $\limsup_{\delta\downarrow 0} \mu_1(\mathsf{D}_\delta(u)) \leq \Pi_{\mathcal{Q},p}$.

Step 3 For any $\delta > 0$, there exists $\varepsilon > 0$ such that for all $u, u' \in \mathbb{S}^{d-1}$ satisfying $\|u - u'\| \leq \varepsilon$, we have $\mathsf{D}_{\delta/2}(u) \subseteq \mathsf{D}_\delta(u') \subseteq \mathsf{D}_{2\delta}(u)$.

Step 4 For $\eta \in E$, $\eta = ru$ with $r \in \mathbb{R}_+$ and $u \in \mathbb{S}^{d-1}$, we have, for any $\delta > 0$,

$$\ell_\alpha^{\mathsf{K}}(\eta) := \int_{\mathsf{K}} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \le \left[\frac{\exp(-r\delta)}{\nu(\mathsf{D}_\delta(u))}\right]^\alpha \mu_{1-\alpha}(\mathsf{K}_0) + \mu_1(\mathsf{D}_{2\delta}(u))^{1-\alpha}.$$

Step 5 The previous steps imply (26).

Proof of Step 1. Let $u \in \mathbb{S}^{d-1}$ and $\delta > 0$. The set $\{y \in \mathsf{Y} \ : \ \langle S(y), u\rangle > M(u) - \delta\}$ is open by (A1). Moreover, its intersection with K_0 corresponds to $\mathsf{D}_\delta(u)$, which is non-empty since the continuous function $y \mapsto \langle S(y), u\rangle$ attains the value $M(u)$ on K_0. To conclude, apply the contraposition of Lemma 3 to the measure $\tilde{\nu}$, whose support is K_0, and the aforementioned open set. As $\mathsf{D}_\delta(u) \subseteq \mathsf{K}_0 \subseteq \mathsf{K}$, we get the claimed result.

Proof of Step 2. Let $u \in \mathbb{U}$. By (2), the measure μ_1 defined in (3) is a finite measure on the Borel sets of Y. The sets $\mathsf{D}_\delta(u)$ are decreasing as δ decreases and their intersection over all $\delta > 0$ is the set $\{y \in \mathsf{K}_0 \ : \ \langle S(y), u\rangle = M(u)\}$. Hence, by the dominated convergence theorem,

$$\lim_{\delta \downarrow 0} \mu_1\left(\mathsf{D}_\delta(u)\right) = \mu_1\left(\{y \in \mathsf{K}_0 \ : \ \langle S(y), u\rangle = M(u)\}\right).$$

This limit is smaller than or equal to the constant $\Pi_{\mathcal{Q},p}$ defined in (13), which proves **Step 2**.

Proof of Step 3. We have $\mathsf{D}_\delta(u) = \{y \in \mathsf{K}_0 \ : \ \zeta(u, y) < \delta\}$ where the function ζ, defined on the compact set $\mathbb{S}^{d-1} \times \mathsf{K}_0$ by $\zeta : (u, y) \mapsto M(u) - \langle S(y), u\rangle$, is continuous with respect to both of its variables and valued in $\mathbb{R}_+$. In particular, there exists $\varepsilon > 0$ such that $|\zeta(u, y) - \zeta(u', y)| < \delta/2$ for all $y \in \mathsf{K}_0$ and $u, u' \in \mathbb{S}^{d-1}$ satisfying $\|u - u'\| \le \varepsilon$. The result follows easily.

Proof of Step 4. Since $\mathsf{D}_\delta(u)$ is a subset of Y, by definition of q_η in (5) with $\eta = r\,u$, we have

$$q_\eta(y) \le \frac{\exp(r\langle S(y), u\rangle)}{\int_{\mathsf{D}_\delta(u)} \exp(r\langle S(y'), u\rangle)\nu(\mathrm{d}y')} \le \frac{\exp(r(\langle S(y), u\rangle - M(u) + \delta))}{\nu(\mathsf{D}_\delta(u))}.$$

Since $\langle S(y), u\rangle - M(u) \le -2\delta$ for all $y \in \mathsf{K}_0 \setminus \mathsf{D}_{2\delta}(u)$, we deduce that for such a y,

$$q_\eta(y) \le \frac{\exp(r(-2\delta + \delta))}{\nu(\mathsf{D}_\delta(u))} = \frac{\exp(-r\delta)}{\nu(\mathsf{D}_\delta(u))}. \tag{27}$$

The definitions of $\tilde{\nu}$, K_0 and $\ell_\alpha^{\mathsf{K}}(\eta)$ imply $\ell_\alpha^{\mathsf{K}}(\eta) = \ell_\alpha^{\mathsf{K}_0}(\eta)$. By splitting the integral on the sets $\mathsf{K}_0 \setminus \mathsf{D}_{2\delta}(u)$ and $\mathsf{D}_{2\delta}(u)$, and by using (27), we get

$$\ell_\alpha^{\mathsf{K}}(\eta) \le \left[\frac{\exp(-r\delta)}{\nu(\mathsf{D}_\delta(u))}\right]^\alpha \mu_{1-\alpha}(\mathsf{K}_0) + \int_{\mathsf{D}_{2\delta}(u)} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y).$$

Hölder's inequality and the fact that q_η is a probability density with respect to ν further imply

$$\int_{\mathsf{D}_{2\delta}(u)} q_\eta(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \le \left(\int_{\mathsf{D}_{2\delta}(u)} q_\eta \,\mathrm{d}\nu\right)^\alpha \left(\int_{\mathsf{D}_{2\delta}(u)} p \,\mathrm{d}\nu\right)^{1-\alpha} \le \mu_1(\mathsf{D}_{2\delta}(u))^{1-\alpha}.$$

Combining the last two inequalities yields the claimed result.

Proof of Step 5. Suppose, by way of contradiction, that (26) does not hold. Then there exists $L > \Pi_{\mathcal{Q},p}^{1-\alpha}$ and $(\eta_n) \in E^{\mathbb{N}}$ such that, for all n, $r_n := \|\eta_n\| \ge 2^n$ and $\ell_\alpha^{\mathsf{K}}(\eta_n) \ge L$. Introduce $u_n = \eta_n / r_n \in \mathbb{S}^{d-1}$. Taking an adequate subsequence, we can assume without loss of generality that the sequence (u_n) converges to some $u_\star \in \mathbb{S}^{d-1}$. By definition of $\mathbb{U}$ in (7), we also have $u_\star \in \mathbb{U}$. Applying **Step 2**, we can find $\delta > 0$ such that $\mu_1(\mathsf{D}_{4\delta}(u_\star))^{1-\alpha} < L$. Hence with the previous condition involving L, to exhibit a contradiction, it suffices to show that

$$\limsup_{n \to +\infty} \ell_\alpha^{\mathsf{K}}(\eta_n) \le \mu_1(\mathsf{D}_{4\delta}(u_\star))^{1-\alpha}. \tag{28}$$

Applying **Step 3**, for sufficiently large n we have

$$\mathsf{D}_{\delta/2}(u_\star) \subseteq \mathsf{D}_\delta(u_n) \quad \text{and} \quad \mathsf{D}_{2\delta}(u_n) \subseteq \mathsf{D}_{4\delta}(u_\star).$$

Thus, applying **Step 4**, when n is large enough we also have

$$\ell_\alpha^{\mathsf{K}}(\eta_n) \leq \left[\frac{\exp(-r_n\delta)}{\nu(\mathsf{D}_{\delta/2}(u_\star))}\right]^\alpha \mu_{1-\alpha}(\mathsf{K}_0) + \mu_1(\mathsf{D}_{4\delta}(u_\star))^{1-\alpha}. \tag{29}$$

By **Step 1**, we know that $\nu(\mathsf{D}_{\delta/2}(u_\star)) > 0$. Therefore, Hölder's inequality ensures

$$\mu_{1-\alpha}(\mathsf{K}_0) \leq \left(\int_{\mathsf{K}_0} p \, \mathrm{d}\nu\right)^{1-\alpha} \nu(\mathsf{K}_0)^\alpha < +\infty.$$

By letting $n \to +\infty$ with $r_n \geq 2^n$ in (29), we obtain (28), which concludes the proof. □

D.3 General properties of the iteration maps and their fixed points

The following lemma is quoted from [Bertholom et al., 2024]. It provides a characterization of the local maximizers of ℓ_α.

Lemma 9. *Let $\eta_\star \in \mathrm{Fix}(\mathcal{M})$, where $\mathrm{Fix}(\mathcal{M})$ is defined in (18). Let $\|\cdot\|_\star$ denote the norm on H defined for $x \in \mathsf{H}$ by $\|x\|_\star = \left\|\mathbf{v}_1(\eta_\star)^{1/2}\, x\right\|$, and let $\rho_\star := \rho(\eta_\star)$, where $\rho(\eta_\star)$ is defined in (14). For all $\gamma \in (0,1]$ and any norm $\|\cdot\|_\bullet$ on H, we have*

$$\limsup_{\substack{\eta,\eta' \to \eta_\star \\ \eta \neq \eta'}} \frac{\|\mathcal{M}_\gamma(\eta) - \mathcal{M}_\gamma(\eta')\|_\star}{\|\eta - \eta'\|_\star} = 1 - \gamma(1-\rho_\star) \leq \limsup_{\substack{\eta,\eta' \to \eta_\star \\ \eta \neq \eta'}} \frac{\|\mathcal{M}_\gamma(\eta) - \mathcal{M}_\gamma(\eta')\|_\bullet}{\|\eta - \eta'\|_\bullet}. \tag{30}$$

Moreover, $\rho_\star < 1$ if and only if there exists a norm on H such that for all $\gamma \in (0,1]$, the mapping $\mathcal{M}_\gamma$ is a contraction in a neighborhood of $\eta_\star$. In that case, $\eta_\star$ is a strict local maximizer of ℓ_α.

In the following lemma, we have collected the properties already derived in Section 2.3.

Lemma 10. *Let $0 < \gamma \leq 1$. We have the following assertions:*

- *(i) For $0 < s \leq 1$, $\mathbf{m}_s$ is a $\mathcal{C}^\infty$-local diffeomorphism on E whose Jacobian matrix $\mathbf{J}_{\mathbf{m}_s} = s\,\mathbf{v}_s$ is symmetric positive definite.*
- *(ii) The matrix $\alpha\,\mathbf{v}_1(\eta)^{-1}\,\mathbf{v}_\alpha(\eta)$ is diagonalizable with positive eigenvalues.*
- *(iii) The mapping ℓ_α is of class $\mathcal{C}^\infty$ on E, with gradient given by (19).*
- *(iv) The definition set $\mathcal{D}_\gamma$ of $\mathcal{M}_\gamma$ defined in (11) is an open subset of E.*
- *(v) The mapping $\mathcal{M}_\gamma$ is a $\mathcal{C}^\infty$-local diffeomorphism on the open set $\mathcal{D}_\gamma$.*
- *(vi) The set $\mathrm{Fix}(\mathcal{M})$ of its fixed points coincides with set of critical points of ℓ_α defined in (18), thus does not depend on γ.*
- *(vii) For all $\eta \in \mathrm{Fix}(\mathcal{M})$, the Jacobian matrix $\mathbf{J}_{\mathcal{M}_\gamma}(\eta)$ is given by (21). In other words, it is a $(1-\gamma,\gamma)$-barycenter between the identity matrix and $\alpha\,\mathbf{v}_1(\eta)^{-1}\,\mathbf{v}_\alpha(\eta)$.*
- *(viii) For all $\eta \in \mathrm{Fix}(\mathcal{M})$, counting multiplicities, the number of positive (resp. zero, resp. negative) eigenvalues of $\mathbf{H}_{\ell_\alpha}(\eta)$ is equal to the number of eigenvalues of $\alpha\,\mathbf{v}_1(\eta)^{-1}\,\mathbf{v}_\alpha(\eta)$ larger than (resp. equal to, resp. smaller than) 1.*

Moreover, Assertion (vii) implies that, if $\eta \in \mathrm{Fix}(\mathcal{M})$, then $\mathbf{J}_{\mathcal{M}_\gamma}(\eta)$ and $\alpha\,\mathbf{v}_1(\eta)^{-1}\,\mathbf{v}_\alpha(\eta)$ have the same eigenvectors and their corresponding eigenvalues are on the same side of 1. In particular, for such points η, (viii) still holds with $\mathbf{J}_{\mathcal{M}_\gamma}(\eta)$ replacing $\alpha\,\mathbf{v}_1(\eta)^{-1}\,\mathbf{v}_\alpha(\eta)$.

Corollary 1. *Under Assumption (A3), the points in $\mathrm{Fix}(\mathcal{M})$ are isolated and for every compact set $\mathsf{K} \subset E$, the set $\mathsf{K} \cap \mathrm{Fix}(\mathcal{M})$ is finite or empty.*

Proof. To prove that the points in $\mathrm{Fix}(\mathcal{M})$ are isolated, we suppose that $\mathrm{Fix}(\mathcal{M})$ contains a non-isolated point η and show that (A3) does not hold. Let $(\eta_n)_{n\in\mathbb{N}}$ be a sequence valued in $\mathrm{Fix}(\mathcal{M}) \setminus \{\eta\}$, converging to η. By Taylor's formula, we have $\mathcal{M}_1(\eta_n) = \mathcal{M}_1(\eta) + \mathbf{J}_{\mathcal{M}_1}(\eta_n)(\eta_n - \eta) + o(\|\eta_n - \eta\|)$ and by Lemma 10 (vi), these points are all fixed points of $\mathcal{M}_1$. Thus, letting $\eta'_n = (\eta_n - \eta)/\|\eta_n - \eta\|$, as $n \to +\infty$ we have

$$\mathbf{J}_{\mathcal{M}_1}(\eta)\, \eta'_n = \eta'_n + o(1).$$

Observe that the sequence $(\eta'_n)_{n\in\mathbb{N}}$ lies in the unit sphere of H by definition of η'_n, hence it admits a converging subsequence in this sphere. The corresponding limit is an eigenvector of $\mathbf{J}_{\mathcal{M}_1}(\eta)$ for the eigenvalue 1. By Lemma 10-(vii), it contradicts Assumption (A3).

Now, we let $\mathsf{K} \subset E$ be a compact set. To show that $\mathsf{K} \cap \mathrm{Fix}(\mathcal{M})$ is a finite or empty set, we suppose that $\mathsf{K} \cap \mathrm{Fix}(\mathcal{M})$ is infinite and show that $\mathrm{Fix}(\mathcal{M})$ admits a non-isolated point. Let $(\eta_n)_{n\in\mathbb{N}}$ be a sequence of distinct points in $\mathsf{K} \cap \mathrm{Fix}(\mathcal{M})$. It admits a subsequence that converges in K, and by continuity of $\mathcal{M}_1$ the limit must also be in $\mathrm{Fix}(\mathcal{M})$. So this limit is non-isolated in $\mathrm{Fix}(\mathcal{M})$. □

D.4 Step-size selection

In a minimal regular exponential family, the set of mean parameters $F = \nabla A(E)$ is an open convex set, and E and F are the definition sets of two Lagrange conjugates, see [Rockafellar, 1970, Theorem 26.5]. Beside, it is shown in [Wainwright and Jordan, 2008, Theorem 3.3] that F is the interior of the convex set

$$\tilde{F} := \left\{ \int S \,\mathrm{d}P \ : \ P \text{ is a probability dominated by } \nu \text{ such that } \int \|S\| \,\mathrm{d}P < \infty \right\}.$$

We use these results to specify the definition set $\mathcal{D}_\gamma$ of the mapping $\mathcal{M}_\gamma$, which is the open subset of E defined in (11).

Lemma 11. *For all $\gamma \in [0, 1)$, we have*

$$\mathrm{Fix}(\mathcal{M}) \subset \mathcal{D}_1 \subseteq \mathcal{D}_\gamma = E.$$

Proof. The first inclusion is a direct consequence of (11) and (18).

There only remains to show that $\mathcal{D}_\gamma = E$, since the inclusion $\mathcal{D}_1 \subseteq \mathcal{D}_\gamma$ will follow immediately. Let $\eta \in E$. By definition of $\mathbf{m}_\alpha$ in (8), we have

$$\mathbf{m}_\alpha(\eta) = \int S\, \pi_{\alpha,\eta} \,\mathrm{d}\nu \in \tilde{F}.$$

On the other hand, from what precedes, $\mathbf{m}_1(\eta)$ belongs to the interior of the convex set $\tilde{F}$. Since $\gamma \in [0, 1)$, we conclude with Lemma 6 that $\gamma\mathbf{m}_\alpha(\eta) + (1-\gamma)\mathbf{m}_1(\eta) \in (\mathbf{m}_\alpha(\eta), \mathbf{m}_1(\eta)] \subset F$. Hence $\eta \in \mathcal{D}_\gamma$, which concludes the proof. □

An immediate consequence of this result is that the mapping β defined in Assumption (D2) satisfies

$$\beta(\eta) = \beta^{\mathbb{1}_{\mathcal{D}_1}(\eta)} = \begin{cases} 1 & \text{if } \eta \in \mathcal{D}_1, \\ \beta & \text{otherwise.} \end{cases} \tag{31}$$

D.5 Avoiding local minima and saddle points

Intuitively, the reason why the MA defined in (9) avoids saddle points and local minima of the objective ℓ_α is that the set of initial points that may lead to such limits is "small enough". That is, these points can be avoided by properly randomizing the initialization, even without knowing where the flawed initial points are in the parameter space.

To formalize this idea, we introduce the following two subsets of E.

$E_{\eta_\star}(\Gamma)$: For a given $\eta_\star \in E$ and a set $\Gamma \subset (0, 1]$, we let $E_{\eta_\star}(\Gamma)$ denote the set of initial points $\eta_0 \in E$ for which there exist sequences $(\eta_k) \in E^{\mathbb{N}}$ and $(\gamma_k) \in \Gamma^{\mathbb{N}}$ satisfying (9) and $\lim \eta_k = \eta_\star$.

$E'_{\eta_\star}(\Gamma)$: We also define $E'_{\eta_\star}(\Gamma)$ as the subset of $E_{\eta_\star}(\Gamma)$ obtained by adding the constraint that (γ_k) is eventually constant, i.e., there exists $T \in \mathbb{N}$ such that $\gamma_k = \gamma_T$ for all $k \geq T$.

We start with a lemma useful for proving that the MA does not converge to local minimizers of ℓ_α.

Lemma 12. *Let Γ be a countable subset of $(0,1]$ and let $\eta_\star$ be a local minimizer of ℓ_α. The set $E_{\eta_\star}(\Gamma)$ is at most countable.*

The proof of this lemma relies on the following result.

Lemma 13. *Let U be an open set of $\mathbb{R}^d$, let $k \geq 1$, and let $\phi : U \to \mathbb{R}^d$ be a $\mathcal{C}^k$-local diffeomorphism. Then, there exists an at most countable collection of open sets $(U_i)_{i\in I}$ covering U, such that ϕ is a $\mathcal{C}^k$-diffeomorphism from each U_i onto $\phi(U_i)$. Moreover, for any set $A \subset \mathbb{R}^d$, the preimage $\phi^{-1}(A)$ is included in a countable union of preimages of A by $\mathcal{C}^k$-diffeomorphisms.*

Proof. We can cover U with a collection $(\mathsf{K}_n)_{n\in\mathbb{N}}$ such that each K_n is a compact set included in U (take an increasing covering (K'_n) of $\mathbb{R}^d$ and intersect each K'_n with the closed set of points in U that are at a distance larger than or equal to 2^{-n} from the complement set of U). Now, each $x \in U$ admits an open neighborhood on which ϕ is a $\mathcal{C}^k$-diffeomorphism onto its image, and each K_n can be covered by finitely many such open neighborhoods. Taking the union of these coverings over $n \in \mathbb{N}$, we obtain the collection of open sets $(U_i)_{i\in I}$, which is at most countable.

To conclude, define for all $i \in I$ the $\mathcal{C}^k$-diffeomorphism $\phi_i = \phi_{|U_i} : U_i \to \phi(U_i)$. Then, for any set $A \subset \mathbb{R}^d$, we have

$$\phi^{-1}(A) \subseteq \bigcup_{i\in I} \phi_i^{-1}(A).$$

This proves the second claim. □

Proof of Lemma 12. Let $\eta_0 \in E_{\eta_\star}(\Gamma)$, that is, take $(\eta_k) \in E^{\mathbb{N}}$ and $(\gamma_k) \in \Gamma^{\mathbb{N}}$ satisfying (9) and such that (η_k) converges to a local minimizer $\eta_\star$ of ℓ_α. By Theorem 2, the sequence $(\ell_\alpha(\eta_k))_{k\in\mathbb{N}}$ is either strictly increasing or it remains constant for sufficiently large k. Since $\eta_\star$ is a local minimum of ℓ_α, the latter case is the only possibility. Hence, there exists $T \in \mathbb{N}$ such that $\mathcal{M}_{\gamma_T} \circ \cdots \circ \mathcal{M}_{\gamma_0}(\eta_0) = \eta_\star$. Consequently, we have the inclusion

$$E_{\eta_\star}(\Gamma) \subset \bigcup_{\substack{T\in\mathbb{N} \\ (\gamma_0,\ldots,\gamma_T)\in\Gamma^{T+1}}} \{\eta \in E \ : \ \mathcal{M}_{\gamma_T} \circ \cdots \circ \mathcal{M}_{\gamma_0}(\eta) = \eta_\star\}.$$

Since, for all $\gamma > 0$, $\mathcal{M}_\gamma$ is a $\mathcal{C}^\infty$-local diffeomorphism on its (open) definition set, by Lemma 13 the preimage of a countable set by $\mathcal{M}_\gamma$ is countable. Thus, for all $(\gamma_0, \ldots, \gamma_T) \in \Gamma^T$, the mapping $\mathcal{M}_{\gamma_T} \circ \cdots \circ \mathcal{M}_{\gamma_0}^{-1}(\{\eta_\star\})$ is at most countable. Since Γ is countable, we conclude that $E_{\eta_\star}(\Gamma)$ is at most countable. □

To prove that the MA avoids strict saddle points, we now study the set $E'_{\eta_\star}(\Gamma)$, hence in contrast to limits that are local minimizers, we now require the step-sizes to remain at a fixed value after a sufficiently large number of iterations. This condition is satisfiable in our setting, see Appendix D.4.

Lemma 14. *Let Γ be a countable subset of $(0,1]$ and let $\eta_\star$ be a critical point of ℓ_α. If the Hessian matrix of ℓ_α at $\eta_\star$ has at least one positive eigenvalue, then the set $E'_{\eta_\star}(\Gamma)$ is included in a countable union of manifolds of dimension at most $d-1$.*

The statement of Lemma 14 is shown by applying the stable manifold theorem, a fundamental result from dynamical systems theory. In this paper, we will a formulation that follows directly from an application of the local stable manifold theorem as stated in [Shub, 2013, Theorem III.7], with assumptions adapted to our framework.

Corollary 2. *Let U be an open subset of $\mathbb{R}^d$ and $\phi : U \to \mathbb{R}^d$ be a $\mathcal{C}^k$-local diffeomorphism, where k is a positive integer. Suppose that $x \in U$ is a fixed point of ϕ such that the Jacobian $\mathbf{J}_\phi(x)$ is diagonalizable. We let E^{sc} denote the sum of eigenspaces of $\mathbf{J}_\phi(x)$ corresponding to eigenvalues of modulus less than or equal to 1.*

There exist a $\mathcal{C}^k$-embedded open ball W^{sc}_{loc} and an open set B that contain x and are such that

(i) $T_x(W^{sc}_{\mathrm{loc}}) = E^{sc}$, *where* $T_x(W^{sc}_{\mathrm{loc}})$ *is the tangent space of* W^{sc}_{loc} *at* x,

(ii) $W^{sc}_{\mathrm{loc}} \supset \{y \in B \ : \ \forall n \in \mathbb{N},\ \phi^n(y) \in B\}$.

This result requires further explanation. First, note that the space E^{sc} is formed by the directions corresponding to eigenvalues of $\mathbf{J}_\phi(x)$ with absolute value less than or equal to one. Intuitively, these directions are stable under the iterative application of ϕ in the neighborhood of the fixed point.

When we say that W^{sc}_{loc} is a $\mathcal{C}^k$-embedded open ball, it means that there exists an embedding of class $\mathcal{C}^k$ from an open ball in a Euclidean space whose image is W^{sc}_{loc}. An embedding is a map that is both an immersion, i.e., its derivative is injective at every point; and a homeomorphism. This guarantees that the local stable manifold W^{sc}_{loc} is geometrically well-behaved.

The tangent space $T_x(W^{sc}_{\mathrm{loc}})$ is the real vector space that encompasses all possible directions along which one can move within W^{sc}_{loc}, while tangentially passing through x. In other words, it is the set of all velocity vectors of smooth curves contained in W^{sc}_{loc} that pass through x. It can be identified with the fiber above x in the tangent bundle $T(W^{sc}_{\mathrm{loc}})$.

In the first assertion of Corollary 2, the equation $T_x(W^{sc}_{\mathrm{loc}}) = E^{sc}$ means that, at the fixed point x, the manifold W^{sc}_{loc} aligns with the stable directions of ϕ in the neighborhood of the fixed point. The second assertion of Corollary 2 states that W^{sc}_{loc} indeed contains all local stable points, that is, those that remain in neighborhood of the fixed point when iterating infinitely many times the mapping ϕ.

With these elements in mind, we can prove Lemma 14.

Proof of Lemma 14. Let $\eta_0 \in E'_{\eta_\star}(\Gamma)$, that is, take $(\eta_k) \in E^{\mathbb{N}}$ and $(\gamma_k) \in \Gamma^{\mathbb{N}}$ satisfying (9), such that (η_k) converges to $\eta_\star$ and $\gamma_k = \gamma_T$ for all $k \geq T$.

Lemma 10 ensures that for all $\gamma \in \Gamma$, the mapping $\mathcal{M}_\gamma$ is a $\mathcal{C}^\infty$-local diffeomorphism. More precisely, by Assertions (ii) and (vii) in Lemma 10, we know that $\mathbf{J}_{\mathcal{M}_\gamma}(\eta_\star)$ is diagonalizable with positive eigenvalues. We are particularly interested in the mapping $\mathcal{M}_{\gamma_T}$, and apply Corollary 2 with $\phi = \mathcal{M}_{\gamma_T}$, $x = \eta_\star$, and $U = \mathcal{D}_{\gamma_T}$. Let W^{sc}_{loc} and B denote respectively the local stable manifold and the open neighborhood associated with the fixed point $\eta_\star$ in Corollary 2.

The dimension of the manifold W^{sc}_{loc} is equal to the dimension of $T_{\eta_\star}(W^{sc}_{\mathrm{loc}})$, its tangent space at $\eta_\star$. Thus, by Corollary 2, counting multiplicities, the dimension of W^{sc}_{loc} equals the number of eigenvalues of $\mathbf{J}_\phi(\eta_\star)$ with modulus strictly less than or equal to 1. Since we have assumed in Lemma 14 that $\mathbf{H}_{\ell_\alpha}(\eta_\star)$ has at least one negative eigenvalue, by Lemma 10-$(viii)$, we deduce that $\alpha\, \mathbf{v}_1(\eta)^{-1}\, \mathbf{v}_\alpha(\eta)$ has at least one eigenvalue larger than 1. Using Lemma 10-(vii) again, the same goes for $\mathbf{J}_\phi(\eta_\star)$. Consequently, the manifold W^{sc}_{loc} has dimension at most $d-1$.

Let $\eta' \in U$ be such that $\lim_{k\to+\infty} \phi^k(\eta') = \eta_\star$. Since B is a neighborhood of $\eta_\star$, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, we have $\phi^k(\eta') \in B$. Writing $k = k_0 + \ell$, this reads

$$\phi^{k_0}(\eta') \in \bigcap_{\ell=0}^{+\infty} \phi^{-\ell}(B) \subset W^{sc}_{\mathrm{loc}},$$

where the inclusion comes from the second assertion in Corollary 2. We deduce that

$$\left\{\eta' \in \mathcal{D}_{\gamma_T} \ : \ \lim_{k\to+\infty} \mathcal{M}^k_{\gamma_T}(\eta') = \eta_\star\right\} \subseteq \bigcup_{k_0=0}^{+\infty} \mathcal{M}^{-k_0}_{\gamma_T}\left(W^{sc}_{\mathrm{loc}}\right).$$

Now, by definition of the sequence (η_k) above, we have

$$\mathcal{M}_{\gamma_{T-1}} \circ \cdots \circ \mathcal{M}_{\gamma_0}(\eta_0) \in \left\{\eta' \in \mathcal{D}_{\gamma_T} \ : \ \lim_{k\to+\infty} \mathcal{M}^k_{\gamma_T}(\eta') = \eta_\star\right\}.$$

We conclude that

$$\eta_0 \in \bigcup_{k_0=0}^{+\infty} \mathcal{M}^{-1}_{\gamma_0} \circ \cdots \circ \mathcal{M}^{-1}_{\gamma_{T-1}} \circ \mathcal{M}^{-k_0}_{\gamma_T}\left(W^{sc}_{\mathrm{loc}}\right).$$

Finally, we obtain

$$E'_{\eta_\star}(\Gamma) \subseteq \bigcup_{\substack{T,k_0\in\mathbb{N}\\(\gamma_0,\dots,\gamma_T)\in\Gamma^{T+1}}} \mathcal{M}_{\gamma_0}^{-1}\circ\cdots\circ\mathcal{M}_{\gamma_{T-1}}^{-1}\circ\mathcal{M}_{\gamma_T}^{-k_0}\left(W_{\mathrm{loc}}^{sc}\right).$$

Using Lemma 13, since W_{loc}^{sc} has dimension at most $d-1$ and $\mathcal{M}_\gamma$ is a $\mathcal{C}^\infty$-local diffeomorphism for all $\gamma\in\Gamma$, we get the claimed result. □

D.6 Convergence results

Proof of Theorem 3. We prove each assertion separately.

Proof of Assertion (i). Let $\eta_\star$ a limit point of (η_k). Then, by compactness of K and of the line segment $[\underline{\gamma},1]$, taking successive subsequences, there exists a strictly increasing function $\chi:\mathbb{N}\to\mathbb{N}$ such that

- $(\eta_{\chi(k)})$ converges to $\eta_\star\in\mathsf{K}$,
- $(\eta_{\chi(k)+1})$ converges to $\eta_{\star\star}\in\mathsf{K}$,
- $(\gamma_{\chi(k)})$ converges to $\gamma_\star\in[\underline{\gamma},1]$.

Since $\eta_{\chi(k)+1}=\mathcal{M}_{\gamma_{\chi(k)}}(\eta_{\chi(k)})$ we have

$$\mathbf{m}_1(\eta_{\chi(k)+1})=\gamma_{\chi(k)}\mathbf{m}_\alpha(\eta_{\chi(k)})+(1-\gamma_{\chi(k)})\mathbf{m}_1(\eta_{\chi(k)}).$$

By continuity of $\mathbf{m}_1$ and $\mathbf{m}_\alpha$, taking the limit as $k\to+\infty$ yields

$$\mathbf{m}_1(\eta_{\star\star})=\gamma_\star\,\mathbf{m}_\alpha(\eta_\star)+(1-\gamma_\star)\,\mathbf{m}_1(\eta_\star).$$

We deduce that $\mathcal{M}_{\gamma_\star}(\eta_\star)=\eta_{\star\star}$. Note that $\chi(k+1)\geq\chi(k)+1$, thus $\eta_{\chi(k+1)}$ is obtained by applying mappings $\mathcal{M}_{\gamma_m}$ successively to $\eta_{\chi(k)+1}$ for $m=\chi(k)+1,\dots,\chi(k+1)-1$. Therefore, the monotonicity property established by Theorem 2 implies $\ell_\alpha(\eta_{\chi(k+1)})\geq\ell_\alpha(\eta_{\chi(k)+1})$. Taking the limit, by continuity of ℓ_α, we get $\ell_\alpha(\eta_\star)\geq\ell_\alpha(\eta_{\star\star})$. But since we showed that $\mathcal{M}_{\gamma_\star}(\eta_\star)=\eta_{\star\star}$ we also have the reverse inequality, hence

$$\ell_\alpha(\eta_\star)=\ell_\alpha(\eta_{\star\star})=\ell_\alpha(\mathcal{M}_{\gamma_\star}(\eta_\star)).$$

The equality case of Theorem 2 guarantees $\eta_\star\in\mathrm{Fix}(\mathcal{M})$.

Proof of Assertion (ii). Suppose that $\mathrm{Fix}(\mathcal{M})\cap\mathsf{K}$ is a finite set, and let $\eta_{\star1},\dots,\eta_{\star p}$ be the distinct limit points of (η_k) in $\mathrm{Fix}(\mathcal{M})\cap\mathsf{K}$, where p is a positive integer. Additionally, we introduce

$$\delta:=\min_{j\neq k}\|\eta_{\star j}-\eta_{\star k}\|>0.$$

Note that $\mathcal{D}_1$ is an open set containing $\mathrm{Fix}(\mathcal{M})$ (see Lemma 11), in particular it contains the points $\{\eta_{\star1},\dots,\eta_{\star p}\}$. Moreover, the mapping $(\gamma,\eta)\mapsto\mathcal{M}_\gamma(\eta)$ is of class $\mathcal{C}^\infty$ on $[0,1]\times\mathcal{D}_1$. Thus, there exists a closed neighborhood $\mathcal{V}$ of $\{\eta_{\star1},\dots,\eta_{\star p}\}$ and a constant $C>0$ such that, for all $\eta,\eta'\in\mathcal{V}$ and $\gamma\in[\underline{\gamma},1]$, we have

$$\|\mathcal{M}_\gamma(\eta)-\mathcal{M}_\gamma(\eta')\|\leq C\,\|\eta-\eta'\|.$$

Thus for a sufficiently small $\varepsilon>0$, we have, for all η in a ball of radius ε centered at the fixed point $\eta_{\star j}$ and all $\gamma\in[\underline{\gamma},1]$,

$$\|\mathcal{M}_\gamma(\eta)-\eta_{\star j}\|=\|\mathcal{M}_\gamma(\eta)-\mathcal{M}_\gamma(\eta_{\star j})\leq C\,\|\eta-\eta_{\star j}\|\leq C\,\varepsilon.$$

Consequently, for all $j'\neq j$ and all $k\in\mathbb{N}$, if $\|\eta_k-\eta_{\star j}\|<\varepsilon$ then

$$\|\eta_{k+1}-\eta_{\star j'}\|\geq\delta-\|\eta_{k+1}-\eta_{\star j}\|\geq\delta-C\,\varepsilon\geq2\varepsilon, \tag{32}$$

where the last inequality holds as soon as $\varepsilon\leq\delta/(C+2)$.

On the other hand, by definition of $\eta_{\star1},\dots,\eta_{\star p}$, for any $\varepsilon>0$, there exists $k_0\in\mathbb{N}$ such that, for all $k\geq k_0$, η_k lies at a distance smaller than ε of the set $\{\eta_{\star1},\dots,\eta_{\star p}\}$ (otherwise we could build another limit point in K).

Choosing ε small enough so that (32) also holds, this implies that η_k must remain at a distance less than ε of the same $\eta_{\star j}$ for all $k \geq k_0$. We conclude that (η_k) converges.

Proof of Assertion (iii). Suppose that

$$\lim_{k\to\infty} \eta_k = \eta_\star, \tag{33}$$

where $\eta_\star \in \mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}$, and that $\mathbf{H}_{\ell_\alpha}(\eta_\star)$ only has negative eigenvalues By Lemma 10-(ii) and Lemma 10-$(viii)$, this implies that the eigenvalues of $\alpha\, \mathbf{v}_1(\eta_\star)^{-1}\, \mathbf{v}_\alpha(\eta_\star)$ all belong to $(0,1)$. In particular, in (14), we have $\rho_\star := \rho(\eta_\star) < 1$.

Let $\mathcal{B}_\varepsilon = \{\eta \in E \ : \ \|\eta - \eta_\star\| < \varepsilon\}$. By Lemma 11, there exists a sufficiently small $\varepsilon > 0$ such that for all $\gamma \in [0,1]$ we have the inclusions $\mathcal{B}_\varepsilon \subseteq \mathcal{D}_1 \subseteq \mathcal{D}_\gamma$, and the identity

$$\sup_{\substack{\eta,\eta' \in \mathcal{B}_\varepsilon \\ \eta \neq \eta'}} \frac{\|\mathcal{M}_\gamma(\eta) - \mathcal{M}_\gamma(\eta')\|_\star}{\|\eta - \eta'\|_\star} = \sup_{\eta \in \mathcal{B}_\varepsilon} \left\|\mathbf{J}_{\mathcal{M}_\gamma}(\eta)\right\|_{\mathrm{op}\star}. \tag{34}$$

In (34), $\|\cdot\|_\star$ denotes the norm on H defined in Lemma 9, and $\|\cdot\|_{\mathrm{op}\star}$ denotes the associated operator norm on linear operators from H to H. Thus, the equality in (30) in Lemma 9 rewrites

$$\lim_{\varepsilon \downarrow 0} \sup_{\eta \in \mathcal{B}_\varepsilon} \left\|\mathbf{J}_{\mathcal{M}_\gamma}(\eta)\right\|_{\mathrm{op}\star} = 1 - \gamma(1 - \rho_\star).$$

By continuity of $(\eta, \gamma) \mapsto \mathbf{J}_{\mathcal{M}_\gamma}(\eta)$ on $\mathcal{D}_1 \times [0,1]$, we deduce that, for any $\gamma' \in [0,1]$,

$$\limsup_{\varepsilon \downarrow 0} \sup_{\gamma \in [\gamma',1]} \sup_{\eta \in \mathcal{B}_\varepsilon} \left\|\mathbf{J}_{\mathcal{M}_\gamma}(\eta)\right\|_{\mathrm{op}\star} \leq 1 - \gamma'(1 - \rho_\star).$$

Using (34) again yields

$$\limsup_{\varepsilon \downarrow 0} \sup_{\gamma \in [\gamma',1]} \sup_{\substack{\eta,\eta' \in \mathcal{B}_\varepsilon \\ \eta \neq \eta'}} \frac{\|\mathcal{M}_\gamma(\eta) - \mathcal{M}_\gamma(\eta')\|_\star}{\|\eta - \eta'\|_\star} \leq 1 - \gamma'(1 - \rho_\star) = \gamma' \rho(\eta_\star) + (1 - \gamma'). \tag{35}$$

Now, let ρ be some constant in $(\gamma_\star\, \rho(\eta_\star) + (1 - \gamma_\star), 1)$. We can find $\gamma' > \gamma_\star$ close enough to $\gamma_\star$ such that $\gamma' \rho(\eta_\star) + (1 - \gamma') < \rho$. Hence, by (35), we can find $\varepsilon > 0$ such that

$$\sup_{\gamma \in [\gamma',1]} \sup_{\substack{\eta,\eta' \in \mathcal{B}_\varepsilon \\ \eta \neq \eta'}} \frac{\|\mathcal{M}_\gamma(\eta) - \mathcal{M}_\gamma(\eta')\|_\star}{\|\eta - \eta'\|_\star} \leq \rho.$$

By definition of $\gamma_\star$ and using (33), there exists $T \in \mathbb{N}$ such that, for all $k \geq T$, we have $\gamma_k \in [\gamma', 1]$ and $\eta_k \in \mathcal{B}_\varepsilon$. The previous equation implies (as in the usual fixed point theorem),

$$\|\eta_{k+1} - \eta_\star\|_\star = \|\mathcal{M}_\gamma(\eta_k) - \mathcal{M}_\gamma(\eta_\star)\|_\star \leq \rho \|\eta_k - \eta'\|_\star.$$

We conclude that $\|\eta_k - \eta_\star\|_\star = \mathcal{O}(\rho^k)$ as $k \to +\infty$. □

We now have all the ingredients to prove Theorem 1.

Proof of Theorem 1. Recall the definitions of $E_{\eta_\star}(\Gamma)$ and $E'_{\eta_\star}(\Gamma)$ at the beginning of Appendix D.5. Analogously, and to specifically treat the setting of Theorem 1, we introduce two subsets of E that depend on the design parameter $\beta \in (0,1)$.

$\tilde{E}_{\eta_\star}(\beta)$: For a given $\eta_\star \in E$, we let $\tilde{E}_{\eta_\star}(\beta)$ denote the set of initial points $\eta_0 \in E$ for which the sequence $(\eta_k) \in E^{\mathbb{N}}$ defined by (9), with step-sizes (γ_k) chosen as in (D2)-(i) or (ii), is such that $\lim \eta_k = \eta_\star$.

$\tilde{E}'_{\eta_\star}(\beta)$: We define $\tilde{E}'_{\eta_\star}(\beta)$ as the subset of $\tilde{E}_{\eta_\star}(\beta)$ on which the step-size sequence (γ_k) is eventually constant.

Note that, by (31) and (D2), the sequence $(\gamma_k)_{k\geq 0}$ is valued in the pair $\{\beta, 1\}$. These two sets are thus respectively included in the sets $E_{\eta_\star}(\{\beta,1\})$ and $E'_{\eta_\star}(\{\beta,1\})$ defined in Appendix D.5. For any $n \in \mathbb{N}$, we also let

$$\varkappa_n := \Pi_{\mathcal{Q},p}^{1-\alpha} + 2^{-n}.$$

For any $n \in \mathbb{N}$, Theorem 4 implies that the set $\mathsf{K}_{\varkappa_n} = \{\eta \in E \ : \ \ell_\alpha(\eta) \geq \varkappa_n\}$ is compact, and it is a subset of E by definition. Theorem 2 guarantees that $(\eta_k)_{k\geq 0} \in \mathsf{K}_{\varkappa_n}^{\mathbb{N}}$ as soon as $\eta_0 \in \mathsf{K}_{\varkappa_n}$. Then, Theorem 3 applies and its first assertion (i) holds, i.e, the limit points of (η_k) are in $\mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}_{\varkappa_n}$. Corollary 1 guarantees that this set is finite, hence assertion (ii) in Theorem 3 also holds, that is, the sequence (η_k) admits a limit in $\mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}_{\varkappa_n}$.

By (D1) and (15), the event $\{\ell_\alpha(\eta_0) \geq \varkappa_n\}$ holds with $\mathbb{P}$-probability 1 for some $n \in \mathbb{N}$. For this n, the conclusions above hold with probability 1. For later reference, we formalize this statement as

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \bigcup_{\eta_\star \in \mathrm{Fix}(\mathcal{M})\cap \mathsf{K}_{\varkappa_n}} \tilde{E}_{\eta_\star}(\beta)\right) = 1. \tag{36}$$

Lemma 11 gives the inclusion $\mathrm{Fix}(\mathcal{M}) \subset \mathcal{D}_1$. Since $\mathcal{D}_1$ is an open set, any sequence (η_k) that converges to a fixed point satisfies $\eta_k \in \mathcal{D}_1$ for all sufficiently large k. In that case, the design guidelines (D2) lead to a constant step-size γ_k for all sufficently large k. With (D2)-(i), the corresponding constant is β, and with (D2)-(ii) we have $\gamma_k = 1$ for all sufficiently large k. This proves that Assertions (ii) and (iv) in Theorem 1 hold with $\mathbb{P}$-probability 1. These conclusions about the step-size sequence imply that (36) can be rewritten as

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \bigcup_{\eta_\star \in \mathrm{Fix}(\mathcal{M})\cap \mathsf{K}_{\varkappa_n}} \tilde{E}'_{\eta_\star}(\beta)\right) = 1. \tag{37}$$

Now, recall that under (A3), the Hessian matrix $\mathbf{H}_{\ell_\alpha}(\eta_\star)$ has no null eigenvalues. By Lemma 10-$(viii)$, this means that $\rho(\eta_\star) \neq 1$ for all $\eta_\star \in \mathrm{Fix}(\mathcal{M})$. Then, according to Assertions (ii) and $(viii)$ in Lemma 10, there are only two cases. For a given $\eta_\star \in \mathrm{Fix}(\mathcal{M})$, either $\rho(\eta_\star) < 1$ and $\mathbf{H}_{\ell_\alpha}(\eta_\star)$ has negative eigenvalues only, or $\rho(\eta_\star) > 1$ and the matrix $\mathbf{H}_{\ell_\alpha}(\eta_\star)$ has at least one positive eigenvalue. We deduce that $\mathrm{Fix}(\mathcal{M})$ can be decomposed as

$$\mathrm{Fix}(\mathcal{M}) = B_1 \sqcup B_2,$$

where we define

$$\begin{aligned} B_1 &:= \mathrm{Fix}(\mathcal{M}) \cap \left\{\eta \in E \ : \ \mathrm{Sp}(\mathbf{H}_{\ell_\alpha}(\eta)) \subset \mathbb{R}_-^*\right\}, \\ B_2 &:= \mathrm{Fix}(\mathcal{M}) \cap \left\{\eta \in E \ : \ \mathrm{Sp}(\mathbf{H}_{\ell_\alpha}(\eta)) \cap \mathbb{R}_+^* \neq \emptyset\right\}. \end{aligned}$$

We are going to conclude by proving that for all $\eta_\star \in B_2$, we have $\mathbb{P}(\tilde{E}'_{\eta_\star}(\beta)) = 0$. This will imply $\rho(\eta_\star) < 1$, which is the first assertion (i) in Theorem 1, and will allow us to apply Theorem 3-(iii) to get the third assertion (iii) in Theorem 1.

Let $\eta_\star \in B_2$, i.e., the Hessian matrix $\mathbf{H}_{\ell_\alpha}(\eta_\star)$ has at least one positive eigenvalue. Lemma 14 states that the set $E'_{\eta_\star}(\{\beta,1\})$ is contained in a countable union of manifolds of dimension at most $d-1$. Then, the design assumption (D1) ensures that $\mathbb{P}(E'_{\eta_\star}(\{\beta,1\})) = 0$. We deduce that $\mathbb{P}(\tilde{E}'_{\eta_\star}(\beta)) = 0$, since $\tilde{E}'_{\eta_\star}(\beta) \subset E'_{\eta_\star}(\{\beta,1\})$. By plugging this into (37) and using the fact that $\mathrm{Fix}(\mathcal{M}) \cap \mathsf{K}_{\varkappa_n}$ is finite for all $n \in \mathbb{N}$ (*c.f.* Corollary 1), we obtain

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \bigcup_{\eta_\star \in B_1\cap \mathsf{K}_{\varkappa_n}} \tilde{E}'_{\eta_\star}(\beta)\right) = 1. \tag{38}$$

In particular, this proves Theorem 1-(i) and implies that the sequence (η_k) converges to a local maximizer of ℓ_α with $\mathbb{P}$-probability 1. Moreover, if η_0 that belongs to the set in (38), we can apply Assertion (iii) of Theorem 3. Therefore, Assertion (iii) of Theorem 1 holds with probability 1. □

E ADDITIONAL EXPERIMENTAL DETAILS

E.1 On the choice of the sample size

Setting an insufficient number of samples typically results in convergence failure, manifesting as exploding or collapsing parameters, most of the time along one dimension. Consequently, strategies involving a gradual increase in sample size from a small initial value, such as the one proposed in [Burroni et al., 2024], may not be as effective within our framework due to the high probability of encountering such failure cases for small sample sizes.

While a heuristic for determining an appropriate sample size on real-world datasets could be $2\,000 \times d$ (this seems to work up to at least $d \sim 100$), our experiments suggest that even small sample sizes can give decent results, at least on synthetic data.

Figure 4 shows the test accuracy for Bayesian Logistic Regression on a synthetic dataset with 10 features, i.e., the parameter space has dimension $d = 20$. The number of data points is set to $2\,000$. We take $\alpha = 0.5$, $\beta = 0.1$ in (D2)-(i) and set the maximum number of global iterations (i.e., the number of times we compute $\widehat{\mathbf{m}}_{\alpha,n}$) to $T = 20$. The same hyperparameters are kept in Figure 5, where the experiment is ran on a dataset with 100 features (i.e., we optimize on a $d = 200$-dimensional space) and $5\,000$ samples.

We set the learning rate for Adam to 0.05, and the maximum number of iterations to 1.000. For Newton's method, we use backtracking line search to improve the overall stability and set the maximum number of iterations to 15. Quite remarkably, Adam does not run particularly slower as the sample size incrases (unlike in Newton's method, we do not compute the inverse Hessian matrix of the log-partition function). It also performs better for all sample sizes on this synthetic dataset, which is different than what we observed on the California Housing dataset.

The main takeaway of this experiment is that a larger sample size allows to dramatically improve the quality of the solution (e.g., for Adam, the median best score with $32\,768$ samples is 5 percentage points above the best score with 256 samples, at equal runtime).

E.2 Further discussion of the California Housing dataset experiment

As shown in Figure 2, the iterations exhibited high variance in their results. With Newton's method, some runs showed minimal improvement in the value of $\widehat{\ell}_\alpha$, however the train and test accuracies always improved, though not always to satisfactory values (some runs ended with less than 50% accuracy). This further emphasizes the fact that, like many other VI methods, the proposed SAA algorithm is quite sensitive to the initial conditions, and that its efficiency should be utilized to perform as many runs as possible.

Note that the third boxplot in Figure 2 shows an overall lower ℓ_α score compared to earlier iterations. This behavior is not abnormal: our framework does not guarantee a monotonic improvement in the alpha-divergence at each iteration of the inner loop. However, the criterion will eventually improve if the optimizer converges sufficiently close to the true global minimum (see Appendix E.5). Given Newton's speed advantage, a sensible strategy would be to run SAA with Newton's method multiple times to then select the parameter yielding the best performance. We saw in Appendix E.1 that Newton's method can be outperformed by other optimizers, even when the increased number of possible runs statistically favors this approach. However, results on the California Housing dataset, where Adam struggled compared to Newton's method, suggest that the choice of the optimizer may be problem-dependent and may require additional hyperparameter tuning.

An important consideration not addressed in this analysis is the influence of the hyperparameter α on the quality of the resulting approximation. While $\alpha = 0.5$ was used in this experiment, we have anecdotally noticed that values of α closer to 1 may grant better results for Bayesian Logistic Regression, but this specific matter requires further investigation.

E.3 Detailed algorithm

In Section 3, we provide a simplified version of our algorithm to avoid overloading the introduction of the practical method with technical details. In Algorithm 1, we assume that the step-size γ_k equals 1 for all $k \in \mathbb{N}$. In other words, we completely omit the second design assumption (D2). The algorithm below describes our complete SAA method, in a way that closely resembles our implementation. We discuss design choices after.

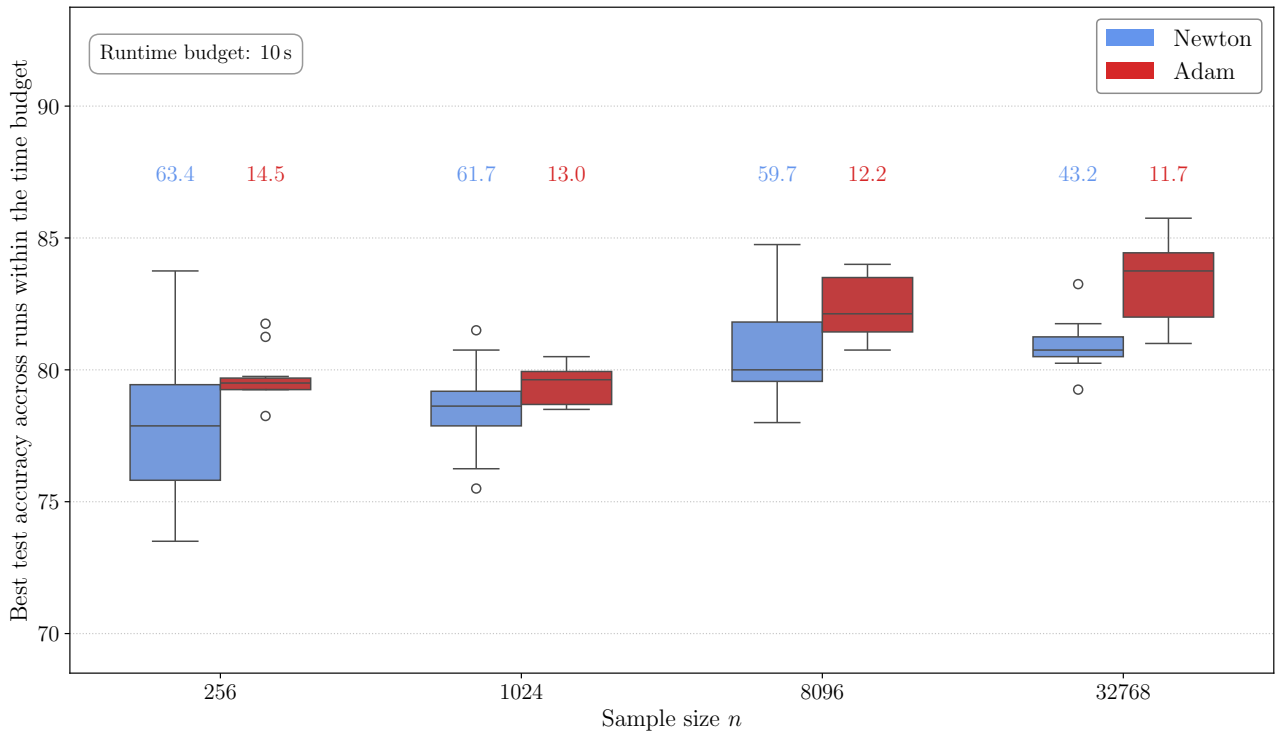

Figure 4: Comparing the effect of different sample sizes using Adam or Newton as the solver for η_{new} in Algorithm 1. For each sample size, we give the algorithms a time budget to run with as many random initializations as possible. Once the runtime budget is exhausted, the parameter with the best test accuracy (shown on the y-axis) is kept. Boxplots show results accross 10 repetitions of this experiment. The numbers above the boxplots show the average number of runs performed by each algorithm within the time limit.

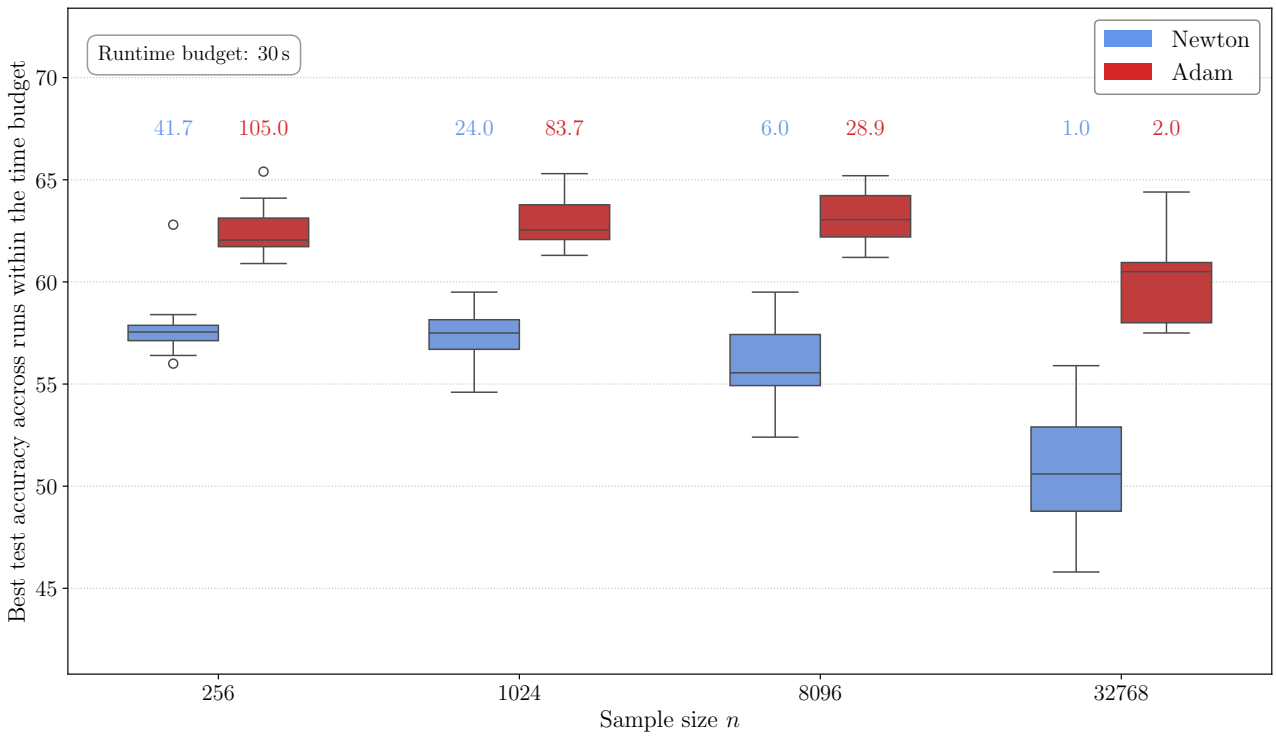

Figure 5: We keep the experimental setting from Figure 4, but $d = 200$ in this example. This example shows that it can be more beneficial to set a lower sample size (and thus degrade the quality of the approximation) to get better initializations. Adam clearly outspeeds Newton's method on this 200-dimensional problem.

Algorithm 3 Sample Average Approximation procedure with improved stability and efficiency

Require: Hyperparameter α, target p, variational family $\mathcal{Q}$, distribution $\mathbb{P}$ on E, number of samples n, number of iterations T, step-size scaler β, runtime budget B, initialization quality threshold Δ, maximum regularization attempts N_{reg}, solution quality criterion C.

Ensure: Approximate maximizer of ℓ_α and/or the criterion C (to maximize, e.g., accuracy).

Set runtime counter to r $\leftarrow 0$ and best parameter $\eta_{\text{best}} \leftarrow$ `None`
while r $<$ B **do**
  Initialize $\eta_\diamond \sim \mathbb{P}$ and $\eta \leftarrow \eta_\diamond$
  Draw i.i.d. samples $y_1, \dots, y_n \sim q_{\eta_\diamond}$
  Compute $\ell \leftarrow \widehat{\ell}_\alpha(\eta)$ with (12)
  Compute $\Pi \leftarrow \Pi_{\widehat{\mathcal{Q}}_n, p}$ with (17)
  if $\Pi - \ell < \Delta$ **then**
    for $k = 1$ **to** T **do**
      Compute $\mathbf{m}_{\text{target}} \leftarrow \widehat{\mathbf{m}}_{\alpha,n}(\eta)$ with (12) and $\mathbf{m}_{\text{current}} \leftarrow \nabla \widehat{A}(\eta)$
      Set $\gamma \leftarrow 1$ and $N = 0$
      while solver failing and $N < N_{\text{reg}}$ **do**
        Update $\gamma \leftarrow \gamma \times \beta$ and $N \leftarrow N + 1$
        Update $\mathbf{m} \leftarrow \gamma \mathbf{m}_{\text{target}} + (1 - \gamma)\mathbf{m}_{\text{current}}$
        Try to solve $\eta_{\text{new}} = \arg\max_{\eta'} \left[\langle \eta', \mathbf{m} \rangle - \widehat{A}_n(\eta') \right]$
      end while
      if solver succeeded **then**
        Update $\eta \leftarrow \eta_{\text{new}}$
      else
        Break the **for** loop
      end if
      Update total runtime r and $\eta_{\text{best}} \leftarrow \eta$ if $C(\eta) > C(\eta_{\text{best}})$
    end for
  end if
end while
return η_{best}

In practice, two concurrent phenomena exert opposing influences on the algorithm's behavior. First, it is highly unlikely that we draw $\eta_\diamond \in E$ and independent samples $y_1, \dots, y_n$ such that the "geometric" mean $\widehat{\mathbf{m}}_{\alpha,n}$ belongs to the boundary of the set of realizable means (from (12), we see that this set is the convex hull of $(S(y_1), \dots, S(y_n))$). On the other hand, optimization algorithms used to find η_{new} can have trouble converging due to numerical issues. For instance in the case of Newton's method, the Hessian matrix of the log-partition may be poorly conditioned. This is why, despite theoretical arguments supporting the choice of $\gamma_k = 1$ for all $k \in \mathbb{N}$, we propose a more complicated design guideline in (D2).

However, keep in mind that like most variational inference methods, the proposed algorithm is very sensitive to the initial conditions. Hence, it is probably not worth to try and optimize an obviously unfavorable initial parameter. To quickly assess the quality of the initial parameter, one can use condition (15). Recall that the constant $\Pi_{\widehat{\mathcal{Q}}_n,p}$ can be computed easily by the formula (17), hence it is possible to compare $\widehat{\ell}_{\alpha,n}(\eta_0)$ and $\Pi_{\widehat{\mathcal{Q}}_n,p}$ before running the costlier optimization process. It is quite rare, however, that the event in (15) is realized (the conditions for $\lim_n \Pi_{\widehat{\mathcal{Q}}_n,p} = 0$ a.s. are not always realistic, see the remark at the end of Appendix B.5). Setting a maximum threshold is certainly a more viable solution.

E.4 On the necessity of inverting $\nabla \widehat{A}_n$

We consider a multivariate Gaussian family with known covariance $\mathcal{Q} = \{q_\eta\}$ in a d-dimensional space. Let A denote the associated log-partition function, and consider some natural parameter $\eta_\diamond$ along with n i.i.d. samples from $q_{\eta_\diamond}$. These samples allow us to compute $\widehat{A}_n(\eta)$ for any η in the parameter space. Figures 6 to 9 show boxplots of the value of $\|\nabla \widehat{A}_n(\eta) - \nabla A(\eta)\|_2$ for different values of n, and for $d \in \{10, 60\}$. Each time, two cases are considered: one where η is close to $\eta_\diamond$ in norm (easy case), and one where η is far away from $\eta_\diamond$ (hard case). Effectively, we add a random noise of a given magnitude to the mean of the Gaussian density q_η. For each sample size, we draw 512 sets of samples and make the boxplot from these 512 runs.

In all cases, there is a significant difference between the gradients of A and $\widehat{A}_n$, which does not vanish as the number of samples increases. While this implies no fundamental issue with our framework, these observations stress the necessity to invert $\nabla \widehat{A}_n$ using an optimization algorithm, and that replacing this step with an inversion of ∇A will result in a completely different algorithm, whose behavior is unpredictable.

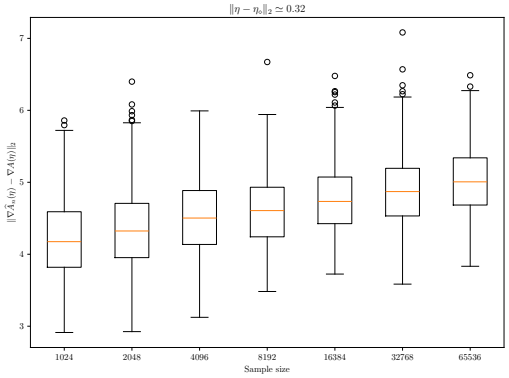

Figure 6: Easy case, $d = 10$.

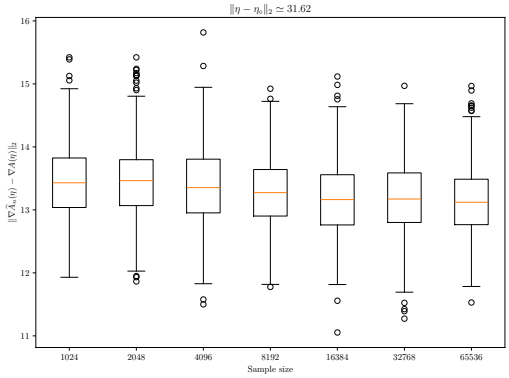

Figure 7: Hard case, $d = 10$.

E.5 Newton's method and Adam for multivariate Gaussian approximations

We first consider the problem of approximating a 10-dimensional multivariate Gaussian distribution with dense covariance matrix by another multivariate Gaussian with dense covariance matrix. The problem has dimension $d + d^2 = 110$. We use Newton's method in the inner optimization loop. Figure 10 shows the evolution of $\log \widehat{\ell}_\alpha$ and ℓ_α (available in closed form) accross iterations, including Newton's algorithm iterations. The main updates (i.e., the points at which Newton's method reaches a stopping criterion) are represented by crosses, and the corresponding iterations are shown on the x-axis. While the Newton iterations do not consistently improve $\widehat{\ell}_\alpha$, the criterion noticeably increases with each main update, as predicted by our theoretical results.

Figure 11 shows the optimization gap at each main update, that is, at the end of each Newton loop. We notice

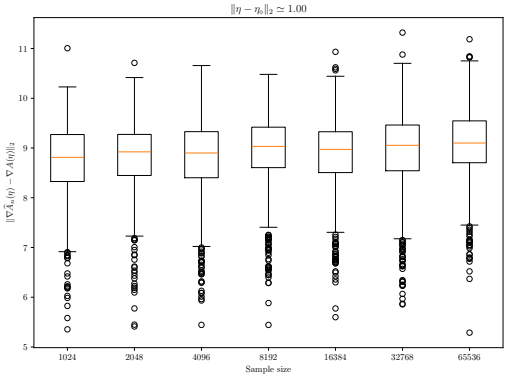

Figure 8: Easy case, $d = 100$.

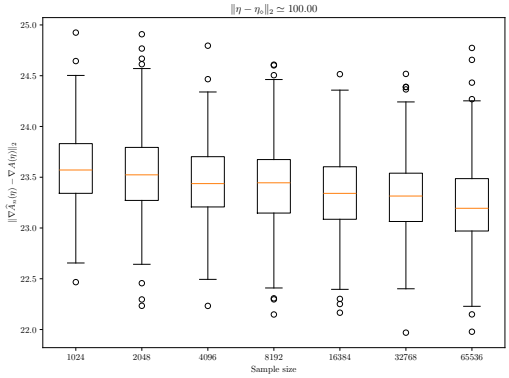

Figure 9: Hard case, $d = 100$.

that the tolerance threshold is consistently met for each of the 30 main updates.

Figures 12 and 13 show the same metrics in a slightly different setup: the approximating density is a multivariate Gaussian with diagonal covariance matrix, and the target is still a multivariate Gaussian with dense covariance matrix (this explains why the algorithm fails to reach $\log \widehat{\ell}_\alpha(\eta) = 0$, which indicates $p = q_\eta$ ν-almost surely: this can only happen if the target belongs to the variational family). We use Adam as the solver. Noticeably, Adam often fails to meet the tolerance stopping criterion (though the approximations it yields are not catastrophic). This can be fixed with a proper step-size strategy, but that means additional hyperparameter tuning.

Note that in this experiment, we set the sample size to $n = 50\,000$. The algorithm managed to converge with sample sizes as low as $5\,000$, and consistently converged with just $10\,000$ samples.

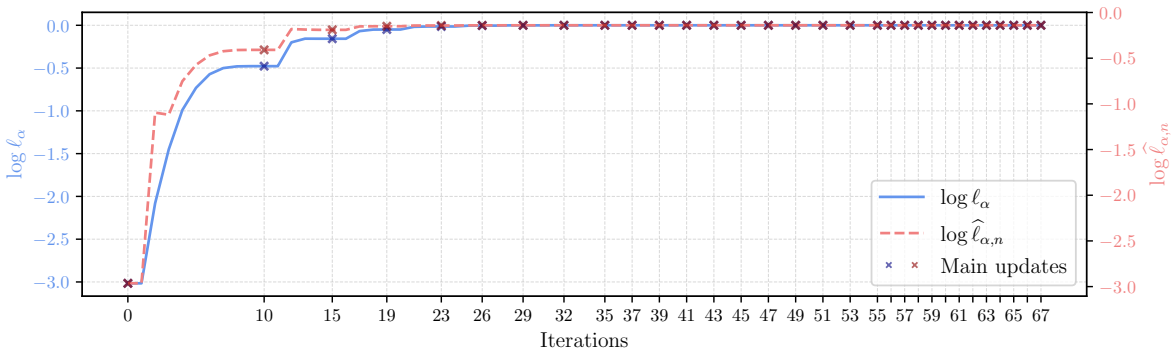

Figure 10: Approximating a 10-dimensional multivariate Gaussian density with dense covariance by another multivariate Gaussian density with dense covariance. Using Newton's algorithm as the solver. The fact that $\log \ell_\alpha$ converges to a value close to 0 indicates that the target was matched very closely.

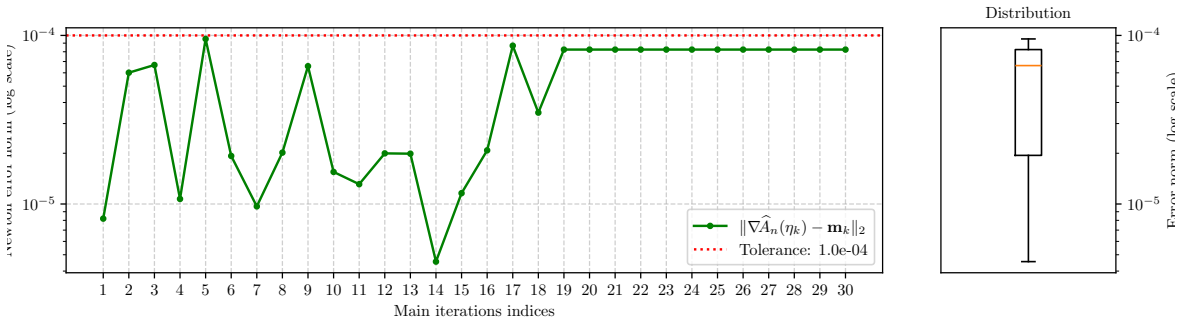

Figure 11: Newton optimization gap in the dense-dense scenario.

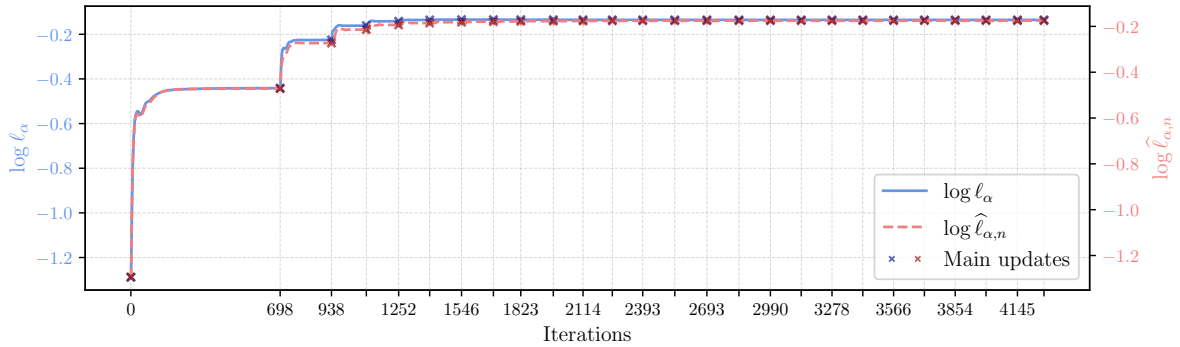

Figure 12: Approximating a 10-dimensional multivariate Gaussian density with dense covariance by a multivariate Gaussian density with diagonal covariance. Using Adam as the solver. Here, $\log \ell_\alpha$ cannot converge to 0 since the target does not belong to the variational family.

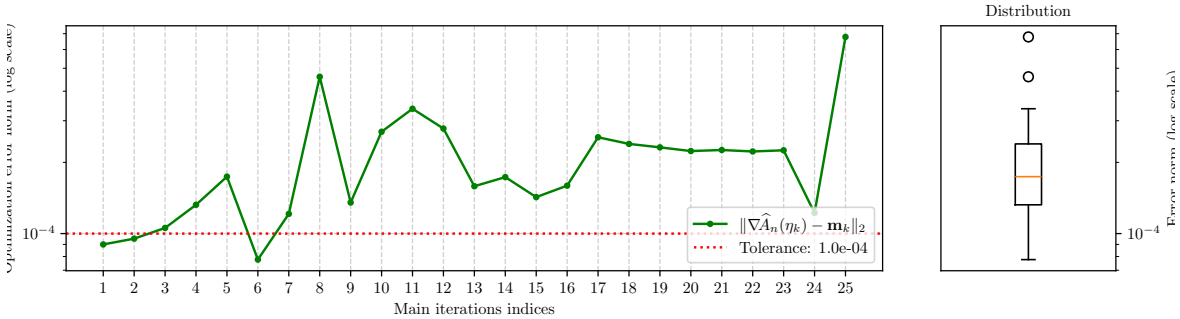

Figure 13: Adam optimization gap in the diagonal-dense scenario.