Convergence of Projected Stochastic Natural Gradient Variational Inference for Various Step Size and Sample or Batch Size Schedules

Thomas Guilmeau **Hadrien Hendrikx** **Florence Forbes**
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Abstract

Stochastic natural gradient variational inference (NGVI) is a popular and efficient algorithm for Bayesian inference. Despite empirical success, the convergence of this method is still not fully understood. In this work, we define and study a projected stochastic NGVI when variational distributions form an exponential family. Stochasticity arises when either gradients are intractable expectations or large sums. We prove new non-asymptotic convergence results for combinations of constant or decreasing step sizes and constant or increasing sample/batch sizes. When all hyperparameters are fixed, NGVI is shown to converge geometrically to a neighborhood of the optimum, while we establish convergence to the optimum with rates of the form $\mathcal{O}\left(\frac{1}{T^\rho}\right)$, possibly with $\rho \geq 1$, for all other combinations of step size and sample/batch size schedules. These rates apply when the target posterior distribution is close in some sense to the considered exponential family. Our theoretical results extend existing NGVI and stochastic optimization results and provide more flexibility to adjust, in a principled way, step sizes and sample/batch sizes in order to meet speed, resources, or accuracy constraints.

1 INTRODUCTION

In Bayesian statistics, posterior distributions are often intractable and require sophisticated inference techniques. Variational inference (VI) methods, which approximate posterior distributions by optimizing over a family of tractable densities the Kullback–Leibler divergence from the true posterior, have emerged as an efficient alternative to Markov chain Monte Carlo (Blei et al., 2017; Zhang et al., 2019). VI algorithms can be easily deployed in many settings (Ranganath et al., 2014), and the properties of the optimal solution are increasingly studied from many perspectives (Wang and Blei, 2018; Alquier and Ridgway, 2020; Margossian and Saul, 2025).

In order to solve the VI optimization problem, natural gradient descent, initially proposed by Amari (1998), has been used successfully in many settings such as latent Dirichlet allocation with very large datasets (Hoffman et al., 2013), Bayesian neural networks (Khan and Nielsen, 2018), or problems with discrete latent variables (Ji et al., 2021). Natural gradient descent has also been used in black-box optimization (Ollivier et al., 2017), machine learning (Martens, 2020), or in the context of the so-called Bayesian learning rule (Khan and Rue, 2023). Natural gradient descent preconditions the standard gradient by the inverse Fisher information matrix of the approximating distribution. The intuition behind its good performance is that it approaches Newton's method (Martens, 2020). In the case of exponential families (Barndorff-Nielsen, 2014), natural gradient descent is equivalent to mirror descent (Raskutti and Mukherjee, 2015; Wu and Gardner, 2024), whose ability to better tailor the geometry of the objective function has been shown by e.g., Bauschke et al. (2017); Lu et al. (2018).

Despite these empirical good performance and intuitions behind its working, the convergence of stochastic NGVI is still poorly understood. In the deterministic setting, let us mention the works of Kumar et al. (2025); Godichon-Baggioni et al. (2025). In the stochastic setting, Hoffman et al. (2013) claim that convergence can be established through results by Bottou (1999), which may not apply without further assumptions. Khan et al. (2016) use the equivalence between NGVI and mirror descent when exponential families are considered to derive a convergence rate to approximate stationarity under a strong

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

convexity assumption which may fail in some parts of the search space, even for Gaussian distributions (Guilmeau et al., 2025). Still in the context of exponential families, Wu and Gardner (2024) leverage a notion of variance from Hanzely and Richtárik (2021) to prove a $\mathcal{O}\left(\frac{1}{T}\right)$ convergence rate when the posterior distribution belongs to the considered exponential family, and they formulate this variance-like quantity only in a conjugate Bayesian linear regression setting with stochasticity coming from subsampling the data. The same notion of variance is also used by Sun et al. (2025) to derive convergence rates in possibly non-convex settings but restricted to mean-field Gaussian approximating families.

In this work, we also use the equivalence between natural gradient descent and mirror descent. We use and refine novel analysis techniques for stochastic mirror descent from Hendrikx (2024) to improve the generality and applicability of existing results for stochastic NGVI. In this endeavor, we make the following contributions. (**1**) We propose a projection step in a geometry that is compatible with the mirror step with novel non-asymptotic convergence rates for the resulting projected stochastic mirror descent algorithm, which may be of independent interest. We establish geometric convergence to a neighborhood of the minimizer for constant step sizes and gradient estimators computed with fixed sample/batch sizes. We establish $\mathcal{O}\left(\frac{1}{T^\rho}\right)$ convergence to the minimizer for either fixed step sizes and increasing sample/batch sizes or diminishing step sizes and fixed or increasing sample/batch sizes, with possibly $\rho \geq 1$. (**2**) We exhibit a new condition on the posterior distribution, which ensures that our general convergence rates apply to our projected stochastic NGVI algorithm, while allowing posterior distributions that do not belong to the considered exponential family. (**3**) Our results involve a variance-like term proposed by Hendrikx (2024), which we upper-bound in case of estimators based on samples from the approximating distribution or on subsampled data batches, thus providing a precise dependence on the step size and sample/batch size in our results. It follows an increased number of possible strategies to jointly schedule step and sample/batch sizes, depending on the desired convergence speed, accuracy, and data-efficiency.

After specifying some background and our stochastic NGVI algorithm in Sections 2 & 3, our non-asymptotic convergence rates for different choices of step sizes and sample/batch sizes are presented in Section 4, with practical and numerical illustrations in Sections 5 & 7. Related work is discussed in Section 6. Proofs and additional results are postponed to the Appendix.

2 BACKGROUND

Notation For a measurable space $\mathcal{X}$ and a measure ν on $\mathcal{X}$, $\mathcal{P}(\mathcal{X}, \nu)$ denotes the set of probability densities p with respect to ν on $\mathcal{X}$ and $\mathbb{E}_p$ the expectation with respect to p. $\mathcal{H}$ denotes a finite-dimensional Hilbert space with scalar product denoted by $\langle \cdot, \cdot \rangle$. $\mathbb{S}^d$ (resp. $\mathbb{S}^d_{>0}$) denotes the space of symmetric (resp. positive definite) $d \times d$ matrices. $\mathcal{N}(\mu, \Sigma)$ is the Gaussian distribution with mean μ and covariance matrix Σ while $\mathcal{U}_M$ is the uniform distribution on $[\![1, M]\!]$. For a proper function $h : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ with domain $\operatorname{dom} h$ which is differentiable on $\operatorname{int} \operatorname{dom} h$ with gradient ∇h, d_h denotes the following quantity

$$d_h(\omega, \omega') = h(\omega) - h(\omega') - \langle \nabla h(\omega'), \omega - \omega' \rangle$$

for all $(\omega, \omega') \in \mathcal{H} \times \operatorname{int} \operatorname{dom} h$. When h is strictly convex then $d_h(\omega, \omega') \geq 0$ with equality if and only if $\omega = \omega'$ and $d_h$is the induced **Bregman divergence**.

Variational inference Variational inference efficiently approximates intractable Bayesian posterior distributions resulting from updating prior belief $X \sim p_0$ upon observing Y. We denote by $\pi \in \mathcal{P}(\mathcal{X}, \nu)$ the density of $X|Y$, given by Bayes rule, $\pi(x) \propto p(y|x)p_0(x)$. VI aims at approximating π by a density from a family $\mathcal{Q} \subset \mathcal{P}(\mathcal{X}, \nu)$. This is done by solving

$$\min_{q \in \mathcal{Q}} \mathrm{KL}(q, \pi), \tag{1}$$

where $\mathrm{KL}(p_1, p_2) = \mathbb{E}_{p_1}[\log p_1(X)] - \mathbb{E}_{p_1}[\log p_2(X)]$, for $p_1, p_2 \in \mathcal{P}(\mathcal{X}, \nu)$, is the Kullback-Leibler divergence.

When $\pi \in \mathcal{Q}$, it is possible to exactly recover π. However, in most settings, $\pi \notin \mathcal{Q}$.

Exponential family We will consider $\mathcal{Q}$ in problem (1) to be an exponential family (Brown, 1986; Barndorff-Nielsen, 2014).

Definition 1. *The exponential family with sufficient statistic $\Gamma : \mathcal{X} \to \mathcal{H}$ is the set $\mathcal{Q} \subset \mathcal{P}(\mathcal{X}, \nu)$ s.t. every $q \in \mathcal{Q}$ satisfies $q = q_\theta$ for some $\theta \in \operatorname{dom} A$ with*

$$q_\theta(x) = \exp(\langle \theta, \Gamma(x) \rangle - A(\theta)),\ \nu\text{-a.e.}, \tag{2}$$

where θ is the natural parameter and A is the log-partition function $A(\theta) = \log\left(\int \exp(\langle \theta, \Gamma(x) \rangle) \nu(dx)\right)$.

For $q_\theta \in \mathcal{Q}$, the associated **expectation parameter** ω is defined as the expected sufficient statistic

$$\omega = \mathbb{E}_{q_\theta}[\Gamma(X)].$$

For our proofs, we will need to assume that $\mathcal{Q}$ is steep. Steepness is defined in (Barndorff-Nielsen, 2014, Chapter 8) and recalled in Appendix A. It is implied by

$\operatorname{dom} A$ being open. For example, Gaussian distributions form a steep exponential family.

The log-partition function benefits from convex analysis properties. Introducing its **convex conjugate**

$$A^*(\omega) = \sup_{\theta \in \mathcal{H}} \{\langle \theta, \omega \rangle - A(\theta)\},$$

the following properties hold.

Proposition 1. *Suppose that* $\operatorname{int} \operatorname{dom} A \neq \emptyset$. *Then,*

(*i*) *A is proper, lower-semicontinuous, strictly convex, and differentiable on* $\operatorname{int} \operatorname{dom} A$. *The gradient of A links the expectation and natural parameters with* $\omega = \mathbb{E}_{q_\theta}[\Gamma(X)] = \nabla A(\theta)$.

The Fisher information matrix of q_θ takes the form $\mathbb{E}_{q_\theta}[\nabla \log q_\theta(X)(\nabla \log q_\theta(X))^\top] = \nabla^2 A(\theta)$.

Furthermore, if $\mathcal{Q}$ is minimal and steep, then

(*ii*) *A^* is proper, lower semi-continuous, strictly convex, differentiable on* $\operatorname{int} \operatorname{dom} A^*$. *If* $\omega \in \operatorname{int} \operatorname{dom} A^*$ *with* $\omega = \nabla A(\theta)$, *then* $A^*(\omega) = \mathbb{E}_{q_\theta}[\log(q_\theta(X))]$.

(*iii*) *∇A is a bijection from* $\operatorname{int} \operatorname{dom} A$ *to* $\operatorname{int} \operatorname{dom} A^*$, *its inverse is ∇A^* and*

$$\omega = \nabla A(\theta) \iff \theta = \nabla A^*(\omega).$$

(*iv*) *For every* $\theta, \theta' \in \operatorname{int} \operatorname{dom} A$ *and their respective expectation parameters* $\omega = \nabla A(\theta), \omega' = \nabla A(\theta')$,

$$d_A(\theta, \theta') = d_{A^*}(\omega', \omega) = KL(q_{\theta'}, q_\theta).$$

Proposition 1 implies that any density in the exponential family $\mathcal{Q}$ can be equivalently represented by q_θ or q_ω, with $\omega = \nabla A(\theta)$ or equivalently $\theta = \nabla A^*(\omega)$, motivating the following definitions.

Definition 2. *Consider an exponential family $\mathcal{Q}$. Then, we define the functions f_π^P and f_π^D as*

$$f_\pi^P(\theta) = KL(q_\theta, \pi), \ \forall \theta \in \operatorname{dom} A,$$
$$f_\pi^D(\omega) = KL(q_\omega, \pi), \ \forall \omega \in \operatorname{dom} A^*.$$

Remark that $f_\pi^P \circ \nabla A^* = f_\pi^D$ while $f_\pi^D \circ \nabla A = f_\pi^P$.

Natural gradient variational inference Natural gradient descent is a preconditioned gradient descent where the conditioning is done by the Fisher information matrix of the current distribution q_θ (Amari, 1998). When q_θ is in an exponential family $\mathcal{Q}$, Proposition 1 (i) states that the Fisher information matrix at q_θ is $\nabla^2 A(\theta)$. To solve (1), or from Definition 2, to equivalently minimize f_π^P on $\operatorname{dom} A$ or f_π^D on $\operatorname{dom} A^*$, a natural gradient descent update for f_π^P, with step size $\eta > 0$, updates $\theta \in \operatorname{dom} A$ to θ_+ set to

$$\theta_+ = \theta - \eta \left(\nabla^2 A(\theta)\right)^{-1} \nabla f_\pi^P(\theta). \tag{3}$$

In addition, $\nabla f_\pi^P(\theta) = \nabla^2 A(\theta) \nabla f_\pi^D(\nabla A(\theta))$ so that (3) is equivalent to

$$\theta_+ = \theta - \eta \nabla f_\pi^D(\omega), \quad \text{where } \omega = \nabla A(\theta). \tag{4}$$

The update (4) is a mirror descent update with mirror map ∇A^*: $\nabla A^*(\omega_+) = \nabla A^*(\omega) - \eta \nabla f_\pi^D(\omega)$ with $\omega_+ = \nabla A(\theta_+)$ (see also (Raskutti and Mukherjee, 2015, Theorem 1) and Wu and Gardner (2024)).

Although (4) alleviates the need to invert a Hessian matrix at each iteration, we still need to compute gradients $\nabla f_\pi^D(\omega)$. From Proposition 1 (ii), it comes

$$f_\pi^D(\omega) = A^*(\omega) - \mathbb{E}_{q_\omega}[\log \pi(X)], \tag{5}$$

so that using (iii), (4) writes as

$$\theta_+ = (1 - \eta)\theta + \eta \nabla \mathbb{E}_{q_\omega}[\log \pi(X)]. \tag{6}$$

The real challenge is then to compute $\nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$, which has been approached from several perspectives. There are two fundamental situations.

1) Monte Carlo approximation: The gradient $\nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$ is intractable but can be approximated by samples from q_ω. Let us mention the gradient estimators based on Fisher's identity used by Ranganath et al. (2014); Ji et al. (2021), gradient estimators based on reparametrization (Domke et al., 2023; Kim et al., 2023), or in the case of Gaussian distributions, gradient estimators based on (Bonnet, 1964; Price, 1958) used by Rezende et al. (2014); Khan and Rue (2023).

2) Data subsampling: The gradient is a finite sum of tractable terms but the sum is too large to be computed. Data subsampling consists in approximating the gradient by only considering a randomly chosen subset of terms appearing in the sum, referred to as a batch. See for instance (Hoffman et al., 2013).

Bregman projection and stochastic mirror descent We outlined above the link between natural gradient descent and mirror descent, which is itself related to the geometry induced by Bregman divergences (Bauschke et al., 2017; Lu et al., 2018). Motivated by these links, we recall some useful notions from the literature of optimization in Bregman geometries.

In this part, we consider a Legendre function h. If h has an open domain, is differentiable and is strictly convex, then it is Legendre, see Appendix B for more details. The Bregman projection below associated to d_h will be useful to enforce constraints while respecting the geometry induced by natural gradient descent.

Definition 3. *For $C \subset \mathcal{H}$, the Bregman projection* proj_C^h *is defined for any* $\omega \in \operatorname{int}\operatorname{dom} h$ *as*

$$\mathrm{proj}_C^h(\omega) = \operatorname*{arg\,min}_{\omega' \in C} \{d_h(\omega', \omega)\}.$$

Proposition 2. *If C is closed and convex, such that $C \cap \operatorname{int}\operatorname{dom} h \neq \emptyset$ with h being Legendre, then* proj_C^h *is well-defined and single-valued on* $\operatorname{int}\operatorname{dom} h$*, and* $\mathrm{proj}_C^h(\omega) \in \operatorname{int}\operatorname{dom} h$ *for any* $\omega \in \operatorname{int}\operatorname{dom} h$.

We review recent results about the convergence of stochastic mirror descent algorithms. Consider the optimization problem of finding ω_* s.t.

$$f(\omega_*) = \min_{\omega \in D} f(\omega),$$

when f is convex and its gradient is unavailable, e.g., because it involves an expectation as in f_π^D. We assume $D \subset \operatorname{dom} h$. The stochastic mirror descent algorithm with mirror map ∇h and step size $\eta > 0$, updates ω to ω_+, assumed to be in D, as

$$\nabla h(\omega_+) = \nabla h(\omega) - \eta G^N(\omega), \tag{7}$$

where $G^N(\omega)$ is an unbiased estimator of $\nabla f(\omega)$ that depends on a parameter $N \in \mathbb{N}_{>0}$ representing e.g., the sample size or the batch size used to compute it. The convergence of such an algorithm is difficult to establish because of the interaction between the noise and the non-Euclidean geometry of the algorithm. In particular, the effect of the noise depends on the location of the current iterate.

In order to account for this specific geometry, several Bregman-based generalizations of Euclidean notions, such as variance, strong convexity, or smoothness have been introduced. A Bregman-based variance-like quantity is defined by Hendrikx (2024).

Definition 4. *Consider the step size $\eta > 0$ and $N \in \mathbb{N}_{>0}$. Then, we define the function $f_{\eta,N} : D \to \mathbb{R}$ as*

$$f_{\eta,N}(\omega) = f(\omega) - \frac{1}{\eta}\mathbb{E}\left[d_h(\omega, \omega_+)\right], \forall \omega \in D,$$

with ω_+ defined as in (7). The variance $\sigma^2_{\eta,N}$ is then

$$\sigma^2_{\eta,N} = \frac{1}{\eta} \sup_{\omega \in D} \{f(\omega_*) - f_{\eta,N}(\omega)\}.$$

It is smaller than previously proposed notions such as that of Hanzely and Richtárik (2021), and Hendrikx (2024, Proposition 2.1) states that $\sigma^2_{\eta,N} \geq 0$, $\forall \eta > 0$.

We recall the notions of relative strong convexity and relative smoothness, which generalize the notions of strong convexity and smoothness to the Bregman geometry (Bauschke et al., 2017; Lu et al., 2018).

Definition 5. *Let $m \in \mathbb{R}_{>0}$, f is said to be m-strongly-convex relatively to h if the following holds: $m\, d_h(\omega, \omega') \leq d_f(\omega, \omega')$, $\forall \omega, \omega' \in D$.*

Similarly, for $\ell \in \mathbb{R}_{>0}$, f is said to be ℓ-smooth relatively to h if $d_f(\omega, \omega') \leq \ell\, d_h(\omega, \omega')$, $\forall \omega, \omega' \in D$.

Relative smoothness allows to derive an upper-bound on $\sigma^2_{\eta,N}$, adapted from Hendrikx (2024, Prop. 2.4).

Proposition 3. *For $\omega \in D$, we define $\nabla h(\overline{\omega}_+) = \nabla h(\omega) - \eta \nabla f(\omega)$, which is the exact counterpart of ω_+ in (7). If f is ℓ-smooth relatively to h and $\eta \leq \frac{1}{\ell}$, then*

$$\sigma^2_{\eta,N} \leq \frac{1}{\eta^2} \sup_{\omega \in D} \{\mathbb{E}\left[d_h(\overline{\omega}_+, \omega_+)\right]\}.$$

Then, Hendrikx (2024, Theorem 3.1) states that if f is m-strongly-convex relatively to h, the sequence $\{\omega_t\}_{t \in \mathbb{N}}$ generated by iterating (7) satisfies at every $t \in \mathbb{N}$

$$\eta\left(\mathbb{E}\left[f_{\eta,N}(\omega_t)\right] - \inf_{\omega \in D} f_{\eta,N}(\omega)\right) + \mathbb{E}\left[d_h(\omega_*, \omega_{t+1})\right] \leq (1 - m\eta)^{t+1} d_h(\omega_*, \omega_0) + \frac{\eta \sigma^2_{\eta,N}}{m}.$$

When $\eta \leq 1/m$, we thus have geometric convergence to a small neighbourhood of ω_*, whose size is controlled by $\sigma^2_{\eta,N}$.

3 PROJECTED STOCHASTIC NGVI

We consider the constrained VI problem

$$\min_{\omega \in C \cap \operatorname{dom} A^*} f_\pi^D(\omega), \tag{8}$$

which is an instance of Problem (1) with approximating family $\{q_\omega \in \mathcal{Q}, \omega \in C \cap \operatorname{dom} A^*\}$, meaning that we restrict densities from $\mathcal{Q}$ to have parameters in C. For Gaussian distributions, C may for instance enforce covariance matrices with constrained eigenvalues.

3.1 Projected stochastic natural gradient

To solve (8), recall from Equation (6) that an idealised natural gradient update for minimizing f_π^D reads

$$\nabla A^*(\omega_+) = (1 - \eta)\nabla A^*(\omega) + \eta \nabla \mathbb{E}_{q_\omega}[\log \pi(X)],$$

where we need to approximate $\nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$. We suppose that we have an estimator $g^N(\omega)$ of $\nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$ where N is a sample or batch size. Using formulation (4), this amounts to approximate $\nabla f_\pi^D(\omega)$ by $G^N(\omega) = \nabla A^*(\omega) - g^N(\omega)$.

We propose the following Algorithm 1, a projected mirror descent algorithm, with possibly varying step

and sample/batch sizes, which alternates stochastic mirror descent steps as in (7) and Bregman projection on C as defined in Definition 3. We detail specific instances of g^N in Sections 5.1 and 5.2. Examples of constraint sets C which admit projection operators $\mathrm{proj}_C^{A^*}$ with closed form are provided in Appendix E.

Algorithm 1: Projected NGVI

Select step sizes $\{\eta_t\}_{t\in\mathbb{N}}$ in $(0,1]$, $\{N_t\}_{t\in\mathbb{N}}$ in $\mathbb{N}_{>0}$, and a starting point $\omega_0 \in \operatorname{int}\operatorname{dom} A^*$.

for $t \geq 0$ **do**

$$\nabla A^*((\omega_t)_+) = (1-\eta_t)\nabla A^*(\omega_t) + \eta_t g^{N_t}(\omega_t)$$
$$\omega_{t+1} = \mathrm{proj}_C^{A^*}((\omega_t)_+).$$

end

3.2 Assumptions

We describe below our main assumptions, which will hold throughout the rest of this work.

(A1) $\operatorname{int}\operatorname{dom} A \neq \emptyset$ and $\mathcal{Q}$ is minimal and steep.

(A2) C is closed, convex, and $C \cap \operatorname{int}\operatorname{dom} A^* \neq \emptyset$.

(A3) Iterates $\{\omega_t\}_{t\in\mathbb{N}}$ from Algorithm 1 satisfy $\omega_t \in \operatorname{int}\operatorname{dom} A^*$ almost surely for any $t \in \mathbb{N}$.

(A4) For any $\omega \in \operatorname{int}\operatorname{dom} A^*$ and $N \in \mathbb{N}_{>0}$, $\mathbb{E}\left[g^N(\omega)\right] = \nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$.

(A5) Problem (8) admits a unique solution $\omega_* \in C \cap \operatorname{int}\operatorname{dom} A^*$.

Assumptions (A1) and (A2) are mild assumptions on $\mathcal{Q}$ and C that allow us to benefit from Propositions 1 and 2. Assumption (A3) ensures that the iterates are well-posed. As shown in Appendix C, (A3) is satisfied if $g^N(\omega) \in \operatorname{int}\operatorname{dom} A$ almost surely for every $\omega \in \operatorname{int}\operatorname{dom} A^*$, $N \in \mathbb{N}_{>0}$, or if the step sizes $\{\eta_t\}_{t\in\mathbb{N}}$ are small enough. Assumption (A4) is a mild unbiasedness assumption on our estimators. Finally, Assumption (A5) requires the considered VI problem to have a unique solution that must belong to $\operatorname{int}\operatorname{dom} A^*$, e.g., excluding cases where the optimum is reached by a singular covariance in the Gaussian case. A sufficient condition for (A5) is provided in subsection 4.1. We stress that $C \cap \operatorname{int}\operatorname{dom} A^*$ will play the role of D in Section 2. In particular, the supremum in the definition of $\sigma^2_{\eta,N}$ will be taken over this set.

4 CONVERGENCE OF PROJECTED NGVI

In this section, we provide two novel general convergence rates for Algorithm 1 under different assumptions on the step sizes and sample/batch sizes. Both results require the relative strong convexity of f_π^D, as defined in Definition 5, and involve the variance-like quantity $\sigma^2_{\eta,N}$, specified in Definition 4. Although Algorithm 1 refers to our projected stochastic NGVI algorithm, we stress that these results apply in the general context of projected stochastic mirror descent.

Fixed step sizes Our first Proposition 4 assumes fixed $\eta_t \equiv \eta$. It states that Algorithm 1 with $N_t \equiv N$ produces iterates that geometrically converge to a Bregman ball around ω_*. If $N_t = (t+1)^\gamma$, our result implies that convergence to ω_* is guaranteed and controlled by a sum of geometric terms and a $\mathcal{O}\left(\frac{1}{T^\gamma}\right)$ term.

Proposition 4. *Assume f_π^D is m-strongly-convex relatively to A^*. Let $\{\omega_t\}_{t\in\mathbb{N}}$ be generated by Algorithm 1 with $\eta_t \equiv \eta \in (0, m^{-1}]$. If $N_t \equiv N \in \mathbb{N}_{>0}$, we have for any $T \in \mathbb{N}$ that*

$$\mathbb{E}\left[d_{A^*}(\omega_*, \omega_T)\right] \leq (1-m\eta)^T d_{A^*}(\omega_*, \omega_0) + \frac{\eta \sigma^2_{\eta,N}}{m}.$$

If there exists $V > 0$ s.t. $\sigma^2_{\eta,N} \leq \frac{V}{N}$ for any $\eta \in (0, m^{-1}]$, and $N_t = (t+1)^\gamma$, $\gamma > 0$, we have for any $T \in \mathbb{N}$ that

$$\mathbb{E}\left[d_{A^*}(\omega_*, \omega_T)\right] \leq (1-m\eta)^T d_{A^*}(\omega_*, \omega_0) + \frac{V}{m\eta}\left((1-m\eta)^{\frac{T+1}{2}} + \frac{2^\gamma}{T^\gamma}\right).$$

Decreasing step sizes Our second result in Proposition 5 is set for decreasing step sizes. It gives convergence rates for the exact convergence of the iterates to ω_*. If $N_t \equiv N$, this convergence is in $\mathcal{O}\left(\frac{1}{T}\right)$, but it can be made faster if the sample/batch size increases with iterations. For instance, if $N_t = (t+1)^\gamma$, for some $\gamma \in (0,1)$, then the convergence is in $\mathcal{O}\left(\frac{1}{T^{1+\gamma}}\right)$.

Proposition 5. *Assume f_π^D is m-strongly-convex relatively to A^* and there exists $V > 0$ s.t. $\sigma^2_{\eta,N} \leq \frac{V}{N}$ for any $\eta \in (0, m^{-1}]$ and $N \in \mathbb{N}_{>0}$. Consider $\{\omega_t\}_{t\in\mathbb{N}}$ generated by Algorithm 1 with $\eta_t = \frac{1}{m(t/2+1)}$. Then, we have for any $T \in \mathbb{N}$ that*

$$\mathbb{E}\left[d_{A^*}(\omega_*, \omega_{T+1})\right] \leq \frac{4V}{m^2(T+2)(T+1)} \sum_{t=0}^{T} \frac{1}{N_t}.$$

4.1 Sufficient conditions

We now introduce a new condition ensuring (A5), that f_π^D is relatively strongly convex, and that allows to control $\sigma^2_{\eta,N}$. This condition requires π to belong to an exponential family that extends $\mathcal{Q}$ in a precise way, without necessarily requiring $\pi \in \mathcal{Q}$ as often done in previous work, e.g., (Wu and Gardner, 2024).

Definition 6. *The posterior π satisfies the linearly extended recoverability condition (LERC) with respect to $\mathcal{Q}$ if π belongs to an exponential family $\mathcal{Q}_\pi$ with sufficient statistic $\Gamma_\pi : \mathcal{X} \to \mathcal{H}_\pi$ for which there exists a linear operator $L_\pi : \mathcal{H} \to \mathcal{H}_\pi$ satisfying $\mathbb{E}_{q_\omega}[\Gamma_\pi(X)] = L_\pi \mathbb{E}_{q_\omega}[\Gamma(X)]$ for all $\omega \in \operatorname{dom} A^*$.*

The LERC extends the exact recoverability situation $\pi \in \mathcal{Q}$, as it is satisfied when $\mathcal{Q}_\pi = \mathcal{Q}$, but also in situations where $\mathcal{Q}_\pi$ is larger than $\mathcal{Q}$.

Example. *If $\mathcal{Q}$ is the family of centered Gaussian distributions with diagonal covariance matrices and $\pi = \mathcal{N}(\mu_\pi, \Sigma_\pi)$ with no restriction on μ_π and Σ_π, then π satisfies the LERC with respect to $\mathcal{Q}$. The linear operator is $L_\pi : \omega \to (0, \operatorname{diag}\omega)$, from $\mathbb{R}^d$ to $\mathbb{R}^d \times \mathbb{S}^d$, where for $\omega \in \mathbb{R}^d$, $\operatorname{diag}\omega$ denotes the diagonal matrix whose diagonal elements are the elements of ω. In the special case $\mu_\pi = 0$, this corresponds to a Gaussian mean-field variational approximation, as the correlation structure of π is removed. More details can be found in Appendix A.*

While generalizing exact recoverability, the LERC also has a direct simplifying impact on solving our VI problem, as shown in the next result.

Proposition 6. *Consider $\pi \in \mathcal{Q}_\pi$ that satisfies the LERC with respect to $\mathcal{Q}$ with θ_π the parameter so that $\pi = q_{\theta_\pi}$.*

(i) Suppose that Assumptions (A1) and (A2) hold. If C is compact, then Assumption (A5) is satisfied. Alternatively, if $L_\pi^\top \theta_\pi \in \operatorname{int}\operatorname{dom} A$, then (A5) is satisfied and $\omega_* = \operatorname{proj}_C^{A^*}(\nabla A(L_\pi^\top \theta_\pi))$.

(ii) Under Assumption (A1), f_π^D is 1-strongly-convex and 1-smooth relatively to A^*.

Proposition 6 (i) shows that the LERC helps ensuring that Assumption (A5) is satisfied, which amounts to showing that our VI problem (8) has a unique well-defined solution ω_*. If $L_\pi^\top \theta_\pi \in \operatorname{int}\operatorname{dom} A$, this solution can be further characterized. Without constraints ($C = \mathcal{H}$), we have in particular that ω_* is such that $\theta_* = \nabla A^*(\omega_*) = L_\pi^\top \theta_\pi$. However, this solution may still be hard to find due to the stochasticity of Algorithm 1. Fortunately, Proposition 6 (ii) establishes that the LERC also implies useful properties for the convergence of Algorithm 1. Indeed, having f_π^D smooth relatively to A^* allows to bound the variance $\sigma_{\eta,N}^2$ using Proposition 3 and the strong convexity of f_π^D relatively to A^* allows to obtain fast convergence rates in Proposition 4 and Proposition 5.

5 ILLUSTRATIONS

In this section, we illustrate into more details two settings in which our new convergence results apply, as stated in Propositions 7 and 8.

5.1 Sampling-based gradients over Gaussians

In this section, no specific assumption is made on π at first but $\mathcal{Q}$ is the family of Gaussian distributions. In this context, we consider Monte Carlo estimators of the gradients $\nabla_\omega \mathbb{E}_{q_\omega}[\log \pi(X)]$. Several estimators of this quantity have been used in VI, such as score-based estimators (Ranganath et al., 2014), or estimators based on the reparametrization trick (Kingma and Welling, 2014). However, we will use another estimator that leverages the fact that $\mathcal{Q}$ is a Gaussian family allowing to obtain better performance (Wu and Gardner, 2024).

For $\omega = (\mu, \Sigma + \mu\mu^\top)$, using theorems from Bonnet (1964) and Price (1958) and the chain rule (see also Khan and Rue (2023)), we have

$$\nabla_{\omega_1} \mathbb{E}_{q_\omega}[\log \pi(X)] = \mathbb{E}_{q_\omega}[\nabla \log \pi(X) - \nabla^2 \log \pi(X)\mu]$$
$$\nabla_{\omega_2} \mathbb{E}_{q_\omega}[\log \pi(X)] = \frac{1}{2}\mathbb{E}_{q_\omega}[\nabla^2 \log \pi(X)].$$

Motivated by these formula, the so-called Bonnet and Price estimator $g^N(\omega) = (g^N(\omega)_1, g^N(\omega)_2)$ is built from N samples $X_n \sim q_\omega$, $n \in [\![1, N]\!]$ by

$$\begin{aligned} g^N(\omega)_1 &= \frac{1}{N}\sum_{n=1}^N \left(\nabla \log \pi(X_n) - \nabla^2 \log \pi(X_n)\mu\right) \\ g^N(\omega)_2 &= \frac{1}{2N}\sum_{n=1}^N \nabla^2 \log \pi(X_n). \end{aligned} \tag{9}$$

This estimator is unbiased. If π is Gaussian, the following proposition states that its associated variance $\sigma_{\eta,N}^2$ is upper-bounded.

Proposition 7. *Suppose that π is a Gaussian distribution with covariance $\Sigma_\pi \in \mathbb{S}_{>0}^d$. Then, g^N is such that Assumptions (A3) and (A4) are satisfied and there exists a constant $V > 0$ that depends only on Σ_π, C, and ω_0 s.t. $\sigma_{\eta,N}^2 \leq \frac{V}{N}$ for all $\eta \in (0, 1]$.*

5.2 Data subsampling

In this section, no assumption on $\mathcal{Q}$ is made at first, but π has the form $\pi(x|y) \propto p_0(x) \prod_{m=1}^M p(y_m|x)$ with $y = \{y_m\}_{m=1}^M$ and M very large. The prior distribution p_0 is in an exponential family $\mathcal{Q}_0$ and satisfies the LERC with respect to $\mathcal{Q}$. Let L_0, Γ_0, A_0 and θ_0 be respectively the linear transformation, sufficient statistic, log-partition, and natural parameter corresponding to p_0, so that $\mathbb{E}_{q_\omega}[\log p_0(X)] =$

$\langle \theta_0, \mathbb{E}_{q_\omega}[\Gamma_0(X)]\rangle - A_0(\theta_0) = \langle L_0^\top \theta_0, \omega\rangle - A_0(\theta_0)$. Equation (5) writes

$$f_\pi^D(\omega) = A^*(\omega) - \langle L_0^\top \theta_0, \omega\rangle + \sum_{m=1}^M \mathbb{E}_{q_\omega}[\log p(y_m|X)] + \text{constant}.$$

and $\nabla f_\pi^D(\omega) = \theta - \left(L_0^\top \theta_0 + \sum_{m=1}^M \theta_{y_m}(\omega)\right)$, where

$$\theta_{y_m}(\omega) = \nabla \mathbb{E}_{q_\omega}[\log p(y_m|X)]$$

is assumed to be tractable. However, for the gradient evaluation, the large sum of M terms is problematic. To circumvent this issue, subsampling only a subset of the data points at every iteration can be considered, thus creating a cheap stochastic estimator of the optimal solution. Consider a uniform random variable $U \sim \mathcal{U}_M$. The gradient can be rewritten as $\nabla f_\pi^D(\omega) = \theta - \left(L_0^\top \theta_0 + M\mathbb{E}_{\mathcal{U}_M}[\theta_{y_U}(\omega)]\right)$, and the expectation approximated by sampling a smaller batch of N data points uniformly, yielding the estimator $g^N(\omega)$ of $\nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$ defined by

$$g^N(\omega) = L_0^\top \theta_0 + \frac{M}{N}\sum_{n=1}^N \theta_{y_{U_n}}(\omega), \tag{10}$$

where $U_n \sim \mathcal{U}_M$ for $n \in [\![1, N]\!]$. This estimator can then be plugged in Algorithm 1 to get an approximation of the optimal θ^*.

Bayesian linear regression As an illustration of the above situation, we investigate a specific Bayesian regression task with the subsampling estimator (10). We consider the same setting as Wu and Gardner (2024), that is Bayesian linear regression with a centered Gaussian prior, Gaussian likelihood, and Gaussian approximating family $\mathcal{Q}$. More specifically,

$X \sim p_0 = \mathcal{N}(0, \Sigma_0)$ with $\theta_0 = -\frac{1}{2}\Sigma_0^{-1}$ and $Y \in \mathbb{R}$ with

$$Y|X = x \sim \mathcal{N}(x^\top z, \sigma^2), \tag{11}$$

with z a vector of fixed covariates values associated to Y. Given M data points $\{y_m\}_{m=1}^M$ and associated covariates $\{z_m\}_{m=1}^M$, the posterior $\pi(x|y_1 \dots y_M)$ is a Gaussian distribution with natural parameter

$$\theta_\pi = \left(\frac{1}{\sigma^2}\sum_{m=1}^M y_m z_m, \theta_0 - \frac{1}{2\sigma^2}\sum_{m=1}^M z_m z_m^\top\right). \tag{12}$$

Assume $\mathcal{Q} = \{q_\omega,\ \omega \in \operatorname{dom} A^*\}$ is the set of Gaussian distributions with expected parameters $\omega = (\omega_1, \omega_2) = (\mu, \Sigma + \mu\mu^\top)$. It follows that p_0 satisfies the LERC since it is Gaussian and thus in $\mathcal{Q}$. For each y_m, $\theta_{y_m}(\omega) = (\nabla_{\omega_1}\mathbb{E}_{q_\omega}[\log p(y_m|X)], \nabla_{\omega_2}\mathbb{E}_{q_\omega}[\log p(y_m|X)])$ is available in closed form:

$$\theta_{y_m}(\omega) = \frac{1}{2\sigma^2}\left(2y_m z_m, -z_m z_m^\top\right).$$

It follows that g^N in (10) writes

$$g^N(\omega) = \frac{M}{N2\sigma^2}\sum_{n=1}^N \left(2y_{U_n} z_{U_n}, -\frac{1}{M}\sigma^2\Sigma_0^{-1} - z_{U_n} z_{U_n}^\top\right).$$

We are in the required setting to define g^N as in (10) and this estimator takes values in $\operatorname{int}\operatorname{dom} A$.

Proposition 8. *When $\pi = q_{\theta_\pi}$ as in Eq. (12) and when g^N is defined as in (10), Assumptions (A3) and (A4) are verified and there exists a constant $V > 0$ which depends only on Σ_0 and on the data such that $\sigma^2_{\eta,N} \le \frac{V}{N}$ for any $\eta \in (0, 1]$.*

5.3 Convergence of Algorithm 1

Both previous Propositions allow to conclude about the convergence of Algorithm 1.

Corollary 1. *Suppose that Algorithm 1 is run in either Propositions 7 or 8 settings, then*

(i) If $\eta_t \equiv \eta \in (0, 1]$ and $N_t \equiv N \in \mathbb{N}_{>0}$, then Algorithm 1 yields iterates $\{\omega_t\}_{t\in\mathbb{N}}$ s.t. for any $T \in \mathbb{N}$,

$$\mathbb{E}[d_{A^*}(\omega^*, \omega_T)] \le (1-\eta)^T d_{A^*}(\omega^*, \omega_0) + \frac{\eta V}{N}.$$

(ii) If $\eta_t \equiv \eta \in (0, 1]$ and $N_t = (t+1)^\gamma, \gamma > 0$, then Algorithm 1 yields iterates $\{\omega_t\}_{t\in\mathbb{N}}$ s.t. for any $T \in \mathbb{N}$,

$$\mathbb{E}[d_{A^*}(\omega_*, \omega_T)] = \mathcal{O}\left(\frac{1}{T^\gamma}\right)$$

(iii) If $\eta_t = \frac{1}{t/2+1}$, then Algorithm 1 yields iterates $\{\omega_t\}_{t\in\mathbb{N}}$ s.t. for any $T \in \mathbb{N}$,

$$\mathbb{E}[d_{A^*}(\omega^*, \omega_{T+1})] = \mathcal{O}\left(\frac{1}{T^2}\sum_{t=0}^T \frac{1}{N_t}\right).$$

6 RELATED WORK

The convergence of stochastic NGVI has seldom been studied. Existing work exploits the connection between natural gradient descent and mirror descent when optimising over an exponential family (see Section 2). Khan et al. (2016) gave convergence rates to stationarity assuming strong convexity of A and a bounded variance, which may fail on some parts of the search space (Guilmeau et al., 2025; Domke, 2019).

Wu and Gardner (2024) exploited a variance-like quantity proposed by Hanzely and Richtárik (2021), which is provably larger (possibly infinite) than $\sigma^2_{\eta,N}$ (Hendrikx, 2024). When $\pi \in \mathcal{Q}$, they showed $\mathcal{O}\left(\frac{1}{T}\right)$ convergence to the minimizer using decreasing step sizes, fixed sample/batch sizes, and Polyak-Ruppert averaging (Polyak and Juditsky, 1992). They demonstrated that their considered variance is finite in a conjugate Bayesian linear regression setting. Sun et al. (2025) also used the variance introduced by Hanzely and Richtárik (2021) to derive convergence rates for convex and non-convex VI problems, covering for instance logistic regression, but which are restricted to mean-field Gaussian approximating families. Also, Sun et al. (2025) used a projection step to a compact constraint set that is necessary for convergence. In Proposition 5, we show a similar rate to Wu and Gardner (2024) without averaging. In addition, we provide in Propositions 4-5 additional possibilities to combine fixed or decreasing step sizes and fixed or increasing sample/batch sizes. These allow to further accelerate convergence and may be closer to what practitioners use. Our result apply for any exponential family $\mathcal{Q}$ and constraint set C (possibly, $C = \mathcal{H}$) and not only for Gaussian mean-field approximations and compact C, although we do require convexity properties. In Section 5, we apply these results to various settings, covering in particular the conjugate Bayesian linear regression task studied by Wu and Gardner (2024).

Our Corollary 1 (i) showcases geometric rates with a bias decreasing inversely with the step size. Similar results were obtained by Lambert et al. (2022) and Diao et al. (2023) for VI algorithms leveraging the Bures-Wasserstein geometry over Gaussian distributions and by Domke et al. (2023) for location-scale VI. However, they considered a sample size $N = 1$, while our result highlight the inverse dependence on $N > 0$ and the case of increasing N_t is covered in Corollary 1 (ii).

Although schedules such as in Propositions 4 and 5 were seldom considered in the VI literature, joint conditions on the step sizes and the sample/batch sizes have been considered within the optimization community, for instance by Ghadimi et al. (2016) for projected gradient descent or by Xiao (2021) for Bregman proximal gradient descent. Our results compliment theirs by using a tighter variance definition and exploiting the relative strong convexity of the objective to yield precise rates. Empirical performance is moreover investigated in Section 7.

Our analysis allows for projection steps (explicit projection operators can be found in Appendix E and additional experiments in Appendix F). Projections have been deemed necessary in the context of stochastic (Euclidean) gradient descent algorithm for VI over location-scale families by Domke et al. (2023) and Kim et al. (2023) but also in the context of natural gradient descent by Sun et al. (2025). Note that in our work, projection is possible, but not mandatory. For natural gradient descent, our projection step can also be seen as an alternative to the Riemannian construction proposed by Lin et al. (2020) and as a way to guarantee the assumptions made by Kumar et al. (2025).

Margossian and Saul (2025) identified conditions on π under which location-scale VI (generally a non-convex problem) admits no local minimizer. These conditions do not force π to belong to the family of approximating distributions, but still require some similarity between π and its approximations. Similarly, our LERC from Definition 6 requires π to be close to $\mathcal{Q}$ and ensures that the VI problem admits a unique solution.

7 NUMERICAL EXPERIMENTS

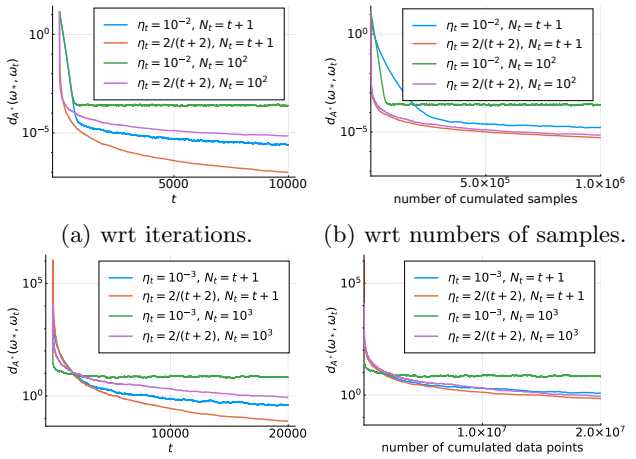
Numerical illustrations are provided in the settings of Section 5: $\mathcal{Q}$ is the family of Gaussian distributions and gradients are estimated either with Bonnet and Price estimators or using subsampling. Projections are not used in this section ($C = \mathcal{H}$). We consider different schedules, namely constant $\eta_t \equiv \eta \in (0, 1]$ as in Proposition 4 and decreasing $\eta_t = \frac{1}{t/2+1}$ as in Proposition 5, as well as constant sample/batch size $N_t \equiv N \in \mathbb{N}_{>0}$ and increasing sample/batch size $N_t = t+1$. Comparing such schedules in terms of iterations can be misleading, as iterations may not involve the same number of data points. Therefore, we also provide plots with respect to the total computational budget used until iteration t, that is, $\sum_{\tau=0}^{t} N_\tau$.

All experiments were run on a personal laptop with 7.6 GB RAM and 8 Intel Core i5-8265U cores. Code can be found at this Github repository.

Bonnet and Price gradients We test Algorithm 1 with the different proposed schedules in the setting of Proposition 7, with π a synthetic Gaussian target in dimension $d = 10$ with condition number $\kappa = 10^2$ generated as in Moré and Toraldo (1989), and g^N is computed using the estimator (9). Figure 1a confirms the result of Propositions 4 and 5: for constant step sizes and sample sizes, the iterates converge geometrically to a neighborhood of ω_*, but if either step sizes decrease or sample sizes increase, the iterates converge to ω_* at a sub-geometric rate. In particular, the convergence is the fastest with decreasing step sizes and increasing sample sizes. Figure 1b shows that in terms of computational budget, it may be beneficial to keep a fixed sample size $N_t \equiv N$, as the schedule with decreasing η_t and fixed N_t achieves the best performance within the given computational budget.

Data subsampling We now test Algorithm 1 in the setting of Proposition 8, where π is a Bayesian linear regression posterior, and g^N is computed by subsampling data points as in (10). We use the CO and NOx (2019) dataset (Kaya et al., 2019) with Gaussian prior $\mathcal{N}(0, 5I)$ and Gaussian likelihood as in (11) with $\sigma^2 = 1$. Here, $M = 36,733$ and $d = 9$.

Figure 1c and 1d showcase results for subsampling-based estimators that are similar to the results obtained with estimators based on Bonnet's and Price's theorems. Namely, quick convergence to a neighborhood of the solution when fixed step and batch sizes are used, and convergence to the solution when either step sizes decrease or batch sizes increase. Again, note that decreasing step sizes and increasing batch sizes may be faster in terms of iterations, but this is not so clear anymore in terms of computational budget. Note also that schedules where $N_t \to +\infty$ may be unrealistic in practice, as batch sizes would reach the total number of data points.

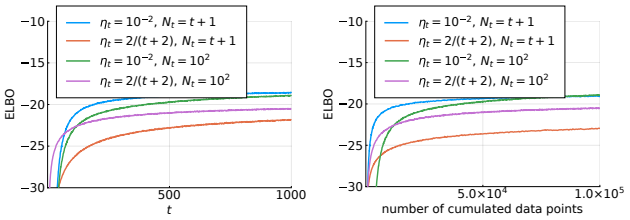


(a) wrt iterations. (b) wrt numbers of samples.

(c) wrt iterations. (d) wrt numbers of samples.

Figure 1: Mean Bregman divergence between current and optimal parameters, over 100 runs, for different NGVI schedules in Prop.7 (a,b) and 8 (c,d) settings.

Logistic regression We then consider a logistic regression task to test Algorithm 1 in a setting when π does not satisfy the LERC with respect to $\mathcal{Q}$. Although π does not satisfy the LERC, $\log \pi$ is still convex and twice-differentiable, implying in particular that (A3) holds (see Appendix C). The logistic regression setting is tested in dimension $d = 5$. $M = 100$ data points $\{z_m\}_{m=1}^M$ are generated uniformly in $[-5, 5]^d$, and for each $m \in [\![1, M]\!]$, y_m follows a Bernoulli distribution with success probability equal to $1/(1 + \exp(-\langle x_*, z_m \rangle))$ where $x_* \in \mathbb{R}^d$ with all its components being equal to 5. Algorithm 1 is run with $\mathcal{Q}$ set to the family of Gaussian distributions with diagonal covariances. The $\mathrm{KL}(q_\omega, \pi)$ evolution is monitored equivalently by the ELBO. A higher ELBO translates to a lower KL (Zhang et al., 2019). The exact ELBO is approximated by a Monte Carlo sum using 100 samples. Figure 2 depicts this Monte-Carlo ELBO averaged over 50 runs, for different sample and step size schedules. Figure 2 shows a slightly different picture from the results obtained in Figure 1. The best ELBO in the transient regime is not reached by the schedule with increasing sample and decreasing step size anymore. Note that our theoretical results only describe the convergence of the Bregman divergence d_{A^*} between the current iterate and the minimizer.



(a) wrt iterations. (b) wrt numbers of samples.

Figure 2: Logistic regression: Average ELBO over 50 runs, for different NGVI schedules.

Additional experiments Additional experiments are deferred to Appendix F. In particular, the impact of potential projection steps is studied on a logistic regression task and on a robust linear regression task where log-convexity of the Bayesian posterior fails.

8 CONCLUSION

In this work, we have proposed new non-asymptotic convergence rates for projected stochastic mirror descent algorithms, which may be of independent interest, and showed that they can be applied to NGVI, including in the case of posterior distributions that cannot be exactly recovered. In particular, our results allow combinations of fixed or decreasing step sizes and fixed or increasing sample/batch sizes. These results extend existing results for NGVI to a wider range of settings and explicit the influence of the choice of step sizes and sample/batch sizes schedules.

Our analysis may apply to slightly extended settings. For instance, generalisations of the exponential families, such as the q-exponential (Amari and Ohara, 2011) and λ-exponential (Wong and Zhang, 2022; Guilmeau et al., 2024) families which include heavy-tailed distributions, or such as mixtures (Lin et al., 2019; Nguyen et al., 2025), represent interesting perspectives to extend our analysis.

Acknowledgements

We are grateful to the reviewers for their constructive comments that helped improve the quality of the paper.

References

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

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

Amari, S.-I. and Ohara, A. (2011). Geometry of the q-exponential family of probability distributions. *Entropy*, 13(6):1170–1185.

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

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

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

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

Bauschke, H. and Combettes, P. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

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

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

Bonnet, G. (1964). Transformations des signaux aléatoires à travers les systèmes non linéaires sans mémoire. *Annales des Télécommunications*, 19(9):203–220.

Bottou, L. (1999). *On-line learning and stochastic approximations*, page 9–42. Cambridge University Press.

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

CO and NOx (2019). Gas Turbine CO and NOx Emission Data Set. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WC95.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. (2023). Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *Internation Conference on Machine Learning (ICML)*.

Domke, J. (2019). Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processings Systems (NeurIPS)*.

Domke, J., Garrigos, G., and Gower, R. (2023). Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dragomir, R., Even, M., and Hendrikx, H. (2021). Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In *International Conference on Machine Learning (ICML)*.

Ghadimi, S., Luan, G., and Zhang, H. (2016). Mini-batch stochastic approximation methods for non-convex stochastic composite optimization. *Mathematical Programming*, 155:267–305.

Godichon-Baggioni, A., Nguyen, D., and Tran, M.-N. (2025). Natural gradient variational Bayes without Fisher matrix analytic calculation and its inversion. *Journal of the American Statistical Association*, 120(550):990–1001.

Guilmeau, T., Branchini, N., Chouzenoux, E., and Elvira, V. (2024). Adaptive importance sampling for heavy-tailed distributions with the λ-exponential family. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Guilmeau, T., Chouzenoux, E., and Elvira, V. (2025). Regularized Rényi divergence minimization through Bregman proximal gradient algorithms. *Journal of Machine Learning Research*.

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

Hendrikx, H. (2024). Investigating variance definitions for stochastic mirror descent with relative smoothness. https://arxiv.org/abs/2404.12213.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Ji, G., Sujono, D., and Sudderth, E. B. (2021). Marginalized stochastic natural gradients for black-box variational inference. In *International Conference on Machine Learning*.

Kaya, H., Tufekci, P., and Uzun, E. (2019). Predicting CO and NOx emissions from gas turbines: Novel data and a benchmark PEMS. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(6):4783–4796.

Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M., and Sugiyama, M. (2016). Faster stochastic variational inference using proximal-gradient methods with general divergence function. In *Conference on Uncertainty in Artifical Intelligence (UAI)*.

Khan, M. E. and Nielsen, D. (2018). Fast yet simple natural-gradient descent for variational inference in complex models. In *International Symposium on Information Theory and Its Applications (ISITA)*.

Khan, M. E. and Rue, H. (2023). The Bayesian learning rule. *Journal of Machine Learning Research*.

Kim, K., Oh, J., Wu, K., Ma, Y.-A., and Gardner, J. R. (2023). On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kingma, D. P. and Welling, M. (2014). Autoencoding variational Bayes. In *International Conference on Learning Representations (ICML)*.

Kumar, N., Möllenhoff, T., Khan, M. E., and Lucchi, A. (2025). Optimization guarantees for square-root natural-gradient variational inference. *Transactions on Machine Learning Research*.

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

Lin, W., Khan, M. E., and Schmidt, M. (2019). Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning (ICML)*.

Lin, W., Schmidt, M., and Khan, M. E. (2020). Handling the positive-definite constraint in the Bayesian learning rule. In *International Conference on Machine Learning (ICML)*.

Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods with applications. *SIAM Journal on Optimization*, 28(1):333–354.

Margossian, C. and Saul, L. K. (2025). Variational inference in location-scale families: Exact recovery of the mean and correlation matrix. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*.

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

Nguyen, D. H., Sakurai, T., and Mamitsuka, H. (2025). Wasserstein gradient flow over variational parameter space for variational inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Nielsen, F. and Nock, R. (2010). Entropies and cross-entropies of exponential families. In *IEEE International Conference on Image Processing (ICIP)*.

Ollivier, Y., Arnold, L., Auger, A., and Hansen, N. (2017). Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Price, R. (1958). A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72.

Ranganath, R., Gerrich, S., and Blei, D. M. (2014). Black box variational inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Raskutti, G. and Mukherjee, S. (2015). The information geometry of mirror descent. *IEEE transactions on Information Theory*, 61(3):1451–1457.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Internation Conference on Machine Learning (ICML)*.

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

Sun, F., Fatkhullin, I., and He, N. (2025). Natural gradient VI: Guarantees for non-conjugate models. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, Y. and Blei, D. M. (2018). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wong, T.-K. L. and Zhang, J. (2022). Tsallis and Rényi deformations linked via new λ-duality. *IEEE Transactions on Information Theory*, 68(8):5353–5373.

Wu, K. and Gardner, J. (2024). Understanding stochastic natural gradient variational inference. In *International Conference on Machine Learning (ICML)*.

Xiao, X. (2021). A unified convergence analysis of stochastic Bregman proximal gradient and extragradient methods. *Journal of Optimization Theory and Applications*, 188:605–627.

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

A EXPONENTIAL FAMILIES

A.1 Steep exponential families

We recall the notion of steepness from Barndorff-Nielsen (2014), which we will use in our proofs.

Definition 7. *Suppose that $\mathcal{Q}$ is an exponential family with log-partition function A that is differentiable on* $\operatorname{int}\operatorname{dom} A$. *We say that $\mathcal{Q}$ is steep if for any sequence $\{\theta_t\}_{t\in\mathbb{N}}$ such that $\theta_t \in \operatorname{int}\operatorname{dom} A$ for any $t \in \mathbb{N}$ and $\theta_t \to \theta$ which belongs to the boundary of* $\operatorname{dom} A$, *then $\|\nabla A(\theta_t)\| \to +\infty$.*

It has been established in (Barndorff-Nielsen, 2014, Theorem 8.2) that if $\operatorname{dom} A$ is open, then $\mathcal{Q}$ is steep.

In this section, we choose the base measure ν to be the Lebesgue measure on $\mathbb{R}^d$ multiplied by $(2\pi)^{-\frac{d}{2}}$.

A.2 Gaussian distributions with full covariance

Proposition 9. *The family of Gaussian distributions on $\mathbb{R}^d$ forms an exponential family with sufficient statistic $\Gamma : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{S}^d$ with $\Gamma(x) = (x, xx^\top)$. For each $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{S}^d_{>0}$, the density of $\mathcal{N}(\mu, \Sigma)$ can be written under the form q_θ outlined in Eq. (2) with*

$$\theta = (\theta_1, \theta_2) = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right),$$
$$A(\theta) = -\frac{1}{4}\theta_1^\top \theta_2^{-1}\theta_1 - \frac{1}{2}\log\det(-2\theta_2).$$

We have $\operatorname{dom} A = \mathbb{R}^d \times \mathbb{S}^d_{<0}$, where $\mathbb{S}^d_{<0}$ denotes the semi-definite negative matrices. Further, the resulting exponential family is steep.

Proof. Denote by $q_{\mu,\Sigma}$ the density of $\mathcal{N}(\mu, \Sigma)$ with respect to ν. Then, at any $x \in \mathbb{R}^d$, we have

$$\begin{aligned} q_{\mu,\Sigma}(x) &= \exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) - \frac{1}{2}\log\det(\Sigma)\right) \\ &= \exp\left(\langle -\frac{1}{2}\Sigma^{-1}, xx^\top\rangle + \langle\Sigma^{-1}\mu, x\rangle - \frac{1}{2}\mu^\top\Sigma^{-1}\mu - \frac{1}{2}\log\det(\Sigma)\right). \end{aligned}$$

We can readily derive the expression of θ, A, and $\operatorname{dom} A$ from here. The steepness property comes from recognising that $\operatorname{dom} A$ is open and applying (Barndorff-Nielsen, 2014, Theorem 8.2). □

Proposition 9 indicates in particular that Gaussian distributions (as an exponential family) enjoy the properties outlined in Proposition 1. In particular, the parameters $\theta = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$ are in bijection with the parameters $\omega = (\mu, \Sigma + \mu\mu^\top)$.

A.3 Gaussian distributions with diagonal covariance matrices

We denote the Hadamard product by $\bullet$.

Proposition 10. *The family of Gaussian distributions on $\mathbb{R}^d$ with diagonal covariance matrices forms an exponential family with sufficient statistic $\Gamma : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ with $\Gamma(x) = (x, x \bullet x)$. For each $\mu \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}^d_{>0}$, the density of $\mathcal{N}(\mu, \operatorname{diag}\sigma^2)$ can be written under the form q_θ outlined in Eq. (2) with*

$$\theta = (\theta_1, \theta_2) = \left(\left(\frac{1}{\sigma^2}\right) \bullet \mu, -\frac{1}{2}\frac{1}{\sigma^2}\right),$$

with the inversion operations being taken point-wise, that is, $\left(\frac{1}{\sigma^2}\right)_i = \frac{1}{(\sigma^2)_i}$ for $i \in [\![1, d]\!]$ with $(\cdot)_i$ denoting the i^{th} element of a vector. Moreover, we have

$$A(\theta) = -\frac{1}{4}\sum_{i=1}^d \frac{(\theta_1^2)_i}{(\theta_2)_i} - \frac{1}{2}\sum_{i=1}^d \log(-2(\theta_2)_i),$$

with $\operatorname{dom} A = \mathbb{R}^d \times \mathbb{R}^d_{<0}$ and the resulting exponential family is steep.

Proof. The proof can be obtained using the same steps as in the proof of Proposition 9. □

Proposition 11. *The family of centered Gaussian distributions with diagonal covariance matrices on $\mathbb{R}^d$ forms an exponential family with sufficient statistic $\Gamma : \mathbb{R}^d \to \mathbb{R}^d$ with $\Gamma(x) = x \bullet x$. For any $\sigma^2 \in \mathbb{R}^d_{>0}$, the density of $\mathcal{N}(0, \operatorname{diag} \sigma^2)$ can be written as in Eq. (2) with $\theta = -\frac{1}{2}\frac{1}{\sigma^2}$ and $A(\theta) = -\frac{1}{2}\sum_{i=1}^d \log(-2\theta_i)$. We have that $\operatorname{dom} A = \mathbb{R}^d_{<0}$ and the resulting exponential family is steep.*

Proof. The proof of this proposition can also be obtained using the same steps as in the proof of Proposition 9. □

We now check that the family of Gaussian distributions over $\mathbb{R}^d$ with full covariance matrices linearly enlarges the family of centered Gaussian distributions with diagonal covariance matrices over $\mathbb{R}^d$. Indeed, consider q_θ in the latter family. Then, $\mathbb{E}_{q_\theta}[X \bullet X] = \sigma^2$ while $\mathbb{E}_{q_\theta}[X] = 0$ and $\mathbb{E}_{q_\theta}[XX^\top] = \operatorname{diag} \sigma^2$. Hence, we have that $\mathbb{E}_{q_\theta}\left[(X, XX^\top)\right] = (0, \operatorname{diag} \mathbb{E}_{q_\theta}[X \bullet X])$, which shows the LERC.

B PROOFS OF SECTION 2

Most proofs of Section 2 are made easier by using the notion of Legendre functions, which we define now.

Definition 8. *Consider a proper convex function $h : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$.*

Then, h is called essentially smooth if $\operatorname{int dom} h \neq \emptyset$, h is differentiable on $\operatorname{int dom} h$, and if for every sequence $\{\omega_t\}_{t\in\mathbb{N}}$ such that $\omega_t \in \operatorname{int dom} h$ for any $t \in \mathbb{N}$ and $\omega_t \to \omega$ in the boundary of $\operatorname{dom} h$, we have the limit $\|\nabla h(\omega_t)\| \to +\infty$.

If h proper, lower-semicontinuous, strictly convex, and essentially smooth, then h will be called Legendre.

Legendre functions benefit from (Rockafellar, 1970, Theorem 26.5), which states the following.

Proposition 12. *Consider a Legendre function h. Then, we have the following:*

(i) h is Legendre if and only if its convex conjugate h^* is Legendre.

(ii) ∇h is a bijection from $\operatorname{int dom} h$ to $\operatorname{int dom} h^*$ and its inverse is ∇h^*.

In this work, we use the notion of steepness from Definition 7. Remark that the steepness of the exponential family $\mathcal{Q}$ is equivalent to the essential smoothness (defined in Definition 8) of its log-partition function A.

Now that we have laid out some preliminary notions, we turn to the proofs of Section 2.

Lemma 1. *Suppose that $\operatorname{int dom} A \neq \emptyset$. Then, A is proper, lower semi-continuous, strictly convex, and all the partial derivatives of A exist on $\operatorname{int dom} A$. Furthermore, if $\mathcal{Q}$ is steep, then A is Legendre.*

Proof. The properness of A comes from the definition of A. The lower semi-continuity and convexity come from (Brown, 1986, Theorem 1.13). The derivability result comes from (Barndorff-Nielsen, 2014, Theorem 8.1). Finally, if $\mathcal{Q}$ is steep, then A is essentially smooth. Combined with the previous properties, we see that stepness implies that A is Legendre. □

B.1 Proof of Proposition 1

Proof. For this proof, we will leverage the results from Lemma 1.

(i) We use the first part of Lemma 1. Next, denoting by $\mathbb{V}_{q_\theta}[\Gamma(X)]$ the variance of $\Gamma(X)$, we have from (Barndorff-Nielsen, 2014, Chapter 8, Eq. (12)-(13)) that

$$\nabla A(\theta) = \mathbb{E}_{q_\theta}[\Gamma(X)] \qquad \text{and} \qquad \nabla^2 A(\theta) = \mathbb{V}_{q_\theta}[\Gamma(X)].$$

Then, the result comes by checking that $\nabla_\theta \log q_\theta(x) = \Gamma(x) - \mathbb{E}_{q_\theta}[\Gamma(X)]$, thus showing that $\nabla^2 A(\theta) = \mathbb{E}_{q_\theta}\left[(\nabla_\theta \log q_\theta(X))(\nabla_\theta \log q_\theta(X))^\top\right]$ which is the Fisher information matrix.

(ii) A is Legendre from Lemma 1, so A^* is also Legendre from Proposition 12. This shows that A^* is proper, lower semi-continuous, strictly convex, and differentiable on $\operatorname{int dom} A^*$. For any $\omega \in \operatorname{int dom} A^*$, we can rewrite

$A^*(\omega) = \langle \theta, \nabla A(\theta) \rangle - A(\theta)$ where $\omega \in \nabla A^*(\theta)$ using the Fenchel-Young equality (Bauschke and Combettes, 2011, Prop. 16.9). It follows that $A^*(\omega) = \langle \theta, \mathbb{E}_{q_\theta}[\Gamma(X)] \rangle - A(\theta) = \mathbb{E}_{q_\theta}[\langle \theta, \Gamma(X) \rangle - A(\theta)] = \mathbb{E}_{q_\theta}[\log q_\theta(X)]$ as desired.

(iii) This result comes from A being Legendre (Lemma 1) and the result of Proposition 12 *(ii)*.

(iv) We first have that $d_A(\theta, \theta') = d_{A^*}(\omega', \omega)$ from (Bauschke and Borwein, 1997, Theorem 3.7 (v)). Then, the equality with the KL divergence comes from Nielsen and Nock (2010). □

B.2 Proof of Proposition 2

Proof. We first show that the domain of proj_C^h is $\operatorname{int}\operatorname{dom} h$, meaning that $\mathrm{proj}_C^h(\omega) \neq \emptyset$ if and only if $\omega \in \operatorname{int}\operatorname{dom} h$. This is done by remarking that h is Legendre so it is essentially strictly convex and that C is closed convex with $C \cap \operatorname{dom} h \neq \emptyset$. Thus, we apply (Bauschke et al., 2003, Proposition 3.31 (vi)) and obtain the result.

We now show that $\mathrm{proj}_C^h(\omega) \subset \operatorname{int}\operatorname{dom} h$ for any $\omega \in \operatorname{int}\operatorname{dom} h$. This is done by remarking that $C \cap \operatorname{int}\operatorname{dom} h \neq \emptyset$ and that h is Legendre so essentially smooth, and using (Bauschke et al., 2003, Proposition 3.33 (v)(b)).

We now demonstrate that proj_C^h is single-valued on $\operatorname{int}\operatorname{dom} h$. To do so, remark that we can apply (Bauschke et al., 2003, Proposition 3.32 (ii)(d)) due to our previous point and the fact that h restricted to $\operatorname{int}\operatorname{dom} h$ is strictly convex. □

Note that when $\mathrm{proj}_C^h(\omega)$ is single-valued, we will consider that $\mathrm{proj}_C^h(\omega)$ is a point of $\mathcal{H}$ instead of being a singleton of $\mathcal{H}$. In the paper, the above result will typically be applied with $h = A^*$.

B.3 Proof of Proposition 3

Proof. We can proceed as in the proof of (Hendrikx, 2024, Prop. 2.4) to obtain that for any $\omega \in D$, we have

$$f(\omega_*) - f_\eta(\omega) \leq \frac{1}{\eta} \mathbb{E}\left[d_h(\bar{\omega}_+, \omega_+) \right].$$

We then obtain our result by taking the supremum over D. □

C PROOFS OF SECTION 3

C.1 Necessary conditions ensuring Assumption (A3)

Assumption (A3) requires the iterates generated by Algorithm 1 to stay in $\operatorname{int}\operatorname{dom} A^*$. Proposition 13 below provides necessary conditions that ensure that any iterate in $\operatorname{int}\operatorname{dom} A^*$ will generate a new iterate in $\operatorname{int}\operatorname{dom} A^*$ as well. Then, this result can be applied recursively to satisfy Assumption (A3).

Proposition 13. *Suppose that Assumptions (A1) and (A2) hold and consider an iterate $\omega_t \in \operatorname{int}\operatorname{dom} A^*$, a step size $\eta_t > 0$, as well as the resulting updated point ω_{t+1} defined by the recursion of Algorithm 1. Then, we have the following:*

(i) If for any $\omega \in \operatorname{int}\operatorname{dom} A^*$, $N \in \mathbb{N}_{>0}$, $g^N(\omega) \in \operatorname{int}\operatorname{dom} A^*$ almost surely, then $\omega_{t+1} \in \operatorname{int}\operatorname{dom} A^*$ almost surely for any step size $\eta_t \in [0, 1]$.

(ii) There exists $\bar{\eta}$ such that if $\eta_t \in (0, \bar{\eta})$, $\omega_{t+1} \in \operatorname{int}\operatorname{dom} A^*$ almost surely.

Note that using similar arguments, point *(i)* holds with $g^N(\omega) \in \operatorname{dom} A^*$ almost surely as long as $\eta_t < 1$.

Proof. In order to prove *(i)* and *(ii)*, we will each time prove that $\omega_+ \in \operatorname{int}\operatorname{dom} A^*$ almost surely. Then, the result on ω_{t+1} will come from applying Proposition 2.

(i) Recall that $(\theta_t)_+ = \nabla A^*((\omega_t)_+)$ and $(\theta_t)_+ = (1 - \eta_t)\theta_t + \eta_t g^N(\omega_t)$, with $\eta_t \in (0, 1]$. Since $\theta_t = \nabla A^*(\omega_t)$ with A^* Legendre and $\omega_t \in \operatorname{int}\operatorname{dom} A^*$, $\theta_t \in \operatorname{int}\operatorname{dom} A$. As this is also assumed to be the case of $g^N(\omega_t)$, we obtain $(\theta_t)_+ \in \operatorname{int}\operatorname{dom} A$ from the convexity of this latter set, and by the Legendre property of A and A^*, $(\omega_t)_+ = \nabla A((\theta_t)_+)$ is in $\operatorname{int}\operatorname{dom} A^*$.

(ii) Similarly as in the proof of (i), we have that $(\theta_t)_+ = (1-\eta_t)\theta_t + \eta_t g^N(\omega_t)$, with $\theta_t \in \operatorname{int}\operatorname{dom} A$. Since this latter set is open, there exists an open ball of radius $r > 0$, centered on θ_t, that is included in $\operatorname{int}\operatorname{dom} A^*$. We can also compute

$$\|(\theta_t)_+ - \theta_t\| \leq \eta_t \|g^N(\omega_t) - \theta_t\|.$$

Since $r > 0$, there exists $\bar{\eta} > 0$ such that $\eta_t \leq \bar{\eta}$ implies $\eta_t \|g^N(\omega_t) - \theta_t\| < r$, thus showing that $(\theta_t)_+ \in \operatorname{int}\operatorname{dom} A^*$ and establishing our result.

□

D PROOFS OF SECTION 4

D.1 Proof of Proposition 4

Proof. Consider a current iterate $\omega_t \in \operatorname{dom} A^*$. We are now going to prove a result analogous to (Dragomir et al., 2021, Lemma 4). In order to do so, we compute

$$\begin{aligned}
&d_{A^*}(\omega_*, \omega_t) - \eta_t d_{f_\pi^D}(\omega_*, \omega_t) - \eta_t (f_\pi^D(\omega_t) - f_\pi^D(\omega_*)) + d_{A^*}(\omega_t, (\omega_t)_+)) \\
&\quad = d_{A^*}(\omega_*, \omega_t) + \eta_t \langle \nabla f_\pi^D(\omega_t), \omega_* - \omega_t \rangle + d_{A^*}(\omega_t, (\omega_t)_+)) \\
&\quad = A^*(\omega_*) - A^*((\omega_t)_+)) - \langle \nabla A^*(\omega_t), \omega_* - \omega_t \rangle - \langle \nabla A^*((\omega_t)_+)), \omega_t - (\omega_t)_+ \rangle \rangle + \eta_t \langle \nabla f_\pi^D(\omega_t), \omega_* - \omega_t \rangle \\
&\quad = A^*(\omega_*) - A^*((\omega_t)_+)) - \langle \nabla A^*(\omega_t), \omega_* - (\omega_t)_+ \rangle \rangle + \eta_t \langle g^N(\omega_t), \omega_t - (\omega_t)_+ \rangle \rangle + \eta_t \langle \nabla f_\pi^D(\omega_t), \omega_* - \omega_t \rangle.
\end{aligned}$$

Now, taking the expectation conditionally on ω_t in the above, and using Assumption (A4), we obtain

$$\begin{aligned}
&d_{A^*}(\omega_*, \omega_t) - \eta_t d_{f_\pi^D}(\omega_*, \omega_t) - \eta_t (f_\pi^D(\omega_t) - f_\pi^D(\omega_*)) + \mathbb{E}[d_{A^*}(\omega_t, (\omega_t)_+))] \\
&\quad = A^*(\omega_*) - \mathbb{E}[A^*((\omega_t)_+))] - \mathbb{E}[\langle \nabla A^*(\omega_t), \omega_* - (\omega_t)_+ \rangle \rangle] + \eta_t \mathbb{E}[\langle \nabla f_\pi^D(\omega_t), \omega_* - (\omega_t)_+ \rangle] \\
&\quad = A^*(\omega_*) - \mathbb{E}[A^*((\omega_t)_+))] - \mathbb{E}[\langle \nabla A^*((\omega_t)_+)), \omega_* - (\omega_t)_+ \rangle] \\
&\quad = \mathbb{E}[d_{A^*}(\omega_*, (\omega_t)_+))].
\end{aligned}$$

The above yields, after rearranging and applying the m-strong convexity property of f_π^D relatively to A^*, the following:

$$\mathbb{E}[d_{A^*}(\omega_*, (\omega_t)_+)] - (1 - m\eta_t) d_{A^*}(\omega_*, \omega_t) \leq \eta_t \left(f_\pi^D(\omega_*) - f_{\pi,\eta_t,N_t}^D(\omega_t) \right),$$

where we have used the definition of f_{π,η_t,N_t}^D in Definition 4. Since $f_\pi^D(\omega_*) - f_{\pi,\eta,N_t}^D(\omega_t) \leq \eta_t \sigma_{\eta_t,N_t}^2$, we finally obtain that $\mathbb{E}[d_{A^*}(\omega_*, (\omega_t)_+)] - (1 - m\eta_t) d_{A^*}(\omega_*, \omega_t) \leq \eta_t^2 \sigma_{\eta_t,N_t}^2$.

We now study the effect of the projection step. We have from Proposition 2 and (Bauschke et al., 2003, Proposition 3.32 (b)) that ω_* is a fixed point of $\operatorname{proj}_C^{A^*}$. Thus, we have from (Bauschke et al., 2003, Prop. 3.3) and (Bauschke et al., 2003, Cor. 3.35) that for any $\omega \in \operatorname{dom} A^*$, $d_{A^*}(\omega_*, \operatorname{proj}_C^{A^*}(\omega)) \leq d_{A^*}(\omega_*, \omega)$. Therefore, $d_{A^*}(\omega_*, \omega_{t+1}) \leq d_{A^*}(\omega_*, (\omega_t)_+)$ from which we deduce that

$$\mathbb{E}\left[d_{A^*}(\omega_*, \omega_{t+1})\right] \leq (1 - m\eta_t) d_{A^*}(\omega_*, \omega_t) + \eta_t^2 \sigma_{\eta_t,N_t}^2. \tag{13}$$

We can obtain the first result with constant sample/batch size $N_t \equiv N$ by iterating inequality (13).

We now turn to the result with varying sample/batch size and under the assumption $\sigma_{\eta,N}^2 \leq \frac{V}{N}$ for some $V > 0$ uniformly in $\eta \in (0, m^{-1}]$. Unrolling the inequality (13) with this additional assumption from $t = 0$ to T yields

$$\mathbb{E}\left[d_{A^*}(\omega_T, \omega_*)\right] \leq (1 - m\eta)^T \mathbb{E}\left[d_{A^*}(\omega_0, \omega_*)\right] + \eta^2 V \sum_{t=0}^{T} \frac{(1 - m\eta)^{T-t}}{N_t}. \tag{14}$$

We can now perform a change of variable $t' = T - t$, so that the sum is equal to:

$$\sum_{t=0}^{T} \frac{(1 - m\eta)^t}{N_{T-t}} = \sum_{t=0}^{\lceil T/2 \rceil} \frac{(1 - m\eta)^t}{N_{T-t}} + \sum_{t=\lceil T/2 \rceil + 1}^{T} \frac{(1 - m\eta)^t}{N_{T-t}} \tag{15}$$

$$\leq \sum_{t=0}^{\lceil T/2 \rceil} \frac{(1 - m\eta)^t}{N_{T-\lceil T/2 \rceil}} + \sum_{t=\lceil T/2 \rceil + 1}^{T} (1 - m\eta)^t. \tag{16}$$

where we used the fact that N_t is monotonically increasing and $N_0 = 1$. In particular, by plugging $N_t = (t+1)^\gamma$, we obtain:

$$\sum_{t=0}^{T} \frac{(1-m\eta)^t}{N_{T-t}} \leq \frac{1}{(T-\lceil T/2\rceil+1)^\gamma} \sum_{t=0}^{\lceil T/2\rceil} (1-m\eta)^t + \left((1-m\eta)^{\lceil T/2\rceil+1}\right) \sum_{t=0}^{T-\lceil T/2\rceil-1} (1-m\eta)^t \tag{17}$$

$$\leq \frac{1}{(T-\lceil T/2\rceil+1)^\gamma} \frac{1}{m\eta} + \left((1-m\eta)^{\frac{T+1}{2}}\right) \frac{1}{m\eta}. \tag{18}$$

Since $\lceil T/2\rceil - 1 < \frac{T}{2} \leq \lceil T/2\rceil$, we can finally obtain that $T - \lceil T/2\rceil + 1 > \frac{T}{2}$, yielding

$$\sum_{t=0}^{T} \frac{(1-m\eta)^t}{N_{T-t}} \leq \frac{1}{m\eta} \left(\frac{2^\gamma}{T^\gamma} + (1-m\eta)^{\frac{T+1}{2}} \right). \tag{19}$$

We finally obtain the result by using the above in Eq. (14). Remark that $T/2$ can actually be replaced by any αT for $0 < \alpha < 1$. □

D.2 Proof of Proposition 5

Proof. We have from Eq. (13) that at any $t \in \mathbb{N}$ the iterates of Algorithm 1 satisfy

$$\mathbb{E}\left[d_{A^*}(\omega_*, \omega_{t+1})\right] \leq (1 - m\eta_t) d_{A^*}(\omega_*, \omega_t) + \eta_t^2 \sigma^2_{\eta_t, N_t}.$$

Taking expectation in the above and dividing by η_t yields

$$0 \leq \eta_t \sigma^2_{\eta_t, N_t} + \left(\frac{1}{\eta_t} - m\right) \mathbb{E}\left[d_{A^*}(\omega_*, \omega_t)\right] - \frac{1}{\eta_t} \mathbb{E}\left[d_{A^*}(\omega_*, \omega_{t+1})\right].$$

Then, using that $\eta_t = \frac{2}{m(t+2)}$, one obtains

$$0 \leq \frac{2}{m(t+2)} \sigma^2_{\eta_t, N_t} + \frac{m}{2} t \mathbb{E}\left[d_{A^*}(\omega_*, \omega_t)\right] - \frac{m}{2}(t+2) \mathbb{E}\left[d_{A^*}(\omega_*, \omega_{t+1})\right].$$

Then, multiplying by $t+1$ yields

$$0 \leq \frac{2(t+1)}{m(t+2)} \sigma^2_{\eta_t, N_t} + \frac{m}{2} t(t+1) \mathbb{E}\left[d_{A^*}(\omega_*, \omega_t)\right] - \frac{m}{2}(t+1)(t+2) \mathbb{E}\left[d_{A^*}(\omega_*, \omega_{t+1})\right].$$

Finally, using our additional assumption on the variance, it comes $\frac{t+1}{t+2} \sigma^2_{\eta_t, N_t} \leq \sigma^2_{\eta_t, N_t} \leq \frac{V}{N_t}$ and summing for $t \in [\![0, T]\!]$, we obtain that

$$0 \leq \frac{4V}{m^2} \sum_{t=0}^{T} \frac{1}{N_t} - (T+1)(T+2) \mathbb{E}\left[d_{A^*}(\omega_*, \omega_{T+1})\right]$$

from which we obtain the result. □

D.3 Proof of Proposition 6

Proof. We first state a preliminary result. Recall from Equation (5) that for any $\omega \in \operatorname{dom} A^*$, $f_\pi^D(\omega) = A^*(\omega) - \mathbb{E}_{q_\omega}[\log \pi(X)]$. Because $\pi = q_{\theta_\pi} \in \mathcal{Q}_\pi$, $\mathbb{E}_{q_\omega}[\log \pi(X)] = \langle \mathbb{E}_{q_\omega}[\Gamma_\pi(X)], \theta_\pi \rangle - A_\pi(\theta_\pi)$. Using the LERC, we have that $\mathbb{E}_{q_\omega}[\Gamma_\pi(X)] = L_\pi \mathbb{E}_{q_\omega}[\Gamma(X)] = L_\pi \omega$. We thus conclude that

$$f_\pi^D(\omega) = A^*(\omega) - \langle L_\pi^\top \theta_\pi, \omega \rangle + A_\pi(\theta_\pi)\,. \tag{20}$$

(i) We first show that our Problem (8) admits solutions that are necessarily in $\operatorname{int} \operatorname{dom} A^*$. Remark that solving Problem (8) is equivalent to minimizing $\omega \longmapsto f_\pi^D(\omega) + \iota_C(\omega)$ where ι_C is such that

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

Because of (A1) and Proposition 1, and using (20), f_π^D is lower semi-continuous. Because of (A2), ι_C is also lower semi-continuous. Therefore, $f_\pi^D + \iota_C$ is lower semi-continuous. Further, $f_\pi^D + \iota_C$ is coercive if C is compact ($\mathcal{H}$ being finite-dimensional, this implies C is bounded) or if $L_\pi^\top \theta_\pi \in \operatorname{int}\operatorname{dom} A$, from (Bauschke and Borwein, 1997, Fact 2.11). Therefore, $f_\pi^D + \iota_C$ is lower semi-continuous and coercive, so there exist minimizers.

Now, let us show that these solutions necessarily belong to $\operatorname{int}\operatorname{dom} A^*$. Assumption (A2) implies that $C \cap \operatorname{int}\operatorname{dom} f_\pi^D \neq \emptyset$ because of the LERC and (20). So we can apply (Bauschke and Combettes, 2011, Proposition 6.19 (vii)) and (Bauschke and Combettes, 2011, Proposition 26.5 (a)) to have that at any minimizer ω_*, $\partial A^*(\omega_*) \neq \emptyset$, where ∂A^* is the subdifferential of A^* (Bauschke and Combettes, 2011, Chapter 16). This necessarily implies that $\omega_* \in \operatorname{int}\operatorname{dom} A^*$, because due to A^* being Legendre (from Lemma 1 and Proposition 12), $\operatorname{dom} \partial A^* = \operatorname{int}\operatorname{dom} A^*$ from (Rockafellar, 1970, Theorem 26.1). Thus, we have shown that solutions of Problem (8) exist and are in $\operatorname{int}\operatorname{dom} A^*$.

Finally, let us prove the uniqueness. Because of the constraint and previous points, possible solutions are in $C \cap \operatorname{int}\operatorname{dom} A^*$. Because of Assumption (A1) and Proposition 1, A^* is strictly convex on $\operatorname{int}\operatorname{dom} A^*$, so f_π^D is also strictly convex on $\operatorname{int}\operatorname{dom} A^*$ due to the LERC and (20). Additionally, Assumption (A2) implies that ι_C is convex on $\mathcal{H}$. Thus, $f_\pi^D + \iota_C$ is strictly convex on $C \cap \operatorname{int}\operatorname{dom} A^*$, showing that there exist only a unique solution ω_* to Problem (8), and establishing the result.

We now show that if $L_\pi^\top \theta_\pi \in \operatorname{int}\operatorname{dom} A$, then $\omega_* = \operatorname{proj}_C^{A^*}(\nabla A(L_\pi^\top \theta_\pi))$. Because of (20), the fact that A^* is Legendre, and that ω_* minimizes $\omega \longmapsto f_\pi^D(\omega) + \iota_C(\omega)$, ω_* satisfies the optimality conditions

$$0 \in \nabla A^*(\omega_*) - \nabla A^*(\nabla A(L_\pi^\top \theta_\pi)) + \partial \iota_C(\omega_*),$$

where $\partial \iota_C$ is the subdifferential of ι_C (Bauschke et al., 2017, Chapter 16). We can thus recognize that ω_* satisfies the optimality conditions associated to the minimization of $\omega \longmapsto d_{A^*}(\omega, \nabla A(L_\pi^\top \theta_\pi)) + \iota_C(\omega)$, which is the optimization problem associated with the computation of $\operatorname{proj}_C^{A^*}(\nabla A(L_\pi^\top \theta_\pi))$. Because of Proposition 2, this problem admits a unique solution that satisfies the optimality conditions, so we can deduce that $\omega_* = \operatorname{proj}_C^{A^*}(\nabla A(L_\pi^\top \theta_\pi))$, showing our result.

(ii) We can use Equation (20) above so that f_π^D is the sum of A^* and an affine term in ω. Since the Bregman divergence induced by an affine function is zero and the Bregman divergence induced by a sum of functions is the sum of the Bregman divergences induced by each function, we obtain that $d_{f_\pi^D} = d_{A^*}$, thus showing the result. □

D.4 Proof of Proposition 7

Proof. The target distribution π is here assumed to be a Gaussian distribution with mean μ_π and covariance Σ_π. As $\mathcal{Q}$ is the Gaussian family, then π satisfies the LERC with respect to $\mathcal{Q}$, see Definition 6. It follows that f_π^D is 1-smooth relatively to A^* due to Proposition 6 *(ii)*. We thus have from Proposition 3 and Proposition 1 (iv) that

$$\sigma_{\eta,N}^2 \leq \frac{1}{\eta^2} \sup_{\omega \in C \cap \operatorname{int}\operatorname{dom} A^*} \left\{ \mathbb{E}\left[d_A(\theta_+, \bar{\theta}_+) \right] \right\},$$

where $\theta_+ = (1-\eta)\theta + \eta g^N(\omega)$ and $\bar{\theta}_+ = (1-\eta)\theta + \eta \nabla \mathbb{E}_{q_\omega}[\log \pi(X)]$, so that $\mathbb{E}[\theta_+] = \bar{\theta}_+$.

Let us first specify these latter expressions. When q_ω is a Gaussian distribution, as already used in Section 5.1, we have more generally with some arbitrary differentiable function l that

$$\begin{aligned}
\nabla_{\omega_1} \mathbb{E}_{q_\omega}[l(X)] &= \mathbb{E}_{q_\omega}[\nabla l(X)] - \mathbb{E}_{q_\omega}[\nabla^2 l(X)]\mu \\
\nabla_{\omega_2} \mathbb{E}_{q_\omega}[l(X)] &= \frac{1}{2}\mathbb{E}_{q_\omega}[\nabla^2 l(X)],
\end{aligned}$$

which can be estimated with the following estimators

$$\begin{aligned}
g^N(\omega)_1 &= \frac{1}{N}\sum_{n=1}^N \left(\nabla l(X_n) - \nabla^2 l(X_n)\mu \right) \\
g^N(\omega)_2 &= \frac{1}{N}\sum_{n=1}^N \frac{1}{2}\nabla^2 l(X_n).
\end{aligned}$$

These estimators can be checked to be unbiased, thus satisfying Assumption (A4).

Thus, for $\log \pi(x) = -\frac{1}{2}(x-\mu_\pi)^\top \Sigma_\pi^{-1}(x-\mu_\pi) + \text{constant}$, it comes

$$\nabla_{\omega_1} \mathbb{E}_{q_\omega}[\log \pi(X)] = -\Sigma_\pi^{-1}(\mu - \mu_\pi) + \Sigma_\pi^{-1}\mu = \Sigma_\pi^{-1}\mu_\pi \qquad g^N(\omega)_1 = \Sigma_\pi^{-1}\left(\mu_\pi + \mu - \frac{1}{N}\sum_{n=1}^N X_n\right)$$

$$\nabla_{\omega_2} \mathbb{E}_{q_\omega}[\log \pi(X)] = -\frac{1}{2}\Sigma_\pi^{-1} \qquad g^N(\omega)_2 = -\frac{1}{2}\Sigma_\pi^{-1}.$$

We can readily notice that $g^N(\omega) \in \operatorname{int}\operatorname{dom} A$ almost surely, thus establishing (A3) using Proposition 13.

We can then also remark that $\bar{\theta}_+ = (1-\eta)\theta + \eta\theta_\pi$ and denoting $\bar{\theta}_+ = (\bar{\theta}_{+,1}, \bar{\theta}_{+,2})$, $\theta_+ = (\theta_{+,1}, \theta_{+,2})$ that $\bar{\theta}_{+,2} = \theta_{+,2}$.

Going back to computing $d_A(\theta_+, \bar{\theta}_+)$ in the above bound, it is equivalent to computing the KL divergence between 2 Gaussian distributions respectively defined by $\bar{\theta}_+$ and θ_+ (Proposition 1 (iv)). The expression of this KL is

$$KL(q_{\bar{\theta}_+}, q_{\theta_+}) = A(\theta_+) - A(\bar{\theta}_+) - \langle \theta_+ - \bar{\theta}_+, \bar{\omega}_+ \rangle \ .$$

For a Gaussian distribution with natural parameter $\theta = (\theta_1, \theta_2)$ the expression of the log-partition function is $A(\theta) = -\theta_1^\top (4\theta_2)^{-1}\theta_1 - \frac{1}{2}\log\det(-2\theta_2)$. It follows that using $\bar{\theta}_{+,2} = \theta_{+,2}$,

$$A(\theta_+) - A(\bar{\theta}_+) = \theta_{+,1}^\top \left(-4\theta_{+,2}\right)^{-1} \theta_{+,1} - \bar{\theta}_{+,1}^\top \left(-4\theta_{+,2}\right)^{-1} \bar{\theta}_{+,1} \ ,$$

and

$$\langle \theta_+ - \bar{\theta}_+, \bar{\omega}_+ \rangle = \langle \theta_{+,1} - \bar{\theta}_{+,1}, \bar{\omega}_{+,1} \rangle$$

where $\bar{\omega}_{+,1} = \bar{\mu}_+ = \left(-2\theta_{+,2}\right)^{-1} \bar{\theta}_{+,1}$.

Combining both terms, it comes

$$d_A(\theta_+, \bar{\theta}_+) = \theta_{+,1}^\top \left(-4\theta_{+,2}\right)^{-1} \theta_{+,1} - \bar{\theta}_{+,1}^\top \left(-4\theta_{+,2}\right)^{-1} \bar{\theta}_{+,1} - \langle 2\left(-4\theta_{+,2}\right)^{-1} \bar{\theta}_{+,1}, \theta_{1,+} - \bar{\theta}_{+,1} \rangle.$$

From here, we deduce, using that $\mathbb{E}[\theta_{1,+}] = \bar{\theta}_{+,1}$,

$$\begin{aligned}
\mathbb{E}\left[d_A(\theta_+, \bar{\theta}_+)\right] &= \mathbb{E}\left[\theta_{+,1}^\top \left(-4\theta_{+,2}\right)^{-1} \theta_{+,1}\right] - \mathbb{E}\left[\theta_{+,1}\right]^\top \left(-4\theta_{+,2}\right)^{-1} \mathbb{E}\left[\theta_{+,1}\right] \\
&= \mathbb{E}\left[\left(\theta_{+,1} - \mathbb{E}\left[\theta_{+,1}\right]\right)^\top \left(-4\theta_{+,2}\right)^{-1} \left(\theta_{+,1} - \mathbb{E}\left[\theta_{+,1}\right]\right)\right] \\
&= \eta^2 \mathbb{E}\left[\left(g^N(\omega)_1 - \theta_{\pi,1}\right)^\top \left(-4\theta_{+,2}\right)^{-1} \left(g^N(\omega)_1 - \theta_{\pi,1}\right)\right] \\
&= \eta^2 \mathbb{E}\left[(\mu - \frac{1}{N}\sum_{n=1}^N X_n)^\top \Sigma_\pi^{-1} \left(-4\theta_{+,2}\right)^{-1} \Sigma_\pi^{-1} (\mu - \frac{1}{N}\sum_{n=1}^N X_n)\right] \\
&= \frac{\eta^2}{N^2} \left\langle \Sigma_\pi^{-1} \left(-4\theta_{+,2}\right)^{-1} \Sigma_\pi^{-1}, \mathbb{E}\left[(\sum_{n=1}^N (X_n - \mu))(\sum_{n=1}^N (X_n - \mu))^\top\right] \right\rangle .
\end{aligned}$$

Since $X_n - \mu \sim \mathcal{N}(0, \Sigma)$ for any $n \in [\![1, N]\!]$ the matrix-valued random variable $(\sum_{n=1}^N (X_n - \mu))(\sum_{n=1}^N (X_n - \mu))^\top$ follows a Wishart distribution with N degrees of freedom and scale matrix Σ, so its expectation is $N\Sigma$. Therefore, we have shown that

$$\mathbb{E}\left[d_A(\theta_+, \bar{\theta}_+)\right] = \frac{\eta^2}{N} \left\langle \Sigma_\pi^{-1} \left(-4\theta_{+,2}\right)^{-1} \Sigma_\pi^{-1}, \Sigma \right\rangle .$$

We now have to bound the above quantity on $\operatorname{int}\operatorname{dom} A^*$. The trace formulation of the previous identity is:

$$\mathbb{E}\left[d_A(\theta_+, \bar{\theta}_+)\right] = \frac{\eta^2}{N} \operatorname{tr}\left(\Sigma_\pi^{-1} \left(-4\theta_{+,2}\right)^{-1} \Sigma_\pi^{-1} \Sigma\right) .$$

We now handle the term in $\theta_{+,2}$. By definition $-4\theta_{+,2} = 2(1-\eta)\Sigma^{-1} + 2\eta\Sigma_\pi^{-1}$. We introduce the map

$$\begin{aligned} g(\Sigma, \eta) &= \frac{1}{2} \operatorname{tr}\left(\Sigma_\pi^{-1}\left((1-\eta)\Sigma^{-1} + \eta\Sigma_\pi^{-1}\right)^{-1}\Sigma_\pi^{-1}\Sigma\right) \\ &= \frac{1}{2} \operatorname{tr}\left(\tilde{\Sigma}\left((1-\eta)I + \eta\tilde{\Sigma}\right)^{-1}\tilde{\Sigma}\right), \end{aligned}$$

with $\tilde{\Sigma} = \Sigma^{\frac{1}{2}}\Sigma_\pi^{-1}\Sigma^{\frac{1}{2}}$. Expressing $\tilde{\Sigma} = Q^\top \Delta Q$, where $Q^\top Q = QQ^\top = I$ and Δ is a diagonal matrix, we obtain:

$$\begin{aligned} g(\Sigma, \eta) &= \frac{1}{2} \operatorname{tr}\left(Q^\top\Delta Q\left((1-\eta)Q^\top Q + \eta Q^\top \Delta Q\right)^{-1} Q^\top \Delta Q\right) \\ &= \frac{1}{2} \operatorname{tr}\left(Q^\top Q \Delta Q\left(Q^\top[(1-\eta)I + \eta\Delta]V\right)^{-1} Q^\top \Delta\right) \\ &= \frac{1}{2} \operatorname{tr}\left(\Delta\left[(1-\eta)I + \eta\Delta\right]^{-1}\Delta\right) \\ &= \frac{1}{2}\sum_{i=1}^{d} \frac{\Delta_{ii}^2}{1-\eta+\eta\Delta_{ii}}. \end{aligned}$$

One can directly notice that $g(\Sigma, \eta)$ remains bounded for any $\eta \in (0, 1]$ as any of the Δ_{ii} goes to 0. However, we need the Δ_{ii} to remain bounded to ensure boundedness of $g(\Sigma, \eta)$. While this is not true in general, we actually do not need to take the supremum over $C \cap \operatorname{int}\operatorname{dom} A^*$, but actually only need to take the value $g(\Sigma, \eta)$ at the point that achieves the supremum in Definition 4 (which is achieved). While possible, this derivation is rather cumbersome.

Instead, suppose that $C = \mathcal{H}$, i.e., no constraints are applied (the constrained case will be treated later on). Then, we notice that at each step, the updates in Σ are such that the updated values of Σ will belong to the set $C' = \{\Sigma \in \mathbb{S}^d,\ \Sigma = ((1-\eta)\Sigma_\pi^{-1} + \eta\Sigma_0)^{-1},\ \eta \in [0, 1]\}$. Because $\Sigma_0, \Sigma_\pi \in \mathbb{S}^d_{>0}$ by assumption, $C' \subset \mathbb{S}^d_{>0}$ and is compact.

The above ensures that the eigenvalues of Σ are bounded from above, so that the Δ_{ii} are also bounded from above. This implies that there exists $\widetilde{V} > 0$ such that

$$\mathbb{E}\left[d_A(\theta_+, \bar{\theta}_+)\right] \leq \frac{\eta^2\widetilde{V}}{N},\ \forall \omega \in \operatorname{int}\operatorname{dom} A^*.$$

We can therefore apply Proposition 3 to obtain the result.

In the case where $C \neq \mathcal{H}$, meaning that additional constraints are imposed, one can simply notice that we take a supremum on $C \cap \operatorname{int}\operatorname{dom} A^*$ in the definition of $\sigma^2_{\eta,N}$ instead of a supremum on $\operatorname{int}\operatorname{dom} A^*$, which is a bigger set. Therefore, the bound that we derived for $C = \mathcal{H}$ also holds when C is smaller.

□

D.5 Proof of Proposition 8

Proof. We can readily check that g^N is an unbiased estimator of $\nabla_\omega \mathbb{E}_{q_\omega}[\log \pi(X)]$ and that it belongs almost surely to $\operatorname{int}\operatorname{dom} A^*$, allowing to apply Proposition 13. Thus, Assumptions (A3) and (A4) are satisfied.

We are going to bound $\sigma^2_{\eta,N}$ using Proposition 3 and the fact that f^D_π is 1-smooth relatively to A^* from Proposition 6 (ii). Proposition 3 leads to

$$\begin{aligned} \sigma^2_{\eta,N} &\leq \frac{1}{\eta^2}\sup_{\omega \in \operatorname{dom} A^*}\left\{\mathbb{E}\left[d_{A^*}(\overline{\omega}_+, \omega_+)\right]\right\} \\ &\leq \frac{1}{\eta^2}\sup_{\omega \in \operatorname{dom} A^*}\left\{\mathbb{E}\left[d_{A^*}(\overline{\omega}_+, \omega_+) + d_{A^*}(\omega_+, \overline{\omega}_+)\right]\right\} \end{aligned}$$

The right-hand side above is then equal to

$$\frac{1}{\eta^2}\sup_{\omega \in \operatorname{dom} A^*}\left\{\mathbb{E}\left[\langle \nabla A^*(\overline{\omega}_+) - \nabla A^*(\omega_+), \overline{\omega}_+ - \omega_+\rangle\right]\right\} = \frac{1}{\eta}\sup_{\omega \in \operatorname{dom} A^*}\left\{\mathbb{E}\left[\langle g^N(\omega) - \nabla f^D_\pi(\omega), \overline{\omega}_+ - \omega_+\rangle\right]\right\}$$

so that eventually

$$\sigma^2_{\eta,N} \leq \frac{1}{\eta} \sup_{\omega \in \operatorname{dom} A^*} \left\{ \mathbb{E}\left[\langle g^N(\omega) - \nabla f^D_\pi(\omega), \overline{\omega}_+ - \omega_+ \rangle \right] \right\}$$

This latter right-hand side expression has been proposed by Hanzely and Richtárik (2021) as a variance-like quantity to analyse mirror descent. In the case of NGVI for Bayesian linear regression with Gaussian prior, likelihood, and approximating family, this quantity has been upper-bounded in (Wu and Gardner, 2024, Lemma 4) by the constant

$$V_2 = (s s_1 + \frac{1}{2} s^2 s_2 + 2 s^2 b \sqrt{s_1 s_2} M + s^3 b^2 s_2 M^2) \frac{M^2}{\sigma^4 N},$$

where s, s_1, s_2, b are given by the following expressions,

$$s = \max\{1, \|\Sigma_0\|\}, \qquad b = \max_{1 \leq n \leq N} \|y_m z_m\|,$$

$$s_1 = \mathbb{E}_{\mathcal{U}_M}\left[\left\| y_X z_X - \frac{1}{N} \sum_{m=1}^M y_m z_m \right\|^2 \right], \qquad s_2 = \mathbb{E}_{\mathcal{U}_M}\left[\left\| z_X z_X^\top - \frac{1}{N} \sum_{m=1}^M z_m z_m^\top \right\|^2 \right].$$

□

E PROJECTION OPERATORS

In Section 3, we have introduced the possibility of further restricting the search space from distributions $q_\theta \in \mathcal{Q}$ with $\theta \in \operatorname{int} \operatorname{dom} A^*$ to distributions $q_\theta \in \mathcal{Q}$ with $\theta \in C \cap \operatorname{int} \operatorname{dom} A^*$. This can be used to safeguard iterates of Algorithm 1 from problematic behaviours (e.g., covariance matrices becoming singular) that may be due, for instance, to a number of samples that is too low. Another interest of this constraint is to incorporate information about the target distribution π. This constraint translates to an additional projection step in Algorithm 1. We now give two situations where this projection operator admits a closed form.

E.1 Gaussian distributions with constrained covariance matrices

In the case of Gaussian distributions, an example of a constraints space C with an explicit projection operator is defined for $0 < \alpha < \beta$ as:

$$C = \{(\mu, \Sigma + \mu\mu^\top) \in \mathcal{H} \text{ s.t. } \alpha I \preccurlyeq \Sigma \preccurlyeq \beta I\}. \tag{21}$$

Proposition 14. *Consider $\mathcal{Q}$ the exponential family of Gaussian distributions, the set C defined in (21), and $\omega = (\mu, \Sigma + \mu\mu^\top)$. Then, $\operatorname{proj}_C^{A^*}(\omega) = (\mu, \Sigma_P + \mu\mu^\top)$ with Σ_P being a version of Σ whose eigenvalues larger (resp. smaller) than β (resp. α) have been replaced by β (resp. α).*

Proof. We recall that $\operatorname{proj}_C^{A^*}(\omega) = \arg\min_{\omega'} \{\iota_C(\omega') + d_{A^*}(\omega', \omega)\}$. Since $\omega \in C$ depends only on Σ, we rewrite the optimisation problem associated to the projection operator in terms of the mean and covariance matrices, where $\omega = (\mu, \Sigma + \mu\mu^\top)$. We then get that $(\mu_P, \Sigma_P + \mu_P \mu_P^\top) = \operatorname{proj}_C^{A^*}(\omega)$ are the solutions to the optimisation problem

$$\min_{\mu' \in \mathbb{R}^d, \Sigma' \in \mathbb{S}^d} \left\{ \iota_{\widetilde{C}}(\Sigma') + \frac{1}{2} \left(-\log\det \Sigma' + (\mu - \mu')^\top \Sigma^{-1} (\mu - \mu') + \langle \Sigma^{-1}, \Sigma \rangle_{\mathbb{S}^d} \right) \right\}.$$

We thus observe that $\mu_P = \mu$ and that Σ_P is a solution to

$$\min_{\Sigma' \in \mathbb{S}^d} \left\{ \iota_{\widetilde{C}}(\Sigma') - \frac{1}{2} \log\det \Sigma' + \langle \Sigma^{-1}, \Sigma' \rangle_{\mathbb{S}^d} \right\}.$$

We will now show using (Benfenati et al., 2020, Theorem 1) that this problem can be reduced to a problem on $\mathbb{R}^d$ involving only the eigenvalues of Σ'.

If we define by $\lambda' \in \mathbb{R}^d$ the set of eigenvalues of Σ', we can remark that $\iota_{\widetilde{C}}(\Sigma') = \sum_{i=1}^d \iota_{[\alpha,\beta]}(\lambda'_i)$ and $-\frac{1}{2}\log\det\Sigma' = -\frac{1}{2}\sum_{i=1}^d \log \lambda_i$, meaning that these two functions only depend on the eigenvalues of Σ'. Further, each of these functions is lower semicontinuous, their sum is coercive, and Σ is such that there exists an

orthonormal matrix Q and vector $\delta \in \mathbb{R}^d$ such that $\Sigma^{-1} = Q\,\mathrm{diag}(\delta)Q^\top$, meaning that we can invoke (Benfenati et al., 2020, Theorem 1) to show that $\Sigma_P = Q\,\mathrm{diag}(\lambda_P)Q^\top$ with λ_P being a solution to

$$\min_{\lambda' \in \mathbb{R}^d} \left\{ \sum_{i=1}^d \iota_{[\alpha,\beta]}(\lambda'_i) - \frac{1}{2}\sum_{i=1}^d \log \lambda'_i + \frac{1}{2}\sum_{i=1}^d \lambda_i \delta_i \right\}.$$

This problem has a separable structure, implying that for each $i \in [\![1, d]\!]$,

$$(\lambda_P)_i = \operatorname*{arg\,min}_{\alpha \le \lambda' \le \beta} \{-\log \lambda' + \delta_i \lambda'\}$$

Since for each $i \in [\![1, d]\!]$, $\delta_i = \frac{1}{\lambda_i}$ where $\lambda \in \mathbb{R}^d_{>0}$ is the vector of eigenvalues of Σ, we can finally check the result by solving the above problem using the Karush-Kuhn-Tucker conditions. □

E.2 Diagonal Gaussians with non-negative means

In this section, we show that when $\mathcal{Q}$ is the family of Gaussians with diagonal covariance, introduced in Section A.3, the projection to

$$C = \{(\mu, \sigma^2 + \mu \bullet \mu) \in \mathbb{R}^d \times \mathbb{R}^d \text{ s.t. } \mu \in \mathbb{R}^d_{\ge 0}\} \tag{22}$$

admits an explicit expression.

Proposition 15. *Consider the exponential family of Gaussians with diagonal covariance, the set C defined in (22), and $\omega = (\mu, \sigma^2 + \mu \bullet \mu)$. Then, $\mathrm{proj}_C^{A^*}(\omega) = (\mu_P, \sigma^2 + \mu_P \bullet \mu_P)$ with $(\mu_P)_i = \max(0, \mu_i)$ for $i \in [\![1, d]\!]$.*

Proof. The projection of $\omega \in \operatorname{int}\operatorname{dom} A^*$ on C as defined in (22) consists in solving

$$\omega_P = \operatorname*{arg\,min}_{\omega' \in \operatorname{int}\operatorname{dom} A^*} \left(g(\omega') + d_{A^*}(\omega', \omega)\right),$$

where $g(\omega) = \iota_{\mathbb{R}^d_{\ge 0}}(\omega_1)$. Therefore, ω_P solves the optimality conditions $0 \in \nabla A^*(\omega_P) - \nabla A^*(\omega) + \partial g(\omega)$ where ∂g is the subdifferential of g (see (Bauschke and Combettes, 2011, Chapter 16)). Since g depends only on $(\omega_P)_1 = \mu_P$ and is separable, we see that ω_P solves

$$\begin{aligned} 0 &\in \frac{1}{(\sigma_P^2)_i}(\mu_P)_i - \frac{1}{(\sigma^2)_i}(\mu)_i + \partial \iota_{\mathbb{R}_{\ge 0}}((\mu_P)_i) \\ 0 &= \frac{1}{(\sigma_P^2)_i} - \frac{1}{(\sigma^2)_i}, \end{aligned}$$

for any $i \in [\![1, d]\!]$. We thus readily obtain that $\sigma_P^2 = \sigma^2$. Now, using that fact that $\partial \iota_{\mathbb{R}_{\ge 0}}(s)$ is the normal cone $N_{\mathbb{R}_{\ge 0}}(s)$ for an scalar s (see (Bauschke and Combettes, 2011, Example 16.12)), and that

$$N_{\mathbb{R}_{\ge 0}}(s) = \begin{cases} \mathbb{R}_{\le 0} \text{ if } s = 0, \\ \emptyset \text{ if } s < 0, \\ \{0\} \text{ if } s > 0, \end{cases}$$

we obtain that if $\mu_i < 0$, $(\mu_P)_i = 0$ and that $\mu_i = (\mu_P)_i$ else. □

F ADDITIONAL NUMERICAL ILLUSTRATIONS

F.1 Additional details about the experiments of Section 7

For a numerical illustration of our results in the case of data subsampling estimators, we consider the CO and NOx (2019) dataset (Kaya et al., 2019). We use the NOx emissions as a response variable Y and the other variables (excluding the CO variable, thus $d = 9$) as fixed covariates.

In the case of experiments performed in the conjugate setting, all the runs are initialized with initial mean simulated uniformly in $[-5, 5]^d$ and initial covariance matrix being equal $10I$. In the case of the logistic regression task, the initialization is similar except that the initial covariance matrix is equal to $0.5I$.

We recall that the source code can be found at https://github.com/tGuilmeau/Projected_Stochastic_NGVI/.

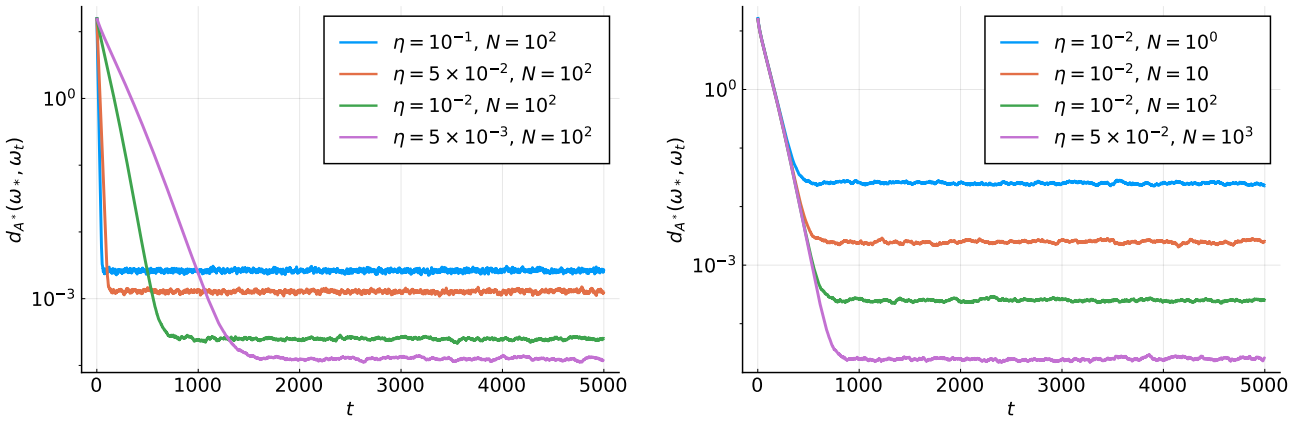
F.2 Bonnet and Price estimators with a Gaussian target

In this Section, we consider the same setting as in Section 7, namely a Gaussian target in dimension $d = 10$ with Gaussian approximating distributions, and Algorithm 1 being run with the Bonnet and Price gradient estimators without performing a projection step ($C = \mathbb{R}^d \times \mathbb{S}^d$). In this setting, we investigate the influence of the hyperparameters of Algorithm 1.

Impact of η and N in the fixed step sizes and fixed sample sizes schedule When $\eta_t \equiv \eta$ and $N_t \equiv N$, Proposition 4 predicts a geometric convergence, with rate $(1-\eta)$, of the iterates of Algorithm 1 to a neighborhood of the optimum ω_*, whose size is controlled by $\sigma^2_{\eta,N}$, defined in Definition 4. We further experimentally investigate the effect of η and N.

Figure 3a shows the performance of Algorithm 1 across different values of the step size η. Figure 3a shows that low values of η yield a slower convergence, but to a better value. Indeed, the slower convergence comes from the fact that the geometrically decreasing term in Proposition 4 decreases with $(1-\eta)^t$. The fact that the iterates are closer to ω_* when η is low indicates that $\sigma^2_{\eta,N}$ decreases with the step size η.

Figure 3b shows the performance of Algorithm 1 when different values of N are considered, keeping η fixed. All tested values of N lead to a geometric decrease with the same rate, since all runs share the same step size η, but the size of the neighborhood around the solution ω_* decreases as N increases. This can be expected from Proposition 7, which exactly predicts that $\sigma^2_{\eta,N} \leq \frac{V}{N}$ for some $V > 0$, uniformly in $\eta \in (0, 1]$.



(a) $N = 100$ and different values of η are tested

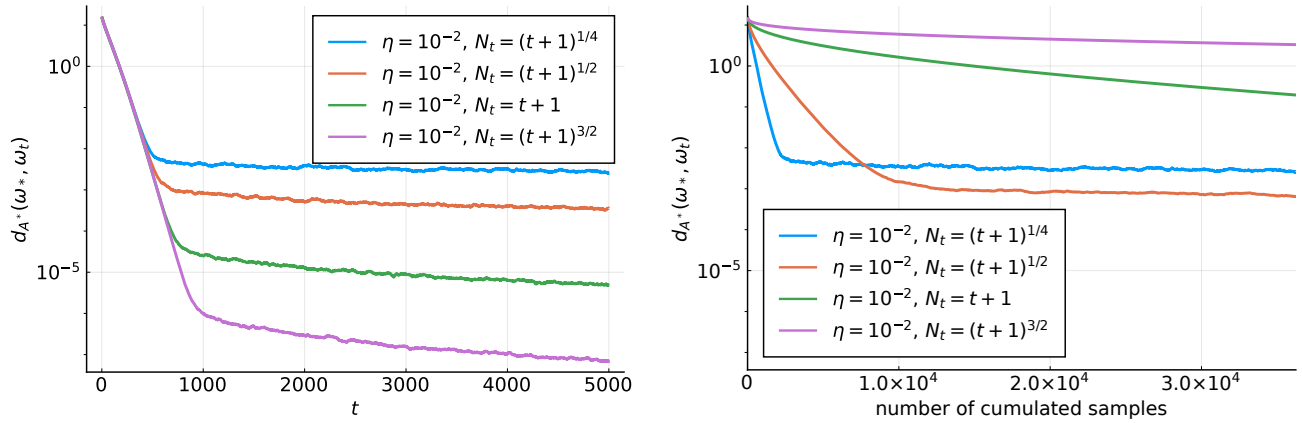
(b) $\eta = 10^{-2}$ and different values of N are tested

Figure 3: Mean Bregman divergence between current and optimal parameters, over 100 runs, for NGVI schedules with constant η and constant sample size N

Influence of the parameter γ in the choice of the sample size In Proposition 4, we showed that when constant step sizes $\eta_t \equiv \eta$ are chosen with increasing sample sizes $N_t = (t+1)^\gamma$, then, the iterates converge to the minimizer at a $\mathcal{O}(\frac{1}{T^\gamma})$ rate. We investigate this effect in the following.

Figure 4a shows that, in terms of iterations, a higher value of γ yields iterates that get closer to the minimizer at a faster rate, as predicted by Proposition 4. Note that in the first iterations, all the choice of γ lead to the same geometric rate, the difference appearing later.

Figure 4b highlights the compromise that needs to be made in terms of computational budget when selecting γ. Indeed, if a larget value of γ increase the performance in terms of iteration count, the resulting algorithm is more expensive in terms of computational budget. Figure 4b suggests that there is a "optimal" value γ such that, for a given computational budget, allows to obtain the best performance.



(a) wrt iterations

(b) wrt number of samples

Figure 4: Mean Bregman divergence between current and optimal parameters, over 100 runs, for NGVI schedules with constant η and increasing sample size $N_t = (t+1)^\gamma$

F.3 Projection impact in a robust regression setting

In this section, we illustrate the positive effect of a projection step in Algorithm 1 in a non-recoverable case, which is not theoretically covered by our previous results. More specifically, we consider a Bayesian Student linear regression similar to the Gaussian Bayesian regression in Section 5.2 but with a Student-distributed noise. The target posterior distribution is not conjugate with the Gaussian prior distribution. We can consider a Gaussian variational family $\mathcal{Q}$, but we are not in a recoverable case as the posterior distribution is not Gaussian and actually not of a known standard form. Nevertheless, the Bonnet and Price estimator g^N of the gradient (9) is tractable and can be computed in order to implement Algorithm 1.

We assume $X \sim p_0 = \mathcal{N}(\mu_0, \Sigma_0)$ with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ with

$$Y|X = x \sim \mathcal{S}(x^\top z, \sigma^2, \varrho),$$

where $z \in \mathbb{R}^d$ is a vector of fixed covariate values associated to Y, $\mathcal{S}(x^\top z, \sigma^2, \varrho)$ denotes the Student distribution with mean $x^\top z$, scale σ^2 and degrees-of-freedom parameter ϱ. Given M data points $\{y_m\}_{m=1}^M$ and associated covariates $\{z_m\}_{m=1}^M$, the goal is to approximate the posterior $\pi(x|y_1, \dots, y_M)$. In this setting, g^N can be computed with (9) using that $\nabla \log \pi(x)$ and $\nabla^2 \log \pi(x)$ are available in closed-form,

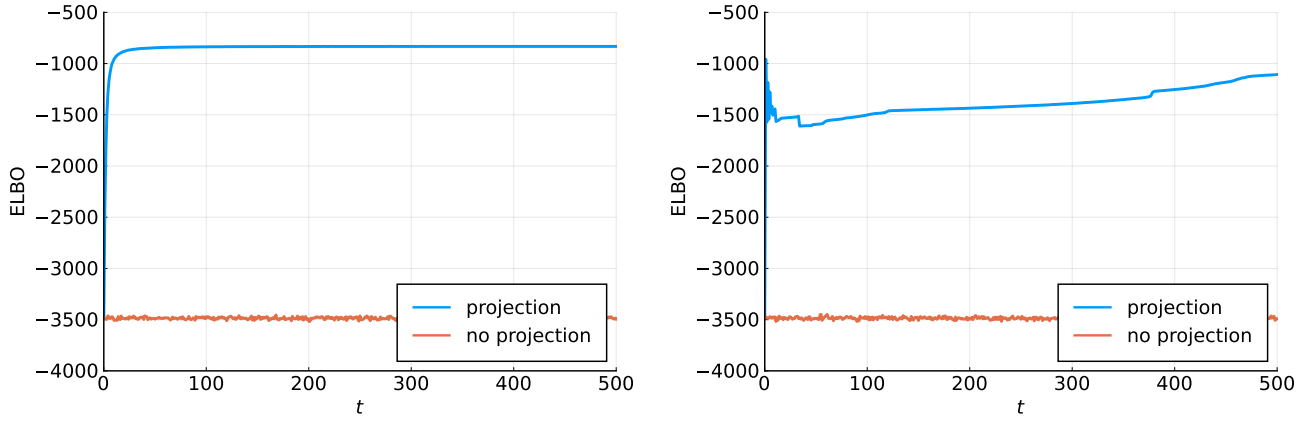
$$\nabla \log \pi(x) = \sum_{m=1}^M \frac{(\varrho+1)\,(y_m - z_m^\top x)}{\varrho\sigma^2 + (y_m - z_m^\top x)^2}\, z_m - \Sigma_0^{-1}(x - \mu_0)$$

$$\nabla^2 \log \pi(x) = -\sum_{m=1}^M \left[\frac{(\varrho+1)\left(\varrho\sigma^2 - (y_m - z_m^\top x)^2\right)}{(\varrho\sigma^2 + (y_m - z_m^\top x)^2)^2}\right] z_m z_m^\top - \Sigma_0^{-1}\,.$$

For a numerical illustration, we consider a subset of the CO and NOx (2019) dataset (Kaya et al., 2019). We use the first $M = 715$ samples given in this dataset for year 2013. The so-called turbine energy yield (TEY) variable is used as a response variable Y and the $d = 10$ other variables available in the dataset as fixed covariates. All variables are standardized for numerical stability. To significantly depart from the Gaussian case, we set $\varrho = 3$ to get an heavy-tailed noise distribution. We then set $\mu_0 = 0$, $\Sigma_0 = 5I$ and $\sigma^2 = 1$.

The main difficulty in this setting is that π is not log-concave, as it can be observed from the expression of $\nabla^2 \log \pi$ above. Therefore, there is a risk that the covariance adaptation fails, leading to singular covariance matrices, or covariance matrices very close to zero. In order to counteract this, we propose to constrain the search to the set $C = \{(\mu, \Sigma + \mu\mu^\top) \in \mathcal{H} \text{ s.t. } \alpha I \preccurlyeq \Sigma \preccurlyeq \beta I\}$ defined in (21). Here, we choose $\alpha = 10^{-4}$ and $\beta = 10^4$. The Bregman projection operator to this set is computed in Section E.1. Projection to this set ensures that we are working with covariance matrices whose conditioning is controlled. We run the algorithms initialized from the prior distribution, which is a Gaussian distribution with mean μ_0 and covariance Σ_0.

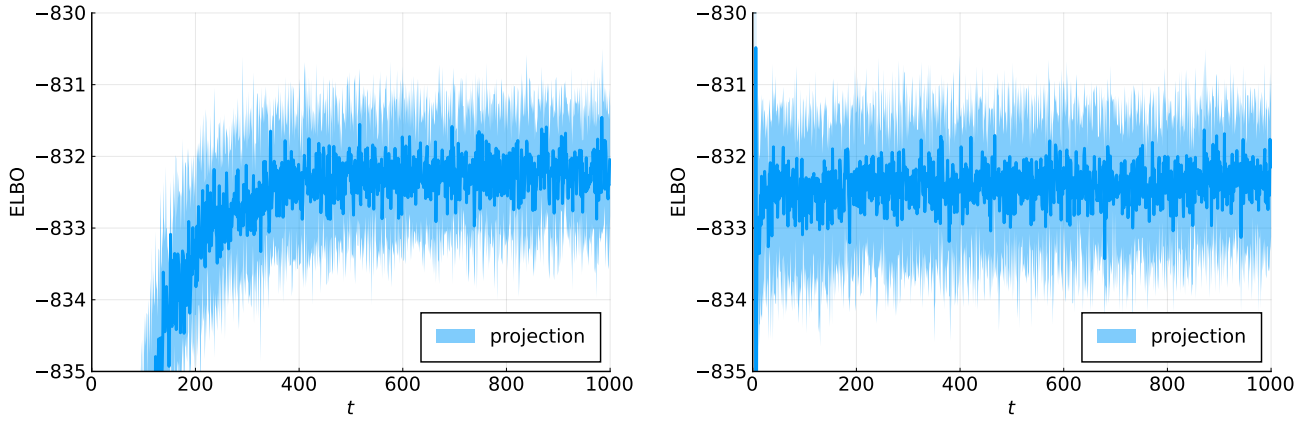
Results in terms of averaged ELBO are shown in Figure 5, comparing the performance of Algorithm 1 with and without a projection step, in case of schedules with constant step sizes and constant or increasing sample sizes. We can observe that when no projection is performed, the ELBO stays constant across iterations, indicating a failure of the algorithm. However, adding the projection step allows to obtain a significant increase in ELBO, showcasing the positive impact of adding a projection step to counteract the lack of log-convexity of the target.



(a) $\eta_t = 5 \cdot 10^{-3}$ and $N_t = 250$ (b) $\eta_t = 2/(t+2)$ and $N_t = 250$

Figure 5: Student linear regression: Averaged ELBO over 50 runs, for different NGVI schedules, comparing for each schedule the algorithm with and without a projection step

Figure 5b shows that although the ELBO increases across iterations, this increase is non-monotonic compared to the one observed in Figure 5a. This behaviour can be explained by the presence of outlier runs among the runs used to compute the averaged ELBO. Figure 6 shows, for the same respective settings, the median and inter-quartile intervals over 50 runs of Algorithm 1 with projection. The observed monotonic convergence confirms that the non-monotonic behaviour observed in Figure 5b is due to the presence of a small number of outlier runs. If needed, these runs can be detected and stopped early. Note that compared to the schedule with constant step size, the schedule with decreasing step size imposes to take an initial step size equal to one, which amounts to forgetting about the initial condition, and may be the cause of this instability.



(a) $\eta_t = 5 \cdot 10^{-3}$ and $N_t = 250$ (b) $\eta_t = 2/(t+2)$ and $N_t = 250$

Figure 6: Student linear regression: Median and intervals between the quantiles of order 0.25 and 0.75 of the ELBO over 50 runs, for different NGVI schedules.

We have not represented the performance obtained by taking sample size schedules of the form $N_t = (1+t)^\gamma$. In a fashion similar to the case of constant sample size and decreasing step size, increasing sample size schedules

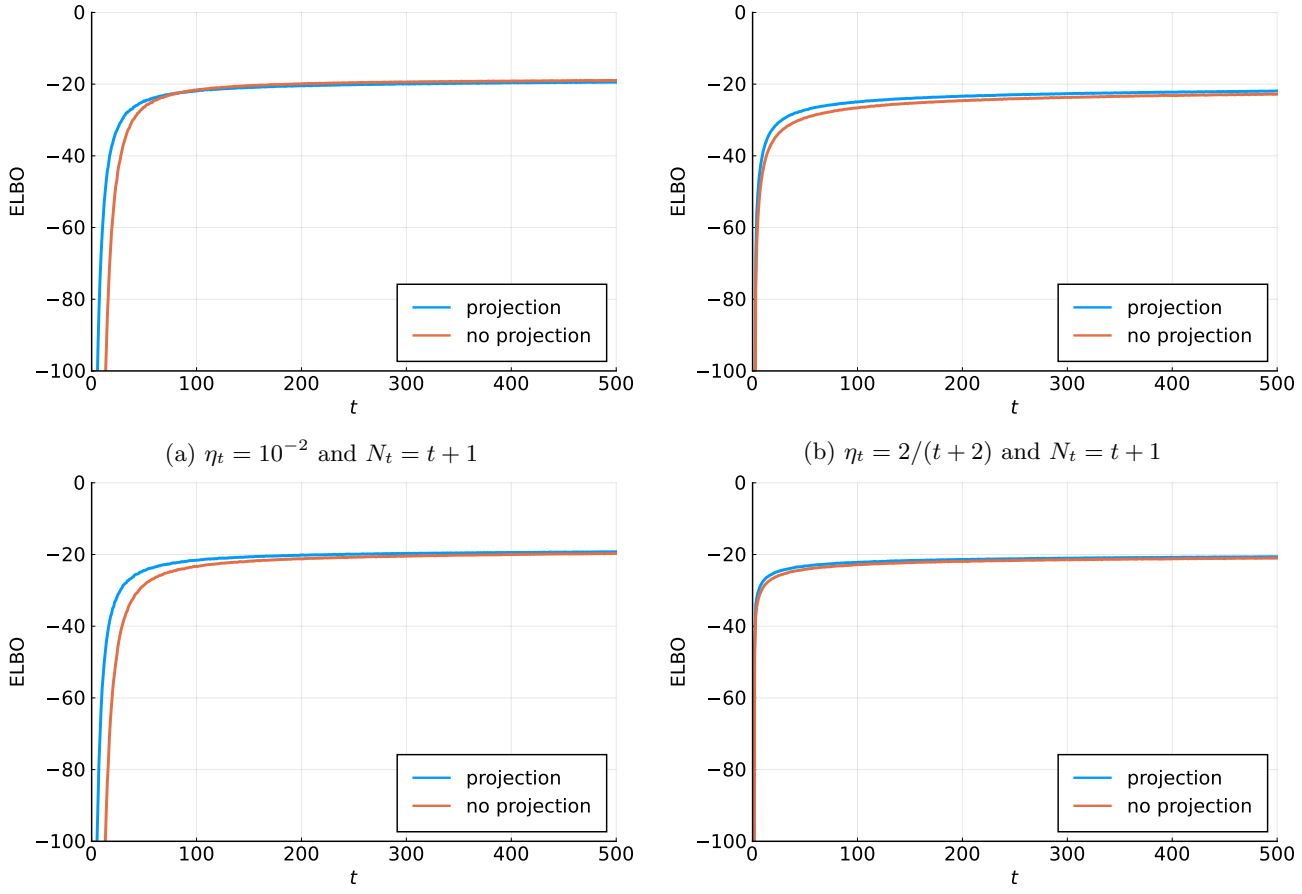
starting with $N_0 = 1$ may exhibit unstable behaviours. This can be avoided by choosing schedules of the form $N_t = \max(N, (1+t)^\gamma)$, which are covered by our theoretical results. Note also that we have not investigated the impact of parameters α and β, which control the size of the set C. This would be an interesting perspective to get more insights on the projection effect.

F.4 Projection impact in a logistic regression setting

We consider again the logistic regression task presented in Section 7. Recall that we use Gaussian distributions with diagonal covariances to approximate the logistic regression posterior with a Gaussian prior. The LERC is not satisfied in this case. The experiment is conducted on synthetic data constructed with an actual regression vector x_* whose components are all fixed being equal to 5. Therefore, $x_* \in \mathbb{R}_{\geq 0}$ in this situation.

For each step size and sample size schedule, we investigate whether an additional projection step can improve the performance of Algorithm 1. We consider the projection on the set $C = \{(\mu, \sigma^2 + \mu \bullet \mu) \in \mathbb{R}^d \times \mathbb{R}^d \text{ s.t. } \mu \in \mathbb{R}^d_{\geq 0}\}$, whose associated projection operator is computed in Section E.2. All the runs are initialized with initial mean simulated uniformly in $[-5, 5]^d$ and initial covariance matrix being equal $0.5I$.

Results are depicted in Figure 7, which shows that in this particular example, adding the projection operator improves upon the performance of the algorithm for every step size and sample size schedule. Indeed, we see that the ELBO obtained when using the projection is higher for the same iteration count than the ELBO obtained without the projection. Eventually, the two curves reach the same value.



(a) $\eta_t = 10^{-2}$ and $N_t = t + 1$

(b) $\eta_t = 2/(t+2)$ and $N_t = t + 1$

(c) $\eta_t = 10^{-2}$ and $N_t = 10^2$

(d) $\eta_t = 2/(t+2)$ and $N_t = 10^2$

Figure 7: Logistic regression: Averaged ELBO over 50 runs, for different NGVI schedules, comparing for each schedule the algorithm with and without a projection step