Bregman Proximal Langevin Monte Carlo via Bregman–Moreau Envelopes

Tim Tsz-Kit Lau [1] **Han Liu** [2 1]

Abstract

We propose efficient Langevin Monte Carlo algorithms for sampling distributions with nonsmooth convex composite potentials, which is the sum of a continuously differentiable function and a possibly nonsmooth function. We devise such algorithms leveraging recent advances in convex analysis and optimization methods involving Bregman divergences, namely the Bregman–Moreau envelopes and the Bregman proximity operators, and in the Langevin Monte Carlo algorithms reminiscent of mirror descent. The proposed algorithms extend existing Langevin Monte Carlo algorithms in two aspects—the ability to sample nonsmooth distributions with mirror descent-like algorithms, and the use of the more general Bregman–Moreau envelope in place of the Moreau envelope as a smooth approximation of the nonsmooth part of the potential. A particular case of the proposed scheme is reminiscent of the Bregman proximal gradient algorithm. The efficiency of the proposed methodology is illustrated with various sampling tasks at which existing Langevin Monte Carlo methods are known to perform poorly.

1. Introduction

The problem of sampling efficiently from high-dimensional log-Lipschitz-smooth and (strongly) log-concave target distributions via discretized Langevin diffusions has been extensively studied in the machine learning and statistics literature lately. A thorough understanding of the nonasymptotic convergence properties of Langevin Monte Carlo (LMC) has been developed, where the log-Lipschitz-smoothness and (strong) log-concavity of the density play a vital role in characterizing its convergence rates. However, such conditions are not always satisifed in applications and there is recent effort to move beyond such scenarios. On the other hand, since the efficiency of LMC algorithms in the stanard Euclidean space heavily hinges on the shape of the target distributions, algorithms based on Riemannian Langevin diffusions (Girolami & Calderhead, 2011) are considered in the case of ill-conditioned target distributions to exploit the local geometry of the log-density. However, algorithms derived by discretizing such Riemannian Langevin diffusions are notoriously hard to analyze, depending on the choice of the Riemannian metric.

In this paper, we propose two Riemannian LMC algorithms based on Bregman divergences to efficiently sample from high-dimensional distributions whose potentials (i.e., negative log-densities) are possibly not strongly convex nor (globally) Lipschitz smooth in the standard Euclidean geometry, but only strongly convex and Lipschitz smooth relative to a Legendre function subsequent to a smooth approximation. To be more precise, potentials can take the form of the sum of a relatively smooth part and a nonsmooth part (which includes the convex indicator function of a closed convex set) in the standard Euclidean geometry. A smooth approximation of the nonsmooth part based on the Bregman divergence is used and we instead sample from the smoothened distribution. By tuning a parameter of the smooth approximation, such a smoothened distribution is sufficiently close to the original target distribution. On the other hand, motivated by the connection between Langevin algorithms and convex optimization, the proposed algorithms can be viewed as the sampling analogue of the Bregman proximal gradient algorithm (Van Nguyen, 2017; Bauschke et al., 2017; Bolte et al., 2018; Bùi & Combettes, 2021; Chizat, 2021) (cf. mirror descent in the smooth case), in which Riemannian structures of the algorithms are induced by the Hessian of some Legendre function. This specific choice of the Riemannian metric also offers us a principled way to analyze the behavior of the proposed algorithms.

1.1. Langevin and Mirror-Langevin Monte Carlo Algorithms

We consider the problem of sampling from a probability measure π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which admits a density, with slight abuse of notation, also denoted by π, with respect to

[1]Department of Statistics and Data Science, Northwestern University, Evanston, IL, USA [2]Department of Computer Science, Northwestern University, Evanston, IL, USA. Correspondence to: Han Liu <hanliu@northwestern.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

the Lebesgue measure

$$(\forall x \in \mathbb{R}^d) \quad \pi(x) = \mathrm{e}^{-U(x)} \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \, \mathrm{d}y, \tag{1}$$

where the *potential* $U \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is measurable and we assume that $0 < \int_{\mathcal{U}} \mathrm{e}^{-U(y)} \, \mathrm{d}y < +\infty$ for $\mathcal{U} \coloneqq \operatorname{dom} U$. We also write $\pi \propto \mathrm{e}^{-U}$ for (1). Usually, the number of dimensions $d \gg 1$.

To perform such a sampling task, the LMC algorithm (see e.g., Dalalyan, 2017b) is arguably the most widely-studied gradient-based MCMC algorithm, which takes the form

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \gamma \nabla U(x_k) + \sqrt{2\gamma} \, \xi_k, \tag{2}$$

where $\xi_k \overset{\text{i.i.d.}}{\sim} \mathsf{N}_d(0, \mathsf{I}_d)$ for all $k \in \mathbb{N}$ and $\gamma > 0$ is a step size. Possibly with varying step sizes, the LMC algorithm is also referred to as the *unadjusted* Langevin algorithm (ULA; Durmus & Moulines, 2017) in the literature, while applying a Metropolis–Hastings correction step at each iteration of (2) the algorithm is often referred to as the *Metropolis-adjusted* Langevin algorithm (MALA; Roberts & Tweedie, 1996). ULA is the discretization of the *overdamped* Langevin diffusion, which is the solution to the stochastic differential equation (SDE)

$$(\forall t \in [0, +\infty[) \quad \mathrm{d}X_t = -\nabla U(X_t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W_t, \tag{3}$$

where $\{W_t\}_{t \in [0, +\infty[}$ is a d-dimensional standard Wiener process (a.k.a. Brownian motion). When U is Lipschitz smooth and strongly convex, it is well known that π has the *unique* invariant measure, which is the Gibbs measure $X_\infty \propto \mathrm{e}^{-U}$. Under such (or weaker) conditions of U, nonasymptotic error bounds of ULA in terms of various dissimilarity measures of probability measures, e.g., total variation and Wasserstein distances, and KL, χ^2- and Rényi divergences, are well studied and established (see e.g., Dalalyan, 2017a; Durmus & Moulines, 2017; 2019; Durmus et al., 2019; Vempala & Wibisono, 2019). To move beyond the Lipschitz smoothness assumption, we consider the case of a possibly nonsmooth composite potential U, which takes the following form

$$(\forall x \in \mathbb{R}^d) \quad U(x) \coloneqq f(x) + g(x), \tag{4}$$

where $f \in \Gamma_0(\mathbb{R}^d)$ is continuously differentiable but possibly not globally Lipschitz smooth (i.e., do not admit a globally Lipschitz gradient) and $g \in \Gamma_0(\mathbb{R}^d)$ is possibly nonsmooth (see Section 1.4 for the definition of $\Gamma_0(\mathbb{R}^d)$).

To demonstrate the sampling counterpart of mirror descent, we consider the smooth case (i.e., $g = 0$) which is well studied in the literature. Introduced in Zhang et al. (2020), under certain assumptions on U, the mirror-Langevin diffusion (MLD) takes the form: for $t \in [0, +\infty[$,

$$\begin{cases} X_t = \nabla \varphi^*(Y_t), \\ \mathrm{d}Y_t = -\nabla U(X_t) \, \mathrm{d}t + \sqrt{2} \big[\nabla^2 \varphi(X_t)\big]^{1/2} \, \mathrm{d}W_t, \end{cases} \tag{5}$$

where φ is a Legendre function and φ^* is the Fenchel conjugate of φ (see Definition 2.2). An Euler–Maruyama discretization scheme yields the Hessian Riemannian LMC (HRLMC) algorithm: for $k \in \mathbb{N}$,

$$x_{k+1} = \nabla \varphi^* \Big(\nabla \varphi(x_k) - \gamma \nabla U(x_k) + \sqrt{2\gamma} \big[\nabla^2 \varphi(x_k)\big]^{1/2} \xi_k \Big). \tag{6}$$

This is the main discretization scheme considered in Zhang et al. (2020) and an earlier draft of Hsieh et al. (2018), and further studied in Li et al. (2022), which is a specific instance of the Riemannian LMC reminiscent of the mirror descent algorithm. Ahn & Chewi (2021) consider an alternative discretization scheme motivated by the mirrorless mirror descent (Gunasekar et al., 2021), called the mirror-Langevin algorithm (MLA):

$$(\forall k \in \mathbb{N}) \quad \begin{aligned} x_{k+1/2} &= \nabla \varphi^*(\nabla \varphi(x_k) - \gamma \nabla U(x_k)), \\ x_{k+1} &= \nabla \varphi^*(Y_{\gamma_k}), \end{aligned} \tag{7}$$

where

$$\begin{cases} \mathrm{d}Y_t = \sqrt{2} \big[\nabla^2 \varphi^*(Y_t)\big]^{-1/2} \, \mathrm{d}W_t \\ Y_0 = \nabla \varphi\big(x_{k+1/2}\big) = \nabla \varphi(x_k) - \gamma \nabla U(x_k). \end{cases} \tag{8}$$

However, the mirror descent-type Langevin algorithms in Hsieh et al. (2018); Zhang et al. (2020); Ahn & Chewi (2021) can only handle relatively smooth potentials (to a Legendre function; see Definition 2.7) but not potentials with relatively smooth plus nonsmooth parts (4) where $g \neq 0$.

1.2. Contributions

We fill this void by extending HRLMC in the following aspects: (i) the target potential U takes the form (4), i.e., $U = f + g$, where f is continuously differentiable but possibly not Lipschitz smooth yet smooth relative to a Legendre function φ, and g is possibly nonsmooth; (ii) the nonsmooth part g is enveloped by its continuously differentiable approximation, which is the Bregman–Moreau envelope (Kan & Song, 2012; Bauschke et al., 2018; Laude et al., 2020; Soueycatt et al., 2020; Bauschke et al., 2006; Chen et al., 2012), in the same vein as using the Moreau envelope (Moreau, 1962; 1965) in Brosse et al. (2017); Durmus et al. (2018); Luu et al. (2021), so that we can adapt recent convergence results for mirror-Langevin algorithms for relatively smooth potentials (Zhang et al., 2020; Ahn & Chewi, 2021; Li et al., 2022; Jiang, 2021).

The proposed sampling algorithm can be viewed as a generalized version of the Moreau–Yosida Unadjusted Langevin Algorithm (MYULA; Durmus et al., 2018; Brosse et al., 2017), and we recover MYULA if both the mirror map and the Legendre function in the smooth approximation are

chosen as $\|\cdot\|^2/2$. Similar to the resemblance of MYULA to the proximal gradient algorithm with specific choice of step sizes, the proposed discretized algorithms is also reminiscent of the Bregman proximal gradient algorithm or the Bregman forward-backward algorithm (Van Nguyen, 2017; Bauschke et al., 2017; Bolte et al., 2018; Bùi & Combettes, 2021, see Section 3.3 for details). The proposed schemes, however, are able to change the geometry of the potential through a mirror map. On the theoretical front, our convergence results reveal a biased convergence guarantee with a bias which vanishes with the step size and the smoothing parameter of the Bregman–Moreau envelope. Numerical experiments also illustrate the efficiency of the proposed algorithms. We perform various nonsmooth (composite) and/or constrained sampling tasks, including sampling from the nonsmooth anisotropic Laplace distributions, at which MYULA is known to underperform ascribed to the anisotropy. To the best of our knowledge, the proposed algorithms are the *first* gradient-based Monte Carlo algorithms based on the *overdamped* Langevin dynamics which are able to sample *nonsmooth* composite distributions while adapting to the *geometry* of such distributions.

1.3. Related Work

1.3.1. Mirror Descent-Type Sampling Algorithms

In addition to Zhang et al. (2020); Ahn & Chewi (2021), Hsieh et al. (2018) introduce the mirrored-Langevin algorithm, which is also reminiscent of mirror descent, but only with I_d instead of $\nabla^2\varphi$ in (5), which entails a standard Gaussian noise in (6). Their convergence guarantee is also based on the assumption that φ is strongly convex. Chewi et al. (2020) analyze the continuous-time MLD (5) and specialize their results to the case when the mirror map is equal to the potential, known as the Newton–Langevin diffusion due to its resemblance to the Newton's method in optimization. Li et al. (2022) improve upon the analysis of Zhang et al. (2020), establishing a vanishing bias with the step size of the mirror Langevin algorithm under more relaxed assumptions.

1.3.2. Nonsmooth Sampling

Sampling efficiently from nonsmooth distributions remains a crucial problem in machine learning, statistics and imaging sciences. In particular, a significant amount of work borrows tools from convex/variational analysis and proximal optimization, i.e., the Moreau envelope and proximity operator, attributing their use to the connection between sampling and optimization, see e.g., Pereyra (2016); Brosse et al. (2017); Bubeck et al. (2018); Durmus et al. (2019; 2018); Mou et al. (2019); Wibisono (2019); Luu et al. (2021); Lee et al. (2021); Lehec (2021); Liang & Chen (2021). Nonasymptotic convergence guarantees are generally obtained from the (Metropolis-adjusted) Langevin algorithms for smooth potentials. A notable exception which does not use the Moreau envelope as a smooth approximation is Chatterji et al. (2020), which applies Gaussian smoothing instead.

1.3.3. Bregman Divergences in Convex Analysis, Optimization and Machine Learning

The origin of convex analysis results involving Bregman divergences (Bregman, 1967) and related optimization methods date backs to more than four decades ago (see e.g., Bauschke & Borwein, 1997; Bauschke & Lewis, 2000; Bauschke et al., 2001; Bauschke & Borwein, 2001; Bauschke, 2003; Bauschke et al., 2003; 2006; 2009; Nemirovski, 1979; Nemirovski & Yudin, 1983). The work by Bauschke et al. (2017) is a major recent breakthrough which revives much interest in developing new optimization algorithms involving Bregman divergences and their convergence results (see e.g., Bùi & Combettes, 2021; Bolte et al., 2018; Bauschke et al., 2019; Dragomir et al., 2021b;a; Hanzely et al., 2021; Teboulle, 2018; Takahashi et al., 2021; Chizat, 2021). Bauschke et al. (2017) relax the globally Lipschitz gradient assumption commonly required in gradient descent or proximal gradient for convergence, by introducing the relative smoothness condition (Definition 2.7). Our proposed sampling algorithms also rely on such an insightful condition. Another long line of work studies the generalization of the notions of the classical Moreau envelope and the proximity operators (Moreau, 1962; 1965; Rockafellar & Wets, 1998; Bauschke & Combettes, 2017) in convex analysis using Bregman divergences, see e.g., Bauschke et al. (2003; 2006); Chen et al. (2012); Kan & Song (2012); Bauschke et al. (2018); Laude et al. (2020); Soueycatt et al. (2020). This line of work motivates our use of the Bregman–Moreau envelopes as smooth approximations of the nonsmooth part of the potential. While there is an extensive amount of literature regarding the applications of Bregman divergences in machine learning other than mirror descent (Bubeck, 2015), we refer to Blondel et al. (2020) which includes useful results for sampling distributions on various convex polytopes such as the probability simplex based on our proposed schemes.

1.4. Notation

We denote by $I_d \in \mathbb{R}^{d\times d}$ the $d \times d$ identity matrix. We also define $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{+\infty\}$. Let $\mathbb{S}^d_{++}$ denote the set of symmetric positive definite matrices of $\mathbb{R}^{d\times d}$. Let $\mathcal{H}$ be a real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$ and a norm $\|\cdot\|$. The domain of a function $f\colon \mathcal{H} \to \overline{\mathbb{R}}$ is $\operatorname{dom} f \coloneqq \{x \in \mathcal{H} : f(x) < +\infty\}$. The set $\Gamma_0(\mathbb{R}^d)$ denotes the class of lower-semicontinuous convex functions from $\mathbb{R}^d$ to $\overline{\mathbb{R}}$ with a nonempty domain (i.e., proper). The convex indicator function $\iota_{\mathcal{C}}(x)$ of a closed convex set $\mathcal{C} \neq \varnothing$ at x equals 0 if $x \in \mathcal{C}$ and $+\infty$ otherwise. We denote by $\mathcal{B}(\mathbb{R}^d)$

the Borel σ-field of $\mathbb{R}^d$. For two probability measures μ and ν on $\mathcal{B}(\mathbb{R}^d)$, the total variation distance between μ and ν is defined by $\|\mu-\nu\|_{\mathrm{TV}} = \sup_{\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)} |\mu(\mathsf{A})-\nu(\mathsf{A})|$. For $k\in\mathbb{N}$, we denote by $\mathscr{C}^k$ the set of k-times continuously differentiable functions $f\colon \mathbb{R}^d\to\mathbb{R}$. If f is a Lipschitz function, i.e., there exists $L>0$ such that for all $x,y\in\mathbb{R}^d$, $|f(x)-f(y)|\leqslant L\|x-y\|$, then we denote $\|f\|_{\mathrm{Lip}} := \inf\{|f(x)-f(y)|/\|x-y\| \,|\, x,y\in\mathbb{R}^d, x\neq y\}$.

2. Preliminaries

In this section, we give definitions of important notions from convex analysis (Rockafellar, 1970; Rockafellar & Wets, 1998; Bauschke & Combettes, 2017), and state some related properties of such notions. In this section, we let $\varphi\in\Gamma_0(\mathbb{R}^d)$ and $\mathcal{X} := \operatorname{int}\operatorname{dom}\varphi$.

Definition 2.1 (Legendre functions)**.** A function φ is called (i) *essentially smooth*, if it is differentiable on $\mathcal{X}\neq\varnothing$ and $\|\nabla\varphi(x_n)\|\to+\infty$ whenever $x_n\to x\in\operatorname{bdry}\operatorname{dom}\varphi$; (ii) *essentially strictly convex*, if it is strictly convex on $\mathcal{X}$; (iii) *Legendre*, if it is both essentially smooth and essentially strictly convex.

Definition 2.2 (Fenchel conjugate)**.** The *Fenchel conjugate* of a proper function f is defined by $f^*(x) := \sup_{y\in\mathbb{R}^d}\{\langle y,x\rangle - f(y)\}$. For a Legendre function φ, it is well known that $\nabla\varphi\colon \mathcal{X}\to\mathcal{X}^* := \operatorname{int}\operatorname{dom}\varphi^*$ with $(\nabla\varphi)^{-1}=\nabla\varphi^*$.

Definition 2.3 (Bregman divergence)**.** The *Bregman divergence* between x and y associated with a Legendre function φ is defined through

$$D_\varphi\colon \mathbb{R}^d\times\mathbb{R}^d\to[0,+\infty]\colon (x,y)\mapsto \begin{cases}\varphi(x)-\varphi(y)-\langle\nabla\varphi(y),x-y\rangle, & \text{if } y\in\mathcal{X},\\ +\infty, & \text{otherwise.}\end{cases} \tag{9}$$

We now assume that φ is a Legendre function in the remaining part of this section.

Definition 2.4 (Bregman–Moreau envelopes)**.** For $\lambda>0$, the *left* and *right Bregman–Moreau envelopes* of $g\in\Gamma_0(\mathbb{R}^d)$ associated with φ are respectively defined by

$$\overleftarrow{\mathrm{env}}{}^{\varphi}_{\lambda,g}(x) := \inf_{y\in\mathbb{R}^d}\left\{g(y)+\frac{1}{\lambda}D_\varphi(y,x)\right\}, \tag{10}$$

and

$$\overrightarrow{\mathrm{env}}{}^{\varphi}_{\lambda,g}(x) := \inf_{y\in\mathbb{R}^d}\left\{g(y)+\frac{1}{\lambda}D_\varphi(x,y)\right\}. \tag{11}$$

Definition 2.5 (Bregman proximity operators)**.** For $\lambda>0$, the *left* and *right Bregman proximity operators* of $g\in\Gamma_0(\mathbb{R}^d)$ associated with φ are respectively defined by

$$\overleftarrow{\mathrm{P}}{}^{\varphi}_{\lambda,g}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{g(y)+\frac{1}{\lambda}D_\varphi(y,x)\right\}, \tag{12}$$

and

$$\overrightarrow{\mathrm{P}}{}^{\varphi}_{\lambda,g}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{g(y)+\frac{1}{\lambda}D_\varphi(x,y)\right\}. \tag{13}$$

We omit the arrows and write $\mathrm{env}^{\varphi}_{\lambda,g}$ and $\mathrm{P}^{\varphi}_{\lambda,g}$ when there is no need to distinguish the left and right Bregman–Moreau envelopes or Bregman proximity operators. When $\varphi = \|\cdot\|^2/2$, we recover the classical *Moreau envelope* and the *Moreau proximity operator* (Moreau, 1962; 1965). Note that $\mathrm{env}^{\varphi}_{\lambda,g}$ envelops g from below and is decreasing in λ, in a sense that $\inf g(\mathbb{R}^d)\leqslant \mathrm{env}^{\varphi}_{\lambda,g}(x)\leqslant \mathrm{env}^{\varphi}_{\kappa,g}(x)\leqslant g(x)$ for any $x\in\mathcal{X}$ and $0<\kappa<\lambda<+\infty$ (Bauschke et al., 2018, Proposition 2.2).

Definition 2.6 (Legendre strongly convex)**.** A function f is *α-Legendre strongly convex* with respect to φ if there exists a constant $\alpha\geqslant 0$ such that $f-\alpha\varphi$ is convex on $\mathcal{X}$.

Definition 2.7 (Relative smoothness)**.** A function f is β-smooth relative to φ if there exists $\beta>0$ such that $\beta\varphi-f$ is convex on $\mathcal{X}$.

3. Bregman Proximal LMC Algorithms

In the case of nonsmooth composite potentials, the mirror-Langevin algorithms (6) and (7) no longer work since the gradient of the nonsmooth part is not available. Based on the mirror Langevin algorithms for relatively smooth potentials (Zhang et al., 2020; Ahn & Chewi, 2021), we devise two possible Bregman proximal LMC algorithms involving the Bregman–Moreau envelopes and the Bregman proximity operators.

3.1. Assumptions and Related Properties

Instead of directly sampling from π, we propose to sample from distributions whose potentials being smooth surrogates of U, defined by

$$\overleftarrow{U}{}^{\psi}_{\lambda} := f+\overleftarrow{\mathrm{env}}{}^{\psi}_{\lambda,g} \quad\text{and}\quad \overrightarrow{U}{}^{\psi}_{\lambda} := f+\overrightarrow{\mathrm{env}}{}^{\psi}_{\lambda,g}, \tag{14}$$

where $\psi\in\Gamma_0(\mathbb{R}^d)$ is a Legendre function possibly different from the Legendre function φ in MLD (5) to allow full flexibility, and $\lambda>0$. Then the corresponding surrogate target densities are

$$\overleftarrow{\pi}{}^{\psi}_{\lambda}\propto\exp\left(-\overleftarrow{U}{}^{\psi}_{\lambda}\right) \quad\text{and}\quad \overrightarrow{\pi}{}^{\psi}_{\lambda}\propto\exp\left(-\overrightarrow{U}{}^{\psi}_{\lambda}\right). \tag{15}$$

We again omit the arrows and write U^{ψ}_{λ} and π^{ψ}_{λ} when we do not need to distinguish the left and right Bregman–Moreau envelopes. In this section, after introducing some required assumptions, we show that they are well-defined (i.e., in $]0,+\infty[$), as close to π by adjusting the (sufficiently small) approximation parameter $\lambda>0$, Legendre strongly log-concave, and relatively smooth. We also give some extra

assumptions of the specific algorithms (Assumption 3.8). Then, with all the assumptions of algorithms originally designed for the relatively smooth potentials satisfied by (14), we enable the capabilities of these algorithms for approximate nonsmooth sampling.

We write $\mathcal{F} \coloneqq \operatorname{dom} f \subseteq \mathbb{R}^d$ and $\mathcal{G} \coloneqq \operatorname{dom} g \subseteq \mathbb{R}^d$. Throughout the whole paper, we assume that $\mathcal{X} \coloneqq \operatorname{int}\operatorname{dom}\varphi$ and $\mathcal{Y} \coloneqq \operatorname{int}\operatorname{dom}\psi$ such that $(\operatorname{int}\mathcal{F}) \cap \mathcal{Y} \subseteq \overline{\mathcal{X}}$, $(\operatorname{int}\mathcal{F}) \cap \mathcal{Y} \cap \mathcal{X} \neq \varnothing$ and $\mathcal{G} \cap \mathcal{Y} \neq \varnothing$. Let us recall that the potential U has the form $f + g$. We make the following assumptions on the functions f and g, the Legendre functions φ and ψ, and their associated Bregman divergences D_φ and D_ψ.

Assumption 3.1. *The function $f\colon \mathbb{R}^d \to \mathbb{R}$ is (i) in $\Gamma_0(\mathbb{R}^d)$, lower bounded and differentiable (i.e., of $\mathscr{C}^1$) but may not admit a* globally Lipschitz gradient*; (ii) β_f-*smooth *relative to φ.*

Assumption 3.2. *The function $g\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is (i) in $\Gamma_0(\mathbb{R}^d)$, lower bounded and possibly* nonsmooth*; either (ii†) such that e^{-g} is* integrable *with respect to the Lebesgue measure, or (ii‡)* Lipschitz*.*

Assumption 3.3. *The function $\varphi \in \Gamma_0(\mathbb{R}^d)$ is (i)* Legendre*; (ii) of $\mathscr{C}^3$ on $\mathcal{X}$; (iii)* supercoercive*, i.e., $\lim_{\|x\|\to+\infty} \varphi(x)/\|x\| = +\infty$; (iv) M_φ-modified self-concordant (Zhang et al., 2020; Li et al., 2022, **(A1)**), i.e., there exists $M_\varphi \in [0, +\infty[$ such that for any $(x, \widetilde{x}) \in \mathcal{X} \times \mathcal{X}$, $\left\|\left|(\nabla^2\varphi(x))^{1/2} - (\nabla^2\varphi(\widetilde{x}))^{1/2}\right|\right\|_{\mathrm{F}} \leqslant \sqrt{M_\varphi}\|\nabla\varphi(x) - \nabla\varphi(\widetilde{x})\|$, where $|||\cdot|||_{\mathrm{F}}$ is the Frobenius norm. (v)* very strictly convex*, i.e., $\nabla^2\varphi(x) \in \mathbb{S}^d_{++}$ for all $x \in \mathcal{X} \neq \varnothing$ (Bauschke & Lewis, 2000).*

Assumption 3.4. *The function $\psi \in \Gamma_0(\mathbb{R}^d)$ is (i)* Legendre*; (ii) of $\mathscr{C}^3$ on $\mathcal{Y}$; (iii)* supercoercive*.*

Assumption 3.5. *The Bregman divergence associated with ψ satisfy the following assumptions: (i) D_ψ is jointly convex, i.e., convex on $\mathbb{R}^d \times \mathbb{R}^d$; (ii) $(\forall y \in \mathcal{Y})$ $D_\psi(y, \cdot)$ is strictly convex on $\mathcal{Y}$, continuous on $\mathcal{Y}$, and coercive, i.e., $(\forall y \in \mathcal{Y})$ $D_\psi(y, z) \to +\infty$ as $\|z\| \to +\infty$.*

Assumptions 3.1 to 3.4 are required for the convergence of the proposed algorithms. Assumption 3.2(ii) is required for (15) to be well-defined. Assumption 3.5 consists of the standard assumptions required for the well-posedness of the Bregman–Moreau envelopes and the Bregman proximity operators of ψ (Bauschke et al., 2018). Proposition 3.6 below implies that the densities (15) are well-defined and as close to the target density π as required when λ is sufficiently small (in total variation distance). We also provide a computable error bound when evaluating exactly the expectation with respect to (15) as opposed to the true target distribution π.

Proposition 3.6. *Suppose that Assumptions 3.1, 3.2, 3.4 and 3.5 hold. Then the following statements hold.*

(a) Let $\lambda > 0$. If either (i) Assumption 3.2(ii†) holds or (ii) Assumption 3.2(ii‡) holds and ψ is ρ-strongly convex, then $\overleftarrow{\pi}^\psi_\lambda$ and $\overrightarrow{\pi}^\psi_\lambda$ define proper densities of probability measures on $\mathbb{R}^d$, i.e., $\exp\Big(-\overleftarrow{U}^\psi_\lambda\Big)$ and $\exp\Big(-\overrightarrow{U}^\psi_\lambda\Big)$ are integrable w.r.t. the Lebesgue measure.

(b) π^ψ_λ converges to π as $\lambda \downarrow 0$, i.e., $\|\pi^\psi_\lambda - \pi\|_{\mathrm{TV}} \to 0$ as $\lambda \downarrow 0$.

(c) If Assumption 3.2(ii‡) holds and ψ is ρ-strongly convex, then for all $\lambda > 0$, $\|\pi^\psi_\lambda - \pi\|_{\mathrm{TV}} \leqslant \lambda\|g\|^2_{\mathrm{Lip}}/\rho$. In addition, for any π- and π^ψ_λ-integrable function $h\colon \mathbb{R}^d \to \mathbb{R}$,

$$\left|\mathbb{E}_{\pi^\psi_\lambda} h - \mathbb{E}_\pi h\right| \leqslant \left(\mathrm{e}^{\lambda\|g\|^2_{\mathrm{Lip}}/\rho} - 1\right) \cdot \min\left\{\mathbb{E}_{\pi^\psi_\lambda}|h|, \mathbb{E}_\pi|h|\right\}.$$

All proofs are postponed to Appendix A. Next we show that the surrogate potentials (14) are indeed continuously differentiable approximations of U under certain conditions. Their gradients and the conditions for them to be Lipschitz are also given. We also assert that the Bregman–Moreau envelopes $\operatorname{env}^\psi_{\lambda,g}(y)$ have desirable asymptotic behavior as λ goes to 0.

Proposition 3.7. *Suppose that Assumptions 3.2, 3.4 and 3.5 hold and $\lambda > 0$. The following statements hold.*

(a) The left and right Bregman–Moreau envelopes are differentiable on $\mathcal{Y}$ and

$$\nabla\overleftarrow{\operatorname{env}}^\psi_{\lambda,g}(y) = \frac{1}{\lambda}\nabla^2\psi(y)\Big(y - \overleftarrow{\mathrm{P}}^\psi_{\lambda,g}(y)\Big), \tag{16}$$

and

$$\nabla\overrightarrow{\operatorname{env}}^\psi_{\lambda,g}(y) = \frac{1}{\lambda}\Big(\nabla\psi(y) - \nabla\psi\Big(\overrightarrow{\mathrm{P}}^\psi_{\lambda,g}(y)\Big)\Big), \tag{17}$$

for any $y \in \mathcal{Y}$, respectively.

(b) If $D_\psi(y, \cdot)$ is continuous and convex on $\mathcal{Y}$ for all $y \in \mathcal{Y}$, and $\nabla\psi$ is Lipschitz on $\mathcal{Y}$, then $\nabla\overleftarrow{\operatorname{env}}^\psi_{\lambda,g}$ and $\nabla\overrightarrow{\operatorname{env}}^\psi_{\lambda,g}$ are Lipschitz on $\mathcal{Y}$.

(c) As $\lambda \downarrow 0$, we have $\overleftarrow{\operatorname{env}}^\psi_{\lambda,g}(y) \uparrow g(y)$ and $\overrightarrow{\operatorname{env}}^\psi_{\lambda,g}(y) \uparrow g(y)$ for all $y \in \mathcal{Y}$.

Finally, we make the following extra assumptions on U^ψ_λ, φ and ψ.

Assumption 3.8. *We assume the following: For $\lambda > 0$, (i) $\overleftarrow{U}^\psi_\lambda$ and $\overrightarrow{U}^\psi_\lambda$ are* α-Legendre strongly convex *with respect to φ; (ii) $\overleftarrow{\operatorname{env}}^\psi_{\lambda,g}$ and $\overrightarrow{\operatorname{env}}^\psi_{\lambda,g}$ are* β_g-smooth *relative to φ.*

Assumption 3.8 is a set of rather generic assumptions but gives us guidance to choose ψ and φ. Note that if f is *α-Legendre strongly convex* with respect to φ, then Assumption 3.8(i) is automatically satisfied. Also note that the constants α and β_g can be different for the left and right versions of their corresponding quantities.

Remark 3.9. Let us define $\beta := \beta_f + \beta_g$. Assumption 3.1(iv) and Assumption 3.8(ii) implies U_λ^ψ is β-smooth relative to φ. Then, U_λ^ψ satisfies (A2) and (A3) of Li et al. (2022), which are required for the convergence of HRLMC.

We propose two mirror-Langevin algorithms which use different discretizations of the mirror-Langevin diffusion. We give the details of the algorithm based on HRLMC in the main text. The algorithm based on MLA, which is essentially (7) with U replaced by U_λ^ψ, coined the Bregman–Moreau mirrorless mirror-Langevin algorithm (BMMMLA), is given in Appendix C.

3.2. The Bregman–Moreau Unadjusted Mirror-Langevin Algorithm

Let $x_0 \in \mathcal{Y}$. A discretization scheme for the case of composite potentials similar to HRLMC (6), called the Bregman–Moreau unadjusted mirror-Langevin algorithm (BMUMLA), iterates, for $k \in \mathbb{N}$,

$$x_{k+1} = \nabla\varphi^*\Big(\nabla\varphi(x_k) - \gamma\nabla U_\lambda^\psi(x_k) + \sqrt{2\gamma}\big[\nabla^2\varphi(x_k)\big]^{1/2}\xi_k\Big). \quad (18)$$

More specifically, when $\varphi = \psi = \|\cdot\|^2/2$, then we have $\nabla\varphi = \mathrm{Id}$, $\nabla\varphi^* = (\nabla\varphi)^{-1} = \mathrm{Id}$, $\nabla^2\varphi = I_d$, and $(\nabla\varphi + \lambda\partial g)^{-1} = (\mathrm{Id} + \lambda\partial g)^{-1} = \mathrm{prox}_{\lambda g}$, so that BMUMLA (18) reduces to MYULA (Durmus et al., 2018). Furthermore, letting $y_k = \nabla\varphi(x_k)$ for all $k \in \mathbb{N}$, then the BMUMLA in the dual space $\nabla\varphi(\mathcal{X})$ takes the form

$$y_{k+1} = y_k - \gamma\nabla U_\lambda^\psi \circ \nabla\varphi^*(y_k) + \sqrt{2\gamma}\big[\nabla^2\varphi^*(y_k)\big]^{1/2}\xi_k. \quad (19)$$

Recent results by Li et al. (2022) show that HRLMC indeed has a vanishing bias with the step size γ, as opposed to what was conjectured in Zhang et al. (2020). An advantage of applying HRLMC over MLA is that an exact simulator of the Brownian motion of varying covariance is not needed, which is usually approximated by inner loops of Euler–Maruyama discretization in practice (Ahn & Chewi, 2021). It is however worth noting that the use of the Bregman–Moreau envelope still incurs bias in our proposed algorithms, but can be controlled via the smoothing parameter λ (see Section 4).

3.3. Reminiscence of Bregman Proximal Gradient Algorithm via Right BMUMLA

The proposed BMUMLA can be simplified by specifying $\psi = \varphi$, regarding the iterates and the assumptions. In particular, the right BMUMLA reduces to

$$\nabla\varphi(x_{k+1}) = \Big(1 - \frac{\gamma}{\lambda}\Big)\nabla\varphi(x_k) - \gamma\nabla f(x_k) + \frac{\gamma}{\lambda}\nabla\varphi\Big(\overrightarrow{\mathrm{P}}_{\lambda,g}^{\varphi}(x_k)\Big) + \sqrt{2\gamma}\big[\nabla^2\varphi(x_k)\big]^{1/2}\xi_k, \quad (20)$$

which can be viewed as the generalization of MYULA (Durmus et al., 2018) with the right Bregman–Moreau envelope, but with a diffusion term of varying covariance. Furthermore, if we let $\gamma = \lambda$, then (20) becomes

$$x_{k+1} = \nabla\varphi^*\Big(\nabla\varphi\Big(\overrightarrow{\mathrm{P}}_{\lambda,g}^{\varphi}(x_k)\Big) - \lambda\nabla f(x_k) + \sqrt{2\gamma}\big[\nabla^2\varphi(x_k)\big]^{1/2}\xi_k\Big), \quad (21)$$

which roughly resembles the iterates of the Bregman proximal gradient algorithm (Van Nguyen, 2017; Bauschke et al., 2017; Bolte et al., 2018; Bùi & Combettes, 2021; Chizat, 2021), which takes the form

$$x_{k+1} = \overleftarrow{\mathrm{P}}_{\lambda,g}^{\varphi}(\nabla\varphi^*(\nabla\varphi(x_k) - \lambda\nabla f(x_k))), \quad (22)$$

or

$$z_{k+1} = \nabla\varphi^*\Big(\nabla\varphi\Big(\overleftarrow{\mathrm{P}}_{\lambda,g}^{\varphi}(z_k)\Big) - \lambda\nabla f\Big(\overleftarrow{\mathrm{P}}_{\lambda,g}^{\varphi}(z_k)\Big)\Big), \quad (23)$$

if we write $z_k = \nabla\varphi^*(\nabla\varphi(x_k) - \lambda\nabla f(x_k))$. The differences between (23) and (21), other than the diffusion term, are the use of different Bregman–Moreau envelopes and the argument of the gradient of the smooth part.

Another advantage of using the same mirror map is that Assumption 3.8 can be made more precise. In particular, regarding Assumption 3.8(ii), since $\overrightarrow{\mathrm{env}}_{\lambda,g}^{\varphi}$ is λ^{-1}-smooth relative to φ (Laude et al., 2020, Proposition 3.8(ii)), for the right BMUMLA with $\psi = \varphi$, i.e., (20), Assumption 3.8(ii) is made precise with a relative smoothness constant $\beta_g = \lambda^{-1}$ for $\overrightarrow{\mathrm{env}}_{\lambda,g}^{\varphi}$.

4. Convergence Analysis

We now state the main convergence results derived from Li et al. (2022). To quantify the convergence, we introduce a modified Wasserstein distance previously introduced by Zhang et al. (2020) and further applied in the analysis of Li et al. (2022).

Definition 4.1. For two probability measures μ and ν on $\mathcal{B}(\mathcal{X})$, the (squared) *modified Wasserstein distance* under the mirror map $\nabla\varphi$ from μ to ν is defined by

$$\mathsf{W}_{2,\varphi}^2(\mu,\nu) := \inf_{u\sim\mu, v\sim\nu} \mathbb{E}\big[\|\nabla\varphi(u) - \nabla\varphi(v)\|^2\big].$$

Note that if $\widetilde{\mu} := (\nabla\varphi)_\sharp\mu$ and $\widetilde{\nu} := (\nabla\varphi)_\sharp\nu$ are the push-forward measures of μ and ν by $\nabla\varphi$ respectively, then $\mathsf{W}_{2,\varphi}^2(\mu,\nu) = \mathsf{W}_2^2(\widetilde{\mu},\widetilde{\nu}) := \inf_{\widetilde{u}\sim\widetilde{\mu},\widetilde{v}\sim\widetilde{\nu}} \mathbb{E}\big[\|\widetilde{u} - \widetilde{v}\|^2\big]$.

The main convergence result is given as follows.

Theorem 4.2. *Let Assumptions 3.1 to 3.5 and 3.8 hold and $M_\varphi < \alpha/2$. Let $x_k \sim \mu_k$ be the iterates of (18) with step size $\gamma \in]0, \gamma_{\max}]$, where $\gamma_{\max} = \mathcal{O}\big((\alpha - 2M_\varphi)^2/\big(\beta^2(1+8M_\varphi)^2\big)\big)$. Then, from any $x_0 \sim \mu_0$, we have*

$$\mathsf{W}_{2,\varphi}(\mu_k, \pi_\lambda^\psi) \leqslant \sqrt{2}\mathrm{e}^{-(\alpha-2M_\varphi)\gamma k}\mathsf{W}_{2,\varphi}(\mu_0, \pi_\lambda^\psi) + C\sqrt{2\gamma}, \quad (24)$$

where $C = \mathcal{O}\Big(\beta(1+8M_\varphi)\sqrt{d}/(\alpha - 2M_\varphi)\Big)$ is a constant. Furthermore, if the stronger Assumption 3.2(ii‡) rather than (ii†) holds and ψ is ρ-strongly convex, then

$$\mathsf{W}_{2,\varphi}(\mu_k, \pi) \leqslant \sqrt{2}\mathrm{e}^{-(\alpha-2M_\varphi)\gamma k}\mathsf{W}_{2,\varphi}(\mu_0, \pi) + C\sqrt{2\gamma} + \Big(1 + \sqrt{2}\mathrm{e}^{-(\alpha-2M_\varphi)\gamma k}\Big)\frac{\eta\lambda}{\rho}\|g\|_{\mathrm{Lip}}^2, \quad (25)$$

where $\eta \coloneqq \sup_{(u,v)\in\mathfrak{X}\times\mathfrak{X}} \|\nabla\varphi(u) - \nabla\varphi(v)\|^2$.

From Theorem 4.2, we can derive a mixing time bound for (18) similar to Corollary 3.2 of Li et al. (2022).

Corollary 4.3. *Suppose that Assumptions 3.1, 3.3 to 3.5 and 3.8 and Assumption 3.2(ii‡) rather than (ii†) hold, ψ is ρ-strongly convex, and $M_\varphi < \alpha/2$. Then, for any target accuracy $\varepsilon > 0$, in order to achieve $\mathsf{W}_2(\widetilde{\mu}_k, \widetilde{\pi}) \leqslant \varepsilon$, it suffices to run BMUMLA in the dual space (19) with step size $\gamma = \varepsilon^2/(18C^2)$ and smoothing parameter $\lambda = \rho\varepsilon/(3\widetilde{\eta}\|g\|_{\mathrm{Lip}}^2)$ for k iterations, where*

$$k \geqslant \frac{1}{(\alpha - 2M_\varphi)\gamma}\log\left(\frac{3\sqrt{2}[\mathsf{W}_2(\widetilde{\mu}_0, \widetilde{\pi}_\lambda^\psi) + \widetilde{\eta}\lambda\|g\|_{\mathrm{Lip}}^2/\rho]}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\frac{\beta^2(1+8M_\varphi)^2 d}{(\alpha - 2M_\varphi)^3\varepsilon^2}\right), \quad (26)$$

where $\widetilde{\eta} \coloneqq \sup_{(u,v)\in\mathfrak{X}\times\mathfrak{X}} \|u - v\|^2$.

Assuming all other constants including α, β, M_φ, η, ρ and $\|g\|_{\mathrm{Lip}}$ are independent of d, then, similar to Li et al. (2022) for the relatively smooth case, (18) has a biased convergence guarantee with a bias incurred by the algorithm which scales as $\mathcal{O}(\sqrt{d\gamma})$. Since essentially we are sampling from the surrogate distribution π_λ^ψ which is different from π, this incurs an additional bias. From (25), this bias attributed to smoothing with the Bregman–Moreau envelope scales as $\mathcal{O}(\lambda)$ for large enough k and $\eta < +\infty$. We then obtain the same mixing time bound of $\widetilde{\mathcal{O}}(d/\varepsilon^2)$ for BMUMLA as the one for MLA in Li et al. (2022). Note that the appearance of η limits the choice of mirror maps in (18) as some choices of φ might not give a bounded η (see Jiang, 2021, for related discussion).

5. Numerical Experiments

We perform numerical experiments of sampling anisotropic Laplace distributions which have nonsmooth potentials. Other additional numerical experiments are given in Appendix D. In this section, we use bold lower case letters $\boldsymbol{\theta} = (\theta_i)_{1\leqslant i\leqslant d}^\top \in \mathbb{R}^d$ to denote vectors. All numerical implementations can be found at `https://github.com/timlautk/bregman_prox_langevin_mc`.

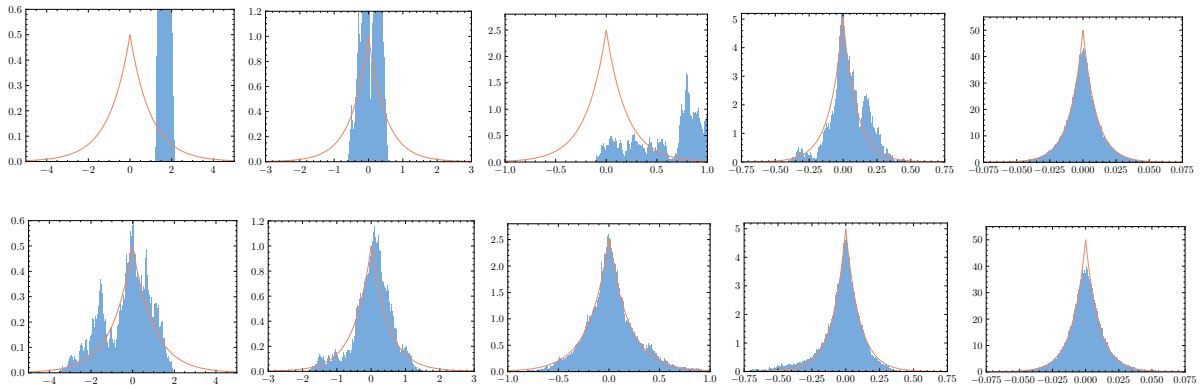
For such a nonsmooth sampling task, inspired by Vorstrup Goldman et al. (2021); Bouchard-Côté et al. (2018), we consider the case where $f = 0$ and $g(\boldsymbol{\theta}) = \|\boldsymbol{\alpha} \odot \boldsymbol{\theta}\|_1 = \sum_{i=1}^d \alpha_i|\theta_i|$ with $\boldsymbol{\alpha} = (1, 2, \ldots, d)^\top$. This is an example in which MYULA is known to perform poorly due to the *anisotropy* (Vorstrup Goldman et al., 2021, §4.1): with a relatively small step size, MYULA mixes fast for the *narrow* marginals, whereas it mixes slowly in the *wide* ones. To alleviate this issue, the mirror map in our proposed scheme allows to adapt to the geometry of the potential, while the square root of the Hessian of φ serves as a preconditioner of the diffusion term. We choose φ to be the $\boldsymbol{\beta}$-hyperbolic entropy (hypentropy; Ghai et al., 2020), defined by

$$\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta}) \coloneqq \sum_{i=1}^d \left[\theta_i \operatorname{arsinh}(\theta_i/\beta_i) - \sqrt{\theta_i^2 + \beta_i^2}\right],$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\beta} \in [0, +\infty[^d$. We allow β_i's to vary across different dimensions to enhance flexibility. The hypentropy interpolates between the squared Euclidean distance and the Boltzmann–Shannon entropy as $\boldsymbol{\beta}$ varies. We choose the associated Legendre function of the Bregman–Moreau envelope to be $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_{\boldsymbol{M}}^2 = \frac{1}{2}\langle\boldsymbol{\theta}, \boldsymbol{M}\boldsymbol{\theta}\rangle$, where $\boldsymbol{M} = \mathrm{Diag}(\boldsymbol{\alpha}/2)$, so that ψ is strongly convex.

We apply the proposed algorithms BMUMLA and BMMMLA, and compare their performance with that of MYULA. We consider $d = 100$, draw $K = 10^5$ samples, with a tight Bregman–Moreau envelope using a small smoothing parameter $\lambda = 10^{-5}$ and a small step size $\gamma = \lambda/2$. The parameter of the hyperbolic entropy is $\boldsymbol{\beta} = (2\sqrt{d - i + 1})_{1\leqslant i\leqslant d}^\top$. Further implementation details and verification of assumptions are given in Appendix B. The marginal empirical densities are given in Figure 1 (Figure D.4 for BMMMLA in Appendix B).

In this example, MYULA does not mix fast for the wide marginals (the lower dimensions, even at the 10^{th} dimension), whereas BMUMLA and BMMMLA are able to mix equally fast across different dimensions. Although our proposed methods require knowledge of the target distribution, we expect even better mixing when $\boldsymbol{\beta}$ is better tuned or adaptively learned using certain auxiliary procedures. Moreover, a quick comparison with methods in Vorstrup Goldman et al. (2021, Figure 2) indicates that, despite being asymptotically biased (since γ and λ are chosen as constants), our pro-



(a) 1st dimension (b) 2nd dimension (c) 5th dimension (d) 10th dimension (e) 100th dimension

Figure 1. Histograms of samples (blue) from MYULA (1st row), BMUMLA (2nd row) and the true densities (orange).

posed algorithms also appear to be comparable to or even outperform some of the asymptotically exact algorithms such as pMALA (Pereyra, 2016) and the bouncy particle sampler (Bouchard-Côté et al., 2018). We however leave a comprehensive comparison with other classes of MCMC algorithms as future work.

6. Discussion

In this paper, we propose two efficient Bregman proximal Langevin algorithms for efficient sampling from nonsmooth convex composite potentials. Our proposed schemes enhance the flexibility of existing (overdamped) LMC algorithms in two aspects: the use of Bregman divergences in (i) altering the *geometry* of the problem and hence the algorithm; (ii) imposing the smooth approximation. Theoretically, our proposed schemes have a vanishing bias with the step size and the smoothing parameter of the Bregman–Moreau envelope, while numerically they outperform MYULA in sampling nonsmooth anisotropic distributions.

There are several interesting directions to extend the current work. Full gradients can be replaced by stochastic or mini-batch gradients in Langevin algorithms (see e.g., Welling & Teh, 2011; Durmus et al., 2019; Salim et al., 2019; Salim & Richtárik, 2020; Nemeth & Fearnhead, 2021) to avoid costly computation of the full gradient in high dimensions. Our proposed algorithm also has potential implications for nonconvex potentials or nonconvex optimization algorithms based on Langevin dynamics (Mangoubi & Vishnoi, 2019; Cheng et al., 2018a; Raginsky et al., 2017; Vempala & Wibisono, 2019), as the Bregman proximial gradient algorithm is able to solve nonconvex optimization algorithms (Bolte et al., 2018). We also refer to recent results of the Bregman–Moreau envelopes of nonconvex functions (Laude et al., 2020) and the use of Moreau envelope in nonsmooth sampling algorithms for computing the Exponential Weighted Aggregation (EWA) estimators (Luu et al., 2021). Recently, Jiang (2021) leverages the assumption of an isoperimetric inequality called the *mirror log-Sobolev inequality* for the target density in mirror Langevin algorithms, which is weaker than assuming a Legendre strongly convex potential. It is however unclear to see how Bregman–Moreau envelopes in the potential would satisfy this assumption and other weaker notions of relative smoothness of the potential introduced in this paper. A natural extension is to consider sampling schemes based on the underdamped Langevin dynamics (Cheng et al., 2018b) or Hamiltonian dynamics (Neal, 1993) with the Bregman–Moreau enveloped potentials, and to include the Metropolis–Hastings adjustment step to accelerate mixing. Other than the Bregman–Moreau envelope, the Bregman forward-backward envelope (Ahookhosh et al., 2021) can also be used to envelop the whole composite potential; see the recent work by Eftekhari et al. (2022) in a similar spirit using the forward-backward envelope with the overdamped Langevin algorithm. More sophisticated discretization scheme such as the explicit stabilized SK-ROCK scheme (Abdulle et al., 2018) in Pereyra et al. (2020) could also constitute new sampling schemes based on MLD. It is also interesting to compare our proposed schemes with gradient-based MCMC algorithms based on piecewise-deterministic Markov processes for nonsmooth sampling as in Vorstrup Goldman et al. (2021), e.g., the zig-zag sampler (Bierkens et al., 2019) and the bouncy particle sampler (Bouchard-Côté et al., 2018).

Acknowledgements

This work is partially supported by NIH R01LM01372201, NSF TRIPOD 1740735, NSF DMS1454377. Part of this work was done when Tim Tsz-Kit Lau was participating in the workshop "Sampling Algorithms and Geometries on

Probability Distributions" at the Simons Institute for the Theory of Computing.

References

Abdulle, A., Almuslimani, I., and Vilmart, G. Optimal explicit stabilized integrator of weak order 1 for stiff and ergodic stochastic differential equations. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):937–964, 2018.

Ahn, K. and Chewi, S. Efficient constrained sampling via the mirror-Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ahookhosh, M., Themelis, A., and Patrinos, P. A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. *SIAM Journal on Optimization*, 31 (1):653–685, 2021.

Bauschke, H. H. Duality for Bregman projections onto translated cones and affine subspaces. *Journal of Approximation Theory*, 121(1):1–12, 2003.

Bauschke, H. H. and Borwein, J. M. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

Bauschke, H. H. and Borwein, J. M. Joint and separate convexity of the Bregman distance. In *Studies in Computational Mathematics*, volume 8, pp. 23–36. Elsevier, 2001.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017.

Bauschke, H. H. and Lewis, A. S. Dykstras algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

Bauschke, H. H., Borwein, J. M., and Combettes, P. L. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3(04):615–647, 2001.

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

Bauschke, H. H., Combettes, P. L., and Noll, D. Joint minimization with alternating Bregman proximity operators. *Pacific Journal of Optimization*, 2:401–424, 2006.

Bauschke, H. H., Wang, X., Ye, J., and Yuan, X. Bregman distances and Chebyshev sets. *Journal of Approximation Theory*, 159(1):3–25, 2009.

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

Bauschke, H. H., Dao, M. N., and Lindstrom, S. B. Regularizing with Bregman–Moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

Bauschke, H. H., Bolte, J., Chen, J., Teboulle, M., and Wang, X. On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, 182(3):1068–1087, 2019.

Bierkens, J., Fearnhead, P., and Roberts, G. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics*, 47(3):1288–1320, 2019.

Blondel, M., Martins, A. F., and Niculae, V. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

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

Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200–217, 1967.

Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Proceedings of the Conference on Learning Theory (COLT)*, 2017.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8 (3-4):231–357, 2015.

Bubeck, S., Eldan, R., and Lehec, J. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Bùi, M. N. and Combettes, P. L. Bregman forward-backward operator splitting. *Set-Valued and Variational Analysis*, 29(3):583–603, 2021.

Chatterji, N., Diakonikolas, J., Jordan, M. I., and Bartlett, P. L. Langevin Monte Carlo without smoothness. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Chen, Y. Y., Kan, C., and Song, W. The Moreau envelope function and proximal mapping with respect to the Bregman distances in Banach spaces. *Vietnam Journal of Mathematics*, 40(2&3):181–199, 2012.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648v4*, 2018a.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the Conference on Learning Theory (COLT)*, 2018b.

Chewi, S., Gouic, T. L., Lu, C., Maunu, T., Rigollet, P., and Stromme, A. Exponential ergodicity of mirror-Langevin diffusions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chizat, L. Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368*, 2021.

Cordero-Erausquin, D. Transport inequalities for log-concave measures, quantitative forms, and applications. *Canadian Journal of Mathematics*, 69(3):481–501, 2017.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. On the lambert W function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

Dalalyan, A. S. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the Conference on Learning Theory (COLT)*, 2017a.

Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(79):651–676, 2017b.

Dragomir, R.-A., d'Aspremont, A., and Bolte, J. Quartic first-order methods for low-rank minimization. *Journal of Optimization Theory and Applications*, 189(2):341–363, 2021a.

Dragomir, R.-A., Taylor, A. B., d'Aspremont, A., and Bolte, J. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, pp. 1–43, 2021b.

Durmus, A. and Moulines, E. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Durmus, A. and Moulines, E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Durmus, A., Moulines, E., and Pereyra, M. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Durmus, A., Majewski, S., and Miasojedow, B. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Eftekhari, A., Vargas, L., and Zygalakis, K. The forward-backward envelope for sampling with the overdamped Langevin algorithm. *arXiv preprint arXiv:2201.09096*, 2022.

Ghai, U., Hazan, E., and Singer, Y. Exponentiated gradient meets gradient descent. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2020.

Gibbs, A. L. and Su, F. E. On choosing and bounding probability metrics. *International Statistical Review*, 70 (3):419–435, 2002.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Gunasekar, S., Woodworth, B., and Srebro, N. Mirrorless mirror descent: A more natural discretization of Riemannian gradient flow. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Hanzely, F., Richtarik, P., and Xiao, L. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2):405–440, 2021.

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jiang, Q. Mirror Langevin Monte Carlo: the case under isoperimetry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kan, C. and Song, W. The Moreau envelope function and proximal mapping in the sense of the Bregman distance. *Nonlinear Analysis: Theory, Methods & Applications*, 75 (3):1385–1399, 2012.

Lambert, J. H. Observationes variae in mathesin puram. *Acta Helvetica*, 3:128–168, 1758.

Laude, E., Ochs, P., and Cremers, D. Bregman proximal mappings and Bregman–Moreau envelopes under relative prox-regularity. *Journal of Optimization Theory and Applications*, 184(3):724–761, 2020.

Lee, Y. T., Shen, R., and Tian, K. Structured logconcave sampling with a restricted Gaussian oracle. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Lehec, J. The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*, 2021.

Li, R., Tao, M., Vempala, S. S., and Wibisono, A. The mirror Langevin algorithm converges with vanishing bias. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2022.

Liang, J. and Chen, Y. A proximal algorithm for sampling from non-smooth potentials. *arXiv preprint arXiv:2110.04597*, 2021.

Luu, T. D., Fadili, J., and Chesneau, C. Sampling from non-smooth distributions through Langevin diffusion. *Methodology and Computing in Applied Probability*, 23(4):1173–1201, 2021.

Mangoubi, O. and Vishnoi, N. K. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In *Proceedings of the Conference on Learning Theory (COLT)*, 2019.

Moreau, J.-J. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 255: 2897–2899, 1962.

Moreau, J.-J. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Mou, W., Flammarion, N., Wainwright, M. J., and Bartlett, P. L. An efficient sampling algorithm for non-smooth composite potentials. *arXiv preprint arXiv:1910.00551*, 2019.

Neal, R. M. Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1993.

Nemeth, C. and Fearnhead, P. Stochastic gradient Markov chain Monte Carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021.

Nemirovski, A. S. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15(1), 1979.

Nemirovski, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

Nesterov, Y. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2nd edition, 2018.

Pereyra, M. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

Pereyra, M., Mieles, L. V., and Zygalakis, K. C. Accelerating proximal Markov chain Monte Carlo by using an explicit stabilized method. *SIAM Journal on Imaging Sciences*, 13(2):905–935, 2020.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the Conference on Learning Theory (COLT)*, 2017.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

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

Rockafellar, R. T. and Wets, R. J.-B. *Variational Analysis*. Springer, 1998.

Salim, A. and Richtárik, P. Primal dual interpretation of the proximal stochastic gradient Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Salim, A., Kovalev, D., and Richtárik, P. Stochastic proximal Langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Soueycatt, M., Mohammad, Y., and Hamwi, Y. Regularization in Banach spaces with respect to the Bregman distance. *Journal of Optimization Theory and Applications*, 185(2):327–342, 2020.

Takahashi, S., Fukuda, M., and Tanaka, M. New Bregman proximal type algorithms for solving DC optimization problems. *arXiv preprint arXiv:2105.04873*, 2021.

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

Van Nguyen, Q. Forward-backward splitting with Bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.

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

Vorstrup Goldman, J., Sell, T., and Singh, S. S. Gradient-based Markov chain Monte Carlo for Bayesian inference with non-differentiable priors. *Journal of the American Statistical Association*, pp. 1–12, 2021.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

Wibisono, A. Proximal Langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*, 2019.

Zhang, K. S., Peyré, G., Fadili, J., and Pereyra, M. Wasserstein control of mirror Langevin Monte Carlo. In *Proceedings of the Conference on Learning Theory (COLT)*, 2020.

APPENDIX

A. Proofs of Main Text

A.1. Proof of Proposition 3.6

Proposition 3.6 includes statements similar to those in Durmus et al. (2018, Proposition 3.1) and Vorstrup Goldman et al. (2021, Theorem 3). We provide the proofs here for self-containedness. In particular, we further need ψ to be ρ-strongly convex in (c).

Proposition A.1. *Let ψ be a Legendre function and ρ-strongly convex ($\rho > 0$), then*

$$(\forall (y, \widetilde{y}) \in \mathcal{Y} \times \mathcal{Y}) \quad \frac{\rho}{2}\|y - \widetilde{y}\|^2 \leqslant D_\psi(y, \widetilde{y}). \tag{A.1}$$

Proof of Proposition A.1. By definition, ψ is ρ-strongly convex if and only if

$$\begin{aligned}(\forall (y, \widetilde{y}) \in \mathcal{Y} \times \mathcal{Y}) \quad \psi(y) \geqslant \psi(\widetilde{y}) + \langle \nabla\psi(\widetilde{y}), y - \widetilde{y}\rangle + \frac{\rho}{2}\|y - \widetilde{y}\|^2 \\ \Leftrightarrow (\forall (y, \widetilde{y}) \in \mathcal{Y} \times \mathcal{Y}) \quad \psi(y) - \psi(\widetilde{y}) - \langle \nabla\psi(\widetilde{y}), y - \widetilde{y}\rangle \geqslant \frac{\rho}{2}\|y - \widetilde{y}\|^2.\end{aligned}$$

Then the result follows from the definition of the Bregman divergence (9).

□

Proof of Proposition 3.6.

(a) (i) We first suppose that Assumption 3.2(ii†) holds. By Bauschke et al. (2018, Proposition 2.2), $U \geqslant U^\psi_\lambda$, which implies

$$0 < \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\,\mathrm{d}x < \int_{\mathbb{R}^d} \mathrm{e}^{-U^\psi_\lambda(x)}\,\mathrm{d}x.$$

It suffices to prove that $\mathrm{e}^{-\mathrm{env}^\psi_{\lambda,g}}$ is integrable (with respect to the Lebesgue measure) which in turn implies $\mathrm{e}^{-U^\psi_\lambda}$ is integrable since f is lower bounded. By Assumption 3.2(i) and Durmus et al. (2018, Lemma A.1), there exist $\rho_g > 0$, $x_g \in \mathbb{R}^d$ and $M_1 \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$,

$$g(x) - g(x_g) \geqslant M_1 + \rho_g\|x - x_g\|.$$

Then, by Definitions 2.4 and 2.5, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\overleftarrow{\mathrm{env}}^\psi_{\lambda,g}(x) - g(x_g) &= g\Big(\overleftarrow{\mathrm{P}}^\psi_{\lambda,g}(x)\Big) - g(x_g) + \frac{1}{\lambda}D_\psi\Big(\overleftarrow{\mathrm{P}}^\psi_{\lambda,g}(x), x\Big) \\ &\geqslant M_1 + \rho_g\Big\|\overleftarrow{\mathrm{P}}^\psi_{\lambda,g}(x) - x_g\Big\| + \frac{1}{\lambda}D_\psi\Big(\overleftarrow{\mathrm{P}}^\psi_{\lambda,g}(x), x\Big) \\ &\geqslant M_1 + \inf_{y\in\mathbb{R}^d}\left\{\rho_g\|y - x_g\| + \frac{1}{\lambda}D_\psi(y, x)\right\} \\ &\geqslant M_1 + \overleftarrow{\mathrm{env}}^\psi_{\lambda,h}(x),\end{aligned} \tag{A.2}$$

where $h\colon \mathbb{R}^d \to \mathbb{R}\colon x \mapsto \rho_g\|x - x_g\|$. Likewise, using the right Bregman–Moreau envelope, we have

$$\begin{aligned}\overrightarrow{\mathrm{env}}^\psi_{\lambda,g}(x) - g(x_g) &= g\Big(\overrightarrow{\mathrm{P}}^\psi_{\lambda,g}(x)\Big) - g(x_g) + \frac{1}{\lambda}D_\psi\Big(x, \overrightarrow{\mathrm{P}}^\psi_{\lambda,g}(x)\Big) \\ &\geqslant M_1 + \rho_g\Big\|\overrightarrow{\mathrm{P}}^\psi_{\lambda,g}(x) - x_g\Big\| + \frac{1}{\lambda}D_\psi\Big(x, \overrightarrow{\mathrm{P}}^\psi_{\lambda,g}(x)\Big) \\ &\geqslant M_1 + \inf_{y\in\mathbb{R}^d}\left\{\rho_g\|y - x_g\| + \frac{1}{\lambda}D_\psi(x, y)\right\} \\ &\geqslant M_1 + \overrightarrow{\mathrm{env}}^\psi_{\lambda,h}(x).\end{aligned} \tag{A.3}$$

Next, using Definition 2.4 again, for all $x \in \mathbb{R}^d$,

$$\overleftarrow{\mathrm{env}}^{\psi}_{\lambda,h}(x) = h\Big(\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x)\Big) + \frac{1}{\lambda} D_\psi\Big(\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x), x\Big) \geqslant h\Big(\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x)\Big) = \rho_g \Big\| \overleftarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x) - x_g \Big\|,$$
$$\overrightarrow{\mathrm{env}}^{\psi}_{\lambda,h}(x) = h\Big(\overrightarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x)\Big) + \frac{1}{\lambda} D_\psi\Big(x, \overrightarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x)\Big) \geqslant h\Big(\overrightarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x)\Big) = \rho_g \Big\| \overrightarrow{\mathrm{P}}^{\psi}_{\lambda,h}(x) - x_g \Big\|.$$

It follows that there exists $M_2 \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$,

$$\min\Big\{\overleftarrow{\mathrm{env}}^{\psi}_{\lambda,h}(x), \overrightarrow{\mathrm{env}}^{\psi}_{\lambda,h}(x)\Big\} \geqslant \rho_g \|x - x_g\| + M_2.$$

Combining this with (A.2) and (A.3) yields the desired result.

(ii) Now we suppose that Assumption 3.2(ii‡) holds and ψ is ρ-strongly convex, then for any $\lambda > 0$,

$$\sup_{x \in \mathbb{R}^d} \Big\{ g(x) - \mathrm{env}^{\psi}_{\lambda,g}(x) \Big\} \leqslant \frac{\lambda}{2\rho} \|g\|^2_{\mathrm{Lip}}. \tag{A.4}$$

If (A.4) holds, then

$$(\forall x \in \mathbb{R}^d) \quad U^{\psi}_{\lambda}(x) \coloneqq f(x) + \mathrm{env}^{\psi}_{\lambda,g}(x) \geqslant f(x) + g(x) - \frac{\lambda}{2\rho}\|g\|^2_{\mathrm{Lip}},$$

which implies

$$\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\lambda}_{\psi}(x)} \,\mathrm{d}x \leqslant \mathrm{e}^{\lambda \|g\|^2_{\mathrm{Lip}}/(2\rho)} \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \,\mathrm{d}x < +\infty.$$

Since Assumption 3.2(ii‡) holds, we have

$$\begin{aligned}
(\forall x \in \mathbb{R}^d) \;\; g(x) - \overleftarrow{\mathrm{env}}^{\psi}_{\lambda,g}(x) &= g(x) - \inf_{y \in \mathbb{R}^d} \left\{ g(y) + \frac{1}{\lambda} D_\psi(y, x) \right\} \\
&= \sup_{y \in \mathbb{R}^d} \left\{ g(x) - g(y) - \frac{1}{\lambda} D_\psi(y, x) \right\} \\
&\leqslant \sup_{y \in \mathbb{R}^d} \left\{ \|g\|_{\mathrm{Lip}} \cdot \|x - y\| - \frac{1}{\lambda} D_\psi(y, x) \right\} \\
&\leqslant \sup_{y \in \mathbb{R}^d} \left\{ \|g\|_{\mathrm{Lip}} \cdot \|x - y\| - \frac{\rho}{2\lambda} \|y - x\|^2 \right\} \qquad \text{by (A.1)} \\
&\leqslant \frac{\lambda}{2\rho} \|g\|^2_{\mathrm{Lip}},
\end{aligned} \tag{A.5}$$

since the maximum of $u \mapsto au - bu^2$ for $a \in [0, +\infty[$ and $b \in]0, +\infty[$ is $a^2/(4b)$. Likewise, we also have the same bound for the right Bregman–Moreau envelope

$$(\forall x \in \mathbb{R}^d) \quad g(x) - \overrightarrow{\mathrm{env}}^{\psi}_{\lambda,g}(x) \leqslant \frac{\lambda}{2\rho} \|g\|^2_{\mathrm{Lip}}.$$

(b) Recall that π has a density with respect to the Lebesgue measure and $U^{\psi}_{\lambda}(x) \leqslant U(x)$ for all $x \in \mathbb{R}^d$. Then we have

$$\int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \,\mathrm{d}x \leqslant \int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(x)} \,\mathrm{d}x. \tag{A.6}$$

This implies that, for all $x \in \mathbb{R}^d$,

$$\begin{aligned}
\pi(x) \leqslant \frac{\pi(x) \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \,\mathrm{d}y}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(y)} \,\mathrm{d}y} &= \frac{\mathrm{e}^{-U(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(y)} \,\mathrm{d}y} \\
&= \frac{\mathrm{e}^{-U^{\psi}_{\lambda}(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(y)} \,\mathrm{d}y} \cdot \mathrm{e}^{-U(x) + U^{\psi}_{\lambda}(x)} = \pi^{\psi}_{\lambda} \cdot \mathrm{e}^{\mathrm{env}^{\psi}_{\lambda,g}(x) - g(x)} \leqslant \pi^{\psi}_{\lambda}(x),
\end{aligned} \tag{A.7}$$

since $\mathrm{env}^{\psi}_{\lambda,g}(x) \leqslant g(x)$ for all $x \in \mathbb{R}^d$. Then for any $\lambda > 0$, we have

$$\begin{aligned}
\|\pi^{\psi}_{\lambda} - \pi\|_{\mathrm{TV}} &= \sup_{\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)} \left| \int_{\mathsf{A}} \pi^{\psi}_{\lambda}(x) - \pi(x)\,\mathrm{d}x \right| \leqslant \sup_{\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)} \int_{\mathsf{A}} \left| \pi^{\psi}_{\lambda}(x) - \pi(x) \right| \mathrm{d}x \\
&\leqslant \int_{\mathbb{R}^d} \left| \pi^{\psi}_{\lambda}(x) - \pi(x) \right| \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(\pi^{\psi}_{\lambda}(x) - \pi(x) \right)^{+} + \left(\pi^{\psi}_{\lambda}(x) - \pi(x) \right)^{-} \mathrm{d}x \\
&= 2 \int_{\mathbb{R}^d} \left(\pi^{\psi}_{\lambda}(x) - \pi(x) \right)^{+} \mathrm{d}x \\
&= 2 \int_{\mathbb{R}^d} \pi^{\psi}_{\lambda}(x) - \pi(x)\,\mathrm{d}x && \text{by (A.7)} \\
&\leqslant 2 \int_{\mathbb{R}^d} \pi^{\psi}_{\lambda}(x) - \pi(x) \frac{\int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\,\mathrm{d}y}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(y)}\,\mathrm{d}y}\,\mathrm{d}x && \text{by (A.6)} \\
&= 2 \left[\frac{1}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(x)}\,\mathrm{d}x} \int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(x)} - \mathrm{e}^{-U(x)}\,\mathrm{d}x \right] \\
&= 2 \int_{\mathbb{R}^d} \pi^{\psi}_{\lambda}(x) \left(1 - \mathrm{e}^{\mathrm{env}^{\psi}_{\lambda,g}(x) - g(x)} \right) \mathrm{d}x && \text{(A.8)} \\
&= 2 \left(1 - \frac{\int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\,\mathrm{d}x}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(x)}\,\mathrm{d}x} \right) \\
&\to 0,
\end{aligned}$$

as $\lambda \downarrow 0$, since, using Proposition 3.7(c) and the monotone convergence theorem, we have

$$\lim_{\lambda\to 0} U^{\psi}_{\lambda}(x) = U(x) \quad \Rightarrow \quad \lim_{\lambda\to 0} \int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(x)}\,\mathrm{d}x = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\,\mathrm{d}x.$$

(c) Since $\mathrm{env}^{\psi}_{\lambda,g}(x) \leqslant g(x)$ for all $x \in \mathbb{R}^d$ and $1 - \mathrm{e}^{-u} \leqslant u$ for all $u \in [0, +\infty[$, by (A.8), if Assumption 3.2(ii‡) holds, then

$$\|\pi^{\psi}_{\lambda} - \pi\|_{\mathrm{TV}} \leqslant 2 \int_{\mathbb{R}^d} \pi^{\psi}_{\lambda}(x) \left(g(x) - \mathrm{env}^{\psi}_{\lambda,g}(x) \right) \mathrm{d}x \leqslant \frac{\lambda}{\rho} \|g\|^2_{\mathrm{Lip}},$$

where the last inequality follows from (A.5).

Now we let $C_U \coloneqq \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\,\mathrm{d}x$. For the second part, we will make use of the inequalities Equation (A.6) and

$$\text{(A.5)} \quad \Rightarrow \quad (\forall x \in \mathbb{R}^d) \quad -U^{\psi}_{\lambda}(x) \leqslant -U(x) + \frac{\lambda}{2\rho} \|g\|^2_{\mathrm{Lip}}, \tag{A.9}$$

which implies

$$\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(x)}\,\mathrm{d}x \leqslant \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \cdot \mathrm{e}^{\lambda\|g\|^2_{\mathrm{Lip}}/(2\rho)}\,\mathrm{d}x = C_U \mathrm{e}^{\lambda\|g\|^2_{\mathrm{Lip}}/(2\rho)}. \tag{A.10}$$

Suppose that $h \geqslant 0$. Then (A.10) and $U^{\psi}_{\lambda}(x) \leqslant U(x)$ for all $x \in \mathbb{R}^d$ imply

$$\begin{aligned}
\mathbb{E}_{\pi^{\psi}_{\lambda}} h &= \int_{\mathbb{R}^d} h(x) \frac{\mathrm{e}^{-U^{\psi}_{\lambda}(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\psi}_{\lambda}(y)}\,\mathrm{d}y}\,\mathrm{d}x \\
&\geqslant C_U^{-1} \mathrm{e}^{-\lambda\|g\|^2_{\mathrm{Lip}}/(2\rho)} \int_{\mathbb{R}^d} h(x) \mathrm{e}^{-U^{\psi}_{\lambda}(x)}\,\mathrm{d}x \\
&\geqslant C_U^{-1} \mathrm{e}^{-\lambda\|g\|^2_{\mathrm{Lip}}/(2\rho)} \int_{\mathbb{R}^d} h(x) \mathrm{e}^{-U(x)}\,\mathrm{d}x \\
&= \mathrm{e}^{-\lambda\|g\|^2_{\mathrm{Lip}}/(2\rho)} \int_{\mathbb{R}^d} h(x)\pi(x)\,\mathrm{d}x \\
&= \mathrm{e}^{-\lambda\|g\|^2_{\mathrm{Lip}}/(2\rho)} \mathbb{E}_{\pi} h.
\end{aligned} \tag{A.11}$$

On the other hand, (A.5) and (A.9) imply

$$\mathbb{E}_{\pi_\lambda^\psi} h \leqslant C_U^{-1} \int_{\mathbb{R}^d} h(x) \mathrm{e}^{-U_\lambda^\psi(x)} \,\mathrm{d}x \leqslant C_U^{-1} \mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \int_{\mathbb{R}^d} h(x) \mathrm{e}^{-U(x)} \,\mathrm{d}x$$
$$= \mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \int_{\mathbb{R}^d} h(x)\pi(x) \,\mathrm{d}x = \mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \mathbb{E}_\pi h. \quad \text{(A.12)}$$

Combining (A.11) and (A.12) yields

$$\mathrm{e}^{-\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \mathbb{E}_\pi h \leqslant \mathbb{E}_{\pi_\lambda^\psi} h \leqslant \mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \mathbb{E}_\pi h. \quad \text{(A.13)}$$

Then, applying (A.13) gives

$$\begin{aligned}
-\Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1\Big)\mathbb{E}_\pi h &= -\max\Big\{\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1, 1 - \mathrm{e}^{-\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)}\Big\}\mathbb{E}_\pi h \\
&= \min\Big\{1 - \mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)}, \mathrm{e}^{-\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1\Big\}\mathbb{E}_\pi h \\
&\leqslant \Big(\mathrm{e}^{-\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1\Big)\mathbb{E}_\pi h \\
&\leqslant \mathbb{E}_{\pi_\lambda^\psi} h - \mathbb{E}_\pi h \\
&\leqslant \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1\Big)\mathbb{E}_\pi h \\
&\leqslant \max\Big\{\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1, 1 - \mathrm{e}^{-\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)}\Big\}\mathbb{E}_\pi h \\
&= \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} - 1\Big)\mathbb{E}_\pi h,
\end{aligned}$$

which implies that, for any $h \geqslant 0$,

$$\Big|\mathbb{E}_{\pi_\lambda^\psi} h - \mathbb{E}_\pi h\Big| \leqslant \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/\rho} - 1\Big)\mathbb{E}_\pi h. \quad \text{(A.14)}$$

Now, for any general integrable function h, we can write $h = h^+ - h^-$, where $h^+ \geqslant 0$ and $h^- \geqslant 0$. We also have $|h| = h^+ + h^-$. Consequently, we have

$$\begin{aligned}
\Big|\mathbb{E}_{\pi_\lambda^\psi} h - \mathbb{E}_\pi h\Big| &= \Big|\Big(\mathbb{E}_{\pi_\lambda^\psi} h^+ - \mathbb{E}_\pi h^+\Big) - \Big(\mathbb{E}_{\pi_\lambda^\psi} h^- - \mathbb{E}_\pi h^-\Big)\Big| \\
&\leqslant \Big|\mathbb{E}_{\pi_\lambda^\psi} h^+ - \mathbb{E}_\pi h^+\Big| + \Big|\mathbb{E}_{\pi_\lambda^\psi} h^- - \mathbb{E}_\pi h^-\Big| \\
&= \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/\rho} - 1\Big)\mathbb{E}_\pi h^+ + \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/\rho} - 1\Big)\mathbb{E}_\pi h^- \qquad \text{by (A.14)} \\
&= \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/\rho} - 1\Big)\mathbb{E}_\pi |h|. \quad \text{(A.15)}
\end{aligned}$$

If we switch the role of π_λ^ψ and π in (A.13), i.e.,

$$\mathrm{e}^{-\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \mathbb{E}_{\pi_\lambda^\psi} h \leqslant \mathbb{E}_\pi h \leqslant \mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/(2\rho)} \mathbb{E}_{\pi_\lambda^\psi} h,$$

then we get the following inequality similar to (A.15):

$$\Big|\mathbb{E}_{\pi_\lambda^\psi} h - \mathbb{E}_\pi h\Big| \leqslant \Big(\mathrm{e}^{\lambda \|g\|_{\mathrm{Lip}}^2/\rho} - 1\Big)\mathbb{E}_{\pi_\lambda^\psi} |h|. \quad \text{(A.16)}$$

Combining (A.15) and (A.16) yields the desired result.

□

A.2. Proof of Proposition 3.7

(a) Recall that g is lower bounded. Then by Bauschke et al. (2018, Fact 2.6), $g(\cdot) + \frac{1}{\lambda} D_\psi(\cdot, y)$ and $g(\cdot) + \frac{1}{\lambda} D_\psi(y, \cdot)$ are both coercive for all $y \in \mathcal{Y}$. Then the gradient formulas of the Bregman–Moreau envelopes follow from Bauschke et al. (2018, Proposition 2.19), which in turn follows from Bauschke et al. (2018, Remark 2.14) and Bauschke et al. (2006, Proposition 3.12).

(b) The Lipschitz continuity of the gradient of the left Bregman–Moreau envelope follows from Soueycatt et al. (2020, Theorem 3.5), whereas the Lipschitz continuity of the gradient of the right Bregman–Moreau envelope holds because, assuming that $\nabla\psi$ is Lipschitz, $y \mapsto \nabla_y D_\psi(y, x) = \nabla\psi(y) - \nabla\psi(x)$ is Lipschitz and we use the fact that the composition of Lipschitz maps is also Lipschitz. We also remark that if we further assume that ψ is very strictly convex, then $\nabla\psi$ is Lipschitz (Bauschke & Lewis, 2000, Proposition 2.10; Laude et al., 2020, Lemma 2.3(iii)).

(c) The asymptotic behavior of the Bregman–Moreau envelopes follow from Bauschke et al. (2018, Theorem 3.3).

□

A.3. Proof of Theorem 4.2

Note that (24) follows from Theorem 3.1 of Li et al. (2022), i.e.,

$$\mathsf{W}_{2,\varphi}(\mu_k, \pi_\lambda^\psi) \leqslant \sqrt{2}\mathrm{e}^{-(\alpha - 2M_\varphi)\gamma k} \mathsf{W}_{2,\varphi}(\mu_0, \pi_\lambda^\psi) + C\sqrt{2\gamma},$$

where $\gamma \in]0, \gamma_{\max}]$, with $\gamma_{\max} = \mathcal{O}\Big(\frac{(\alpha - 2M_\varphi)^2}{(\beta^2(1+8M_\varphi)^2)}\Big)$ and $C = \mathcal{O}\Big(\frac{\beta(1+8M_\varphi)\sqrt{d}}{(\alpha - 2M_\varphi)}\Big)$.

By Gibbs & Su (2002, Theorem 4) with $d(x, y) = \|\nabla\varphi(x) - \nabla\varphi(y)\|^2$, we have the following inequality between the Wasserstein distance and the total variation distance:

$$\mathsf{W}_{2,\varphi}(\mu, \nu) \leqslant \eta \|\mu - \nu\|_{\mathrm{TV}}, \tag{A.17}$$

where $\eta \coloneqq \sup_{u\sim\mu, v\sim\nu} \|\nabla\varphi(u) - \nabla\varphi(v)\|^2$.

Invoking the triangle inequality, (24) and (A.17), we have

$$\begin{aligned}\mathsf{W}_{2,\varphi}(\mu_k, \pi) &\leqslant \mathsf{W}_{2,\varphi}(\mu_k, \pi_\lambda^\psi) + \mathsf{W}_{2,\varphi}(\pi_\lambda^\psi, \pi) \\ &\leqslant \sqrt{2}\mathrm{e}^{-(\alpha - 2M_\varphi)\gamma k} \mathsf{W}_{2,\varphi}(\mu_0, \pi_\lambda^\psi) + C\sqrt{2\gamma} + \eta\|\pi_\lambda^\psi - \pi\|_{\mathrm{TV}}.\end{aligned} \tag{A.18}$$

Recall from Proposition 3.6(c) that if Assumption 3.2(ii‡) holds and ψ is ρ-strongly convex, then

$$\|\pi_\lambda^\psi - \pi\|_{\mathrm{TV}} \leqslant \frac{\lambda}{\rho}\|g\|_{\mathrm{Lip}}^2.$$

Hence, we obtain

$$\mathsf{W}_{2,\varphi}(\pi_\lambda^\psi, \pi) \leqslant \eta\|\pi_\lambda^\psi - \pi\|_{\mathrm{TV}} \leqslant \frac{\eta\lambda}{\rho}\|g\|_{\mathrm{Lip}}^2. \tag{A.19}$$

Then (A.18) becomes

$$\mathsf{W}_{2,\varphi}(\mu_k, \pi) \leqslant \sqrt{2}\mathrm{e}^{-(\alpha - 2M_\varphi)\gamma k} \mathsf{W}_{2,\varphi}(\mu_0, \pi_\lambda^\psi) + C\sqrt{2\gamma} + \frac{\eta\lambda}{\rho}\|g\|_{\mathrm{Lip}}^2. \tag{A.20}$$

On the other hand, applying the triangle inequality again and (A.19), we have

$$\mathsf{W}_{2,\varphi}(\mu_0, \pi_\lambda^\psi) \leqslant \mathsf{W}_{2,\varphi}(\mu_0, \pi) + \mathsf{W}_{2,\varphi}(\pi_\lambda^\psi, \pi) \leqslant \mathsf{W}_{2,\varphi}(\mu_0, \pi) + \frac{\eta\lambda}{\rho}\|g\|_{\mathrm{Lip}}^2.$$

Plugging into (A.20) yields the desired result (25)

$$\mathsf{W}_{2,\varphi}(\mu_k, \pi) \leqslant \sqrt{2}\mathrm{e}^{-(\alpha - 2M_\varphi)\gamma k} \mathsf{W}_{2,\varphi}(\mu_0, \pi) + C\sqrt{2\gamma} + \Big(1 + \sqrt{2}\mathrm{e}^{-(\alpha - 2M_\varphi)\gamma k}\Big)\frac{\eta\lambda}{\rho}\|g\|_{\mathrm{Lip}}^2.$$

A.4. Proof of Corollary 4.3

(26) simply follows from (25) with certain algebraic manipulations.

B. Details of Numerical Experiments

More notation. For any $\boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, $\mathrm{Diag}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ is the diagonal matrix whose diagonal entries are $x_1, \dots, x_d$. We also write $[\![d]\!] \coloneqq \{1, \dots, d\}$.

With the choice of

$$\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \sum_{i=1}^{d} \left[\theta_i \operatorname{arsinh}(\theta_i/\beta_i) - \sqrt{\theta_i^2 + \beta_i^2}\right],$$

and $\psi(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_{\boldsymbol{M}}^2$ with $\boldsymbol{M} \in \mathbb{S}_{++}^d$, simple calculation yields

$$\begin{aligned}
\nabla\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= (\operatorname{arsinh}(\theta_i/\beta_i))_{1\leqslant i\leqslant d},\\
\nabla^2\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= \mathrm{Diag}\left(\left((\theta_i^2+\beta_i^2)^{-1/2}\right)_{1\leqslant i\leqslant d}\right),\\
\varphi_{\boldsymbol{\beta}}^*(\boldsymbol{\theta}) &= \sum_{i=1}^{d} \beta_i \cosh(\theta_i),\\
\nabla\varphi_{\boldsymbol{\beta}}^*(\boldsymbol{\theta}) &= (\beta_i \sinh(\theta_i))_{1\leqslant i\leqslant d},\\
\nabla^2\varphi_{\boldsymbol{\beta}}^*(\boldsymbol{\theta}) &= \mathrm{Diag}((\beta_i \cosh(\theta_i))_{1\leqslant i\leqslant d}),
\end{aligned}$$

and

$$\begin{aligned}
\nabla\psi(\boldsymbol{\theta}) &= \boldsymbol{M}\boldsymbol{\theta},\\
\nabla^2\psi(\boldsymbol{\theta}) &= \boldsymbol{M}.
\end{aligned}$$

Note that for $f = 0$ and $g(\boldsymbol{\theta}) = \sum_{i=1}^d \alpha_i|\theta_i|$, $\mathcal{F} = \mathcal{G} = \mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. It is straightforward to see that Assumptions 3.1 and 3.2(i), (ii†), (ii‡) are satisfied. For Assumption 3.3, we check the modified self-concordance condition since the other assumptions are obvious. Since $\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is separable in a sense that it is in the form $\sum_{i=1}^d \phi_{\beta_i}(\theta_i)$, where $\phi_\beta(\theta) \coloneqq \theta \operatorname{arsinh}(\theta/\beta) - \sqrt{\theta^2+\beta^2}$ with $\beta > 0$, it suffices to show that ϕ_β is a modified self-concordant function. As noted in Zhang et al. (2020), it suffices to check that $[(\phi_\beta^*)'']^{-1/2}$ is Lipschitz.

Since $[(\phi_\beta^*(\theta))'']^{-1/2} = \beta^{-1/2}\sqrt{\operatorname{sech}(\theta)}$, we have

$$\left[\frac{1}{\sqrt{(\phi_\beta^*(\theta))''}}\right]' = -\frac{1}{2\sqrt{\beta}}\sinh(\theta)\operatorname{sech}^{3/2}(\theta) \quad \Rightarrow \quad \left|\left[\frac{1}{\sqrt{(\phi_\beta^*(\theta))''}}\right]'\right| \leqslant \frac{1}{\sqrt{2}\cdot 3^{3/4}}.$$

Hence, $[(\phi_\beta^*)'']^{-1/2}$ is Lipschitz.

It is also obvious to see that ψ satisfies Assumption 3.4.

For $\boldsymbol{M} \in \mathbb{S}_{++}^d$, the Bregman divergence associated to $\frac{1}{2}\|\cdot\|_{\boldsymbol{M}}^2$ is given by $D_\psi(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_{\boldsymbol{M}}^2$, which is indeed a distance since it is symmetric in its arguments.

The choice of ψ implies its associated Bregman divergence D_ψ satisfies all of Assumption 3.5. According to Bauschke & Borwein (2001, Corollary 7.2 and Example 7.3), since $[\nabla^2\psi(\boldsymbol{\theta})]^{-1} = \boldsymbol{M}^{-1}$ is constant for all $\boldsymbol{\theta} \in \mathbb{R}^d$, and thus trivially matrix-concave. Hence D_ψ is jointly convex. In addition, the gradient and Hessian of D_ψ in the second argument are

$$\begin{aligned}
\nabla_{\boldsymbol{\vartheta}} D_\psi(\boldsymbol{\theta}, \boldsymbol{\vartheta}) &= \boldsymbol{M}(\boldsymbol{\vartheta} - \boldsymbol{\theta}),\\
\nabla_{\boldsymbol{\vartheta}}^2 D_\psi(\boldsymbol{\theta}, \boldsymbol{\vartheta}) &= \boldsymbol{M}.
\end{aligned}$$

Since $\boldsymbol{M} \in \mathbb{S}_{++}^d$, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, $D_\psi(\boldsymbol{\theta}, \cdot)$ is strictly convex on $\mathbb{R}^d$. Obviously, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, $D_\psi(\boldsymbol{\theta}, \cdot)$ is also continuous on $\mathbb{R}^d$ and coercive.

To check Assumption 3.8, we first compute the expressions of $\mathrm{P}^{\frac{1}{2}m(\cdot)^2}_{\lambda,\alpha|\cdot|}$. Then the expressions of $\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,g}(\boldsymbol{\theta})$ and $\overrightarrow{\mathrm{P}}^{\psi}_{\lambda,g}(\boldsymbol{\theta})$ are given by

$$\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,g}(\boldsymbol{\theta}) = \left(\overleftarrow{\mathrm{P}}^{\frac{1}{2}m(\cdot)^2}_{\lambda,\alpha_i|\cdot|}(\theta_i)\right)_{1\leqslant i\leqslant d} \quad \text{and} \quad \overrightarrow{\mathrm{P}}^{\psi}_{\lambda,g}(\boldsymbol{\theta}) = \left(\overrightarrow{\mathrm{P}}^{\frac{1}{2}m(\cdot)^2}_{\lambda,\alpha_i|\cdot|}(\theta_i)\right)_{1\leqslant i\leqslant d},$$

attributed to the separable structures of g and D_ψ. Note that $\overleftarrow{\mathrm{P}}^{\psi}_{\lambda,g} = \overrightarrow{\mathrm{P}}^{\psi}_{\lambda,g}$ since D_ψ is symmetric in its arguments.

Simple manipulation yields

$$\begin{aligned}\mathrm{P}^{\frac{1}{2}m(\cdot)^2}_{\lambda,\alpha|\cdot|}(\theta) &= \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\alpha|\vartheta| + \frac{m}{2\lambda}(\theta-\vartheta)^2\right\} = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\frac{\lambda\alpha}{m}|\vartheta| + \frac{1}{2}(\theta-\vartheta)^2\right\} \\ &= \mathrm{prox}_{\lambda\alpha|\cdot|/m}(\theta),\end{aligned}$$

where $\mathrm{prox}_{\mu|\cdot|}(\theta) = \mathrm{sign}(\theta)\max\{|\theta|-\mu, 0\}$ is the soft-thresholding operator, for $\theta\in\mathbb{R}$ and $\mu > 0$. Consequently, we have

$$\mathrm{P}^{\psi}_{\lambda,g}(\boldsymbol{\theta}) = (\mathrm{sign}(\theta_i)\max\{|\theta_i| - \lambda\alpha_i/m_i, 0\})_{1\leqslant i\leqslant d}.$$

It remains to check Assumption 3.8. It appears that $U^{\psi}_{\lambda} = \mathrm{env}^{\psi}_{\lambda,g}$ in this case is Legendre strongly convex with $\alpha = 0$ (i.e., convex but not strongly convex), which does not satisfy the required assumption that $M_\varphi < \alpha/2$. However, for practical purpose, this choice of ψ works well. We will give

We then check that $\mathrm{env}^{\psi}_{\lambda,g}$ is β_g-smooth relative to φ. We check this via the equivalent second-order characterization: $\beta_g\nabla^2\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \nabla^2\mathrm{env}^{\psi}_{\lambda,g}(\boldsymbol{\theta}) \succeq 0$ for all $\boldsymbol{\theta}\in\mathbb{R}^d$.

Let $i\in[\![d]\!]$. Then we have

$$\left[\nabla^2\mathrm{env}^{\psi}_{\lambda,g}(\boldsymbol{\theta})\right]_{i,i} = \begin{cases} m_i/\lambda & \text{if } \theta_i\in[-\lambda\alpha_i/m_i, \lambda\alpha_i/m_i], \\ 0 & \text{otherwise.}\end{cases}$$

Since

$$\nabla^2\varphi_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathrm{Diag}\left(\left(\frac{1}{\sqrt{\theta_i^2+\beta_i^2}}\right)\right)_{1\leqslant i\leqslant d},$$

we can choose

$$\beta_g = \sup_{i\in[\![d]\!]}\sup_{\theta_i\in[-\lambda\alpha_i/m_i, \lambda\alpha_i/m_i]} \left\{\frac{m_i\sqrt{\theta_i^2+\beta_i^2}}{\lambda}\right\}.$$

Given the choice $m_i = \alpha_i/2$ for all $i\in[\![d]\!]$, we then have

$$\beta_g = \sup_{i\in[\![d]\!]}\sup_{\theta_i\in[-2\lambda,2\lambda]} \left\{\frac{\alpha_i\sqrt{\theta_i^2+\beta_i^2}}{2\lambda}\right\} = \sup_{i\in[\![d]\!]}\sup_{\theta_i\in[0,2\lambda]} \left\{\frac{\alpha_i\sqrt{\theta_i^2+\beta_i^2}}{2\lambda}\right\} = \sup_{i\in[\![d]\!]} \frac{\alpha_i\sqrt{4\lambda^2+\beta_i^2}}{2\lambda} < +\infty,$$

which implies that $\overleftarrow{\mathrm{env}}^{\psi}_{\lambda,g}$ is β_g-smooth relative to φ.

B.1. Different Bregman–Moreau Envelopes

To find Bregman–Moreau envelopes which also satisfy Assumption 3.8, we let $\psi = \psi_{\boldsymbol{\sigma}}$ be also the hyperbolic entropy parameterized by $\boldsymbol{\sigma}$. By slight abuse of notation, we also write $\psi_{\boldsymbol{\sigma}}(\boldsymbol{\theta}) = (\psi_{\sigma_i}(\theta_i))_{1\leqslant i\leqslant d}$.

We have the following expression of the associatedl left Bregman proximity operator.

Proposition B.1. *The left Bregman proximity operator of $\alpha|\cdot|$ associated to the Legendre function ψ_σ for $\alpha > 0$ is*

$$\overleftarrow{\mathrm{P}}^{\psi_\sigma}_{\lambda,\alpha|\cdot|}(\theta) = \begin{cases} \sigma\sinh(\mathrm{arsinh}(\theta/\sigma) - \alpha\lambda) & \textit{if } \theta > \sigma\sinh(\alpha\lambda), \\ \sigma\sinh(\mathrm{arsinh}(\theta/\sigma) + \alpha\lambda) & \textit{if } \theta < \sigma\sinh(-\alpha\lambda), \\ \sqrt{\theta^2+\beta^2} & \textit{otherwise.}\end{cases}$$

Proof of Proposition B.1. According to Definition 2.5,

$$\overleftarrow{\mathrm{P}}^{\exp}_{\lambda,\alpha|\cdot|}(\theta) = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\lambda\alpha|\vartheta| + \vartheta(\operatorname{arsinh}(\vartheta/\sigma) - \operatorname{arsinh}(\theta/\sigma)) - \sqrt{\vartheta^2+\sigma^2} + \sqrt{\theta^2+\sigma^2}\right\}.$$

First-order conditions give

$$\begin{cases} \alpha\lambda + \operatorname{arsinh}(\vartheta/\sigma) - \operatorname{arsinh}(\theta/\sigma) = 0 & \text{if } \vartheta > 0, \\ -\alpha\lambda + \operatorname{arsinh}(\vartheta/\sigma) - \operatorname{arsinh}(\theta/\sigma) = 0 & \text{if } \vartheta < 0, \end{cases}$$

which implies

$$\vartheta^\star = \begin{cases} \sigma\sinh(\operatorname{arsinh}(\theta/\sigma) - \alpha\lambda) & \text{if } \vartheta^\star > 0, \\ \sigma\sinh(\operatorname{arsinh}(\theta/\sigma) + \alpha\lambda) & \text{if } \vartheta^\star < 0 \end{cases} = \begin{cases} \sigma\sinh(\operatorname{arsinh}(\theta/\sigma) - \alpha\lambda) & \text{if } \theta > \sigma\sinh(\alpha\lambda), \\ \sigma\sinh(\operatorname{arsinh}(\theta/\sigma) + \alpha\lambda) & \text{if } \theta < \sigma\sinh(-\alpha\lambda). \end{cases} \tag{B.1}$$

On the other hand, if $\vartheta = 0$, then

$$\begin{aligned} &\operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\lambda\alpha|\vartheta| + \vartheta(\operatorname{arsinh}(\vartheta/\sigma) - \operatorname{arsinh}(\theta/\sigma)) - \sqrt{\vartheta^2+\sigma^2} + \sqrt{\theta^2+\sigma^2}\right\} \\ &\qquad = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{-\sqrt{\sigma^2} + \sqrt{\theta^2+\sigma^2}\right\} = \sqrt{\theta^2+\sigma^2} - \sigma, \end{aligned} \tag{B.2}$$

which corresponds to the range $[\sigma\sinh(-\alpha\lambda), \sigma\sinh(\alpha\lambda)]$ for θ. Combining (B.1) and (B.2) yields the desired result. □

The closed-form expression of the right Bregman proximity operator is much more complicated and is not given.

We show that $\overleftarrow{\mathrm{env}}^{\psi_{\boldsymbol{\sigma}}}_{\lambda,g} - \alpha\varphi_{\boldsymbol{\beta}}$ is convex for some $\lambda \in]0, +\infty[$, $\boldsymbol{\beta} \in \mathbb{R}^d_+$, $\boldsymbol{\sigma} \in \mathbb{R}^d_+$ and $\alpha > 2M_{\varphi_{\boldsymbol{\beta}}}$. Also, recall that

$$\left|\left[\frac{1}{\sqrt{(\phi^*_\beta(\theta))''}}\right]'\right| \leqslant \frac{1}{\sqrt{2}\cdot 3^{3/4}} \quad \Rightarrow \quad M_{\varphi_{\boldsymbol{\beta}}} = \left(2\cdot 3^{3/2}\cdot \min_{i\in[\![d]\!]} \beta_i\right)^{-1}.$$

In particular, if we choose $\lambda = 10^{-5}$, $\boldsymbol{\beta} = (2\sqrt{d-i+1})^\top_{1\leqslant i\leqslant d}$, $\boldsymbol{\sigma} = (d, d-1, \ldots, 1)^\top$ and $\alpha = 2M_{\varphi_{\boldsymbol{\beta}}} + 10^{-1}$, then we plot the following:

Figure B.1 shows that the above choices give a convex $\overleftarrow{\mathrm{env}}^{\psi_{\boldsymbol{\sigma}}}_{\lambda,g} - \alpha\varphi_{\boldsymbol{\beta}}$.

Similarly, we also show graphically that $\beta_g\varphi_{\boldsymbol{\beta}} - \overleftarrow{\mathrm{env}}^{\psi_{\boldsymbol{\sigma}}}_{\lambda,g}$ is convex for some $\beta_g > 0$, e.g., $\beta_g = 2500$ (not tight).

Since all Assumptions 3.1 to 3.5 and 3.8 are satisfied, g is Lipschitz and ψ is strongly convex, the convergence results in the main text, i.e., Theorem 4.2 and Corollary 4.3, hold.

C. The Bregman–Moreau Mirrorless Mirror-Langevin Algorithm

In this section, we give the details of the Bregman–Moreau mirrorless mirror-Langevin algorithm (BMMMLA), whose results are mostly taken from Ahn & Chewi (2021).

C.1. Assumptions

We first state the assumptions required in Ahn & Chewi (2021). Instead of the modified self-concordance condition, the Legenedre function φ has to be M_φ*-self-concordant* (Nesterov, 2018, §5.1.3), i.e., for any $x \in \mathfrak{X}$, there exists $M_\varphi \geqslant 0$ such that $\left|\nabla^3\varphi(x)[u,u,u]\right| \leqslant 2M_\varphi\|u\|^3_{\nabla^2\varphi(x)}$ for all $u \in \mathbb{R}^d$. Furthermore, in addition to the α-relative convexity (to φ) and β-relative smoothness (to φ) assumption, U^ψ_λ also has to be L-Lipschitz relative to φ, which is defined as follows.

Definition C.1 (Relative Lipschitz continuity)**.** A function $f \in \mathscr{C}^1$ is L-Lipschitz relative to a very strictly convex (see Assumption 3.3(iii)) Legendre function φ if there exists $L > 0$ such that $\|\nabla f(x)\|_{[\nabla^2\varphi(x)]^{-1}} \leqslant L$ for all $x \in \operatorname{int}\operatorname{dom} f$.

It is worth noting that it is difficult to verify that Bregman–Moreau envelopes would satisfy such a relative Lipschitzness condition in general.

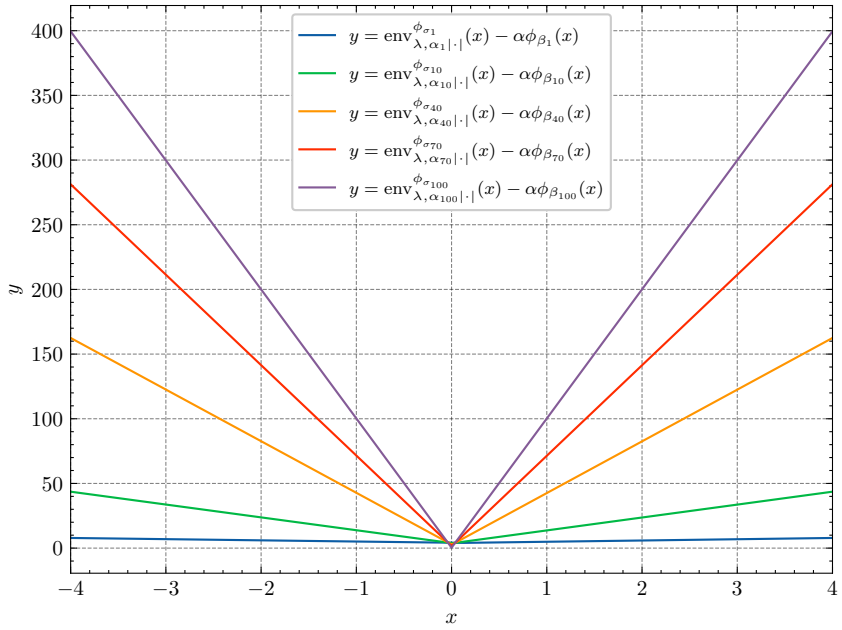


Figure B.1. Plots of $y = \mathrm{env}^{\psi_{\sigma_i}}_{\lambda,\alpha_i|\cdot|}(x) - \alpha\phi_{\beta_i}(x)$, for $i \in \{1, 10, 40, 70, 100\}$.

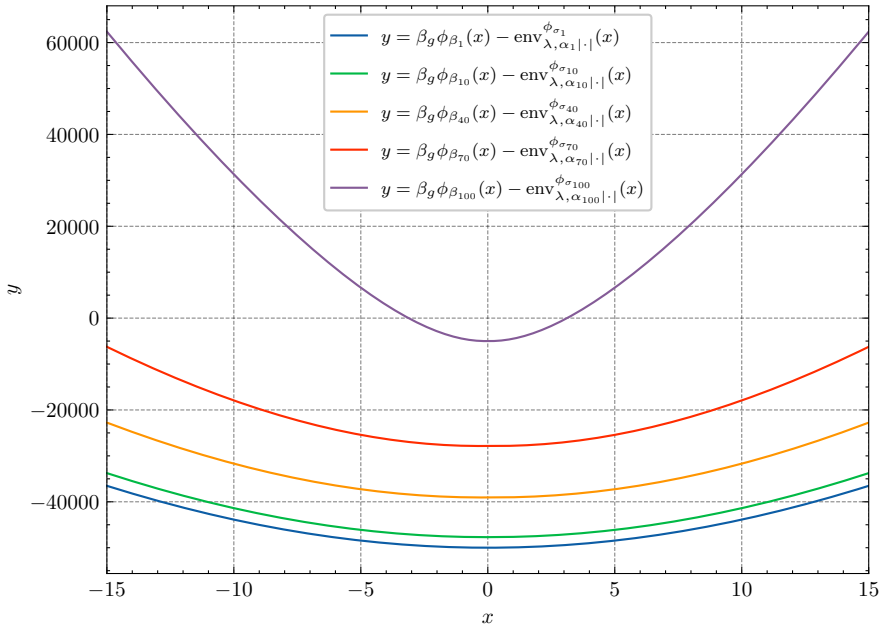


Figure B.2. Plots of $y = \beta_g\phi_{\beta_i}(x) - \mathrm{env}^{\psi_{\sigma_i}}_{\lambda,\alpha_i|\cdot|}(x)$, for $i \in \{1, 10, 40, 70, 100\}$.

Now we let $x_0 \in \mathcal{Y}$. Similar to MLA (7) in Ahn & Chewi (2021), the Bregman–Moreau mirrorless Mirror-Langevin algorithm (BMMMLA) iterates

$$x_{k+1/2} = \nabla\varphi^*\Big(\nabla\varphi(x_k) - \gamma\nabla U^{\psi}_{\lambda}(x_k)\Big),$$
$$x_{k+1} = \nabla\varphi^*(Y_\gamma),$$

where

$$\begin{cases} \mathrm{d}Y_t = \sqrt{2}\big[\nabla^2\varphi^*(Y_t)\big]^{-1/2}\,\mathrm{d}W_t \\ Y_0 = \nabla\varphi\big(x_{k+1/2}\big) = \nabla\varphi(x_k) - \gamma\nabla U_\lambda^\psi(x_k). \end{cases} \tag{C.1}$$

C.2. Convergence Results

We suppose that the assumptions in Appendix C.1 hold. We define the mixture distribution $\overline{\mu}_K \coloneqq \frac{1}{K}\sum_{k=1}^K \mu_k$, and let $\beta' \coloneqq \beta + 2M_\varphi L$. Then we have the following convergence results.

Theorem C.2 (Convex)**.** *Assume $\alpha = 0$ and $\beta' > 0$. Let $X_k \sim \mu_k$ be generated by (C.1) with step size $\gamma = \min\{\varepsilon/(2\beta' d), 1/\beta'\}$. Then for all $\varepsilon > 0$, there exists $\lambda > 0$ such that $D_{\mathrm{KL}}(\overline{\mu}_K \,\|\, \pi_\lambda^\psi) \leqslant \varepsilon$ for*

$$K \geqslant \frac{4d\beta'\mathsf{D}_\varphi(\pi, \mu_0)}{\varepsilon^2}\max\Big\{1, \frac{\varepsilon}{2d}\Big\}.$$

Theorem C.3 (Legendre strongly convex)**.** *Assume $\alpha > 0$ and $\beta' > 0$. Suppose that $X_0 \sim \mu_0$ satisfies $\mathsf{D}_\varphi(\pi, \mu_0) \leqslant \varepsilon$. Let $X_k \sim \mu_k$ be generated by (C.1) with step size $\gamma = \min\{\varepsilon/(2\beta' d), 1/\beta'\}$. Then for all $\varepsilon > 0$, there exists $\lambda > 0$ such that $D_{\mathrm{KL}}(\overline{\mu}_K \,\|\, \pi_\lambda^\psi) \leqslant \varepsilon$ for*

$$K \geqslant \frac{4\beta' d}{\varepsilon}\max\Big\{1, \frac{\varepsilon}{2d}\Big\}.$$

The results follow from Ahn & Chewi (2021, Theorems 1 and 2(b)). Similar bounds on the total variation distance follows from *Pinsker's inequality*: $\|P - Q\|_{\mathrm{TV}}^2 \leqslant \frac{1}{2}D_{\mathrm{KL}}(P \,\|\, Q)$, and also bounds on the total variation distance between $\overline{\mu}_K$ and the target distribution π instead of the surrogate distribution π_λ^ψ. □

Note also that convergence in the Bregman transport cost also holds (Ahn & Chewi, 2021, Theorem 2(a)), where the Bregman transport cost is defined as follows.

Definition C.4 (Bregman transport cost)**.** For two probability measures μ and ν on $\mathcal{B}(\mathbb{R}^d)$, the *Bregman transport cost* (Cordero-Erausquin, 2017) from μ to ν with respect to the Bregman divergence associated with a Legendre function φ is defined by

$$\mathsf{D}_\varphi(\mu, \nu) \coloneqq \inf_{\pi\sim\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} D_\varphi(x, y)\,\mathrm{d}\pi(x, y).$$

We also refer to Ahn & Chewi (2021, Theorem 2.1) for convergence results in terms of the Bregman transport cost (Cordero-Erausquin, 2017). By Pinsker's inequality, we can also obtain similar results in terms of the total variation distance.

Finally, we give the Bregman–Moreau mirrorless mirror-Langevin algorithm (C.1) with an Euler–Maruyama discretization for the second step.

Algorithm 1 The Bregman–Moreau Mirrorless Mirror-Langevin Algorithm (BMMMLA)

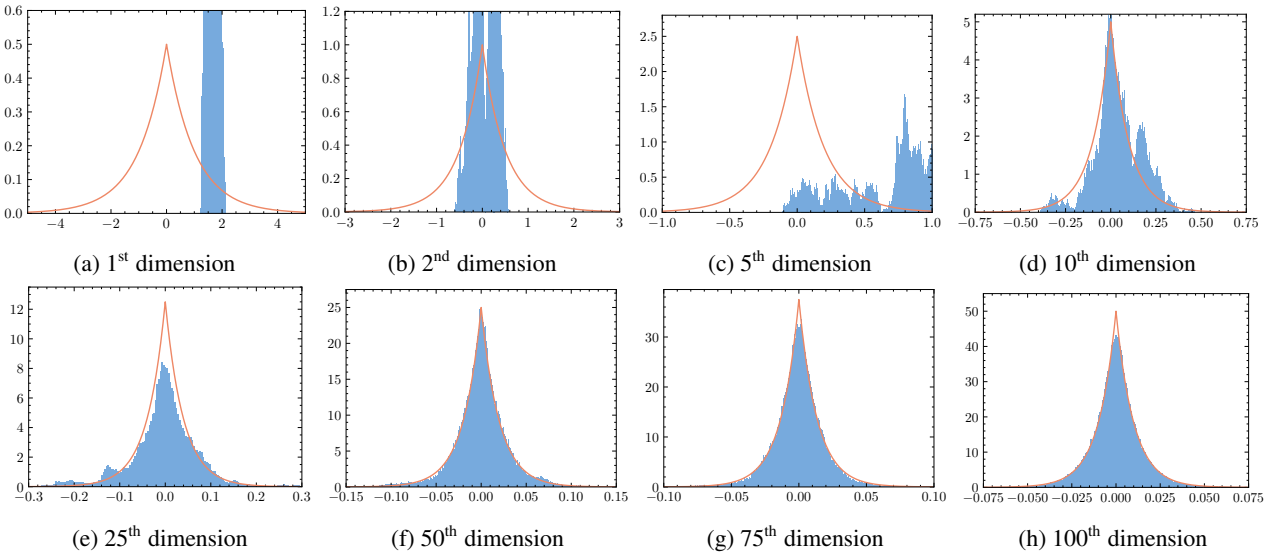
Initialize: Legendre functions φ and ψ, $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, step size $\gamma \in \,]0, +\infty[$, number of samples to be drawn $K \in \mathbb{N}^*$, number of inner steps of Euler–Maruyama discretization $N \in \mathbb{N}^*$.
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $\boldsymbol{\theta}_{k+1/2} = \nabla\varphi^*\Big(\nabla\varphi(\boldsymbol{\theta}_k) - \gamma\nabla U_\lambda^\psi(\boldsymbol{\theta}_k)\Big)$
  $\boldsymbol{y}_0 = \nabla\varphi\big(\boldsymbol{\theta}_{k+1/2}\big) = \nabla\varphi(\boldsymbol{\theta}_k) - \gamma\nabla U_\lambda^\psi(\boldsymbol{\theta}_k)$
  for $n = 0, 1, 2, \ldots, N$ **do**
    $\boldsymbol{\xi}_n \sim \mathsf{N}_d(\mathbf{0}_d, I_d)$
    $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \sqrt{2\gamma/N}\big[\nabla^2\varphi^*(\boldsymbol{y}_n)\big]^{-1/2}\boldsymbol{\xi}_n$
  end for
  $\boldsymbol{\theta}_{k+1} = \nabla\varphi^*(\boldsymbol{y}_{N+1})$
end for

We also give the experimental results of the Bregman–Moreau mirrorless mirror-Langevin algorithm in Appendix D.

D. Additional Numerical Experiments

D.1. Anisotropic Laplace Distribution

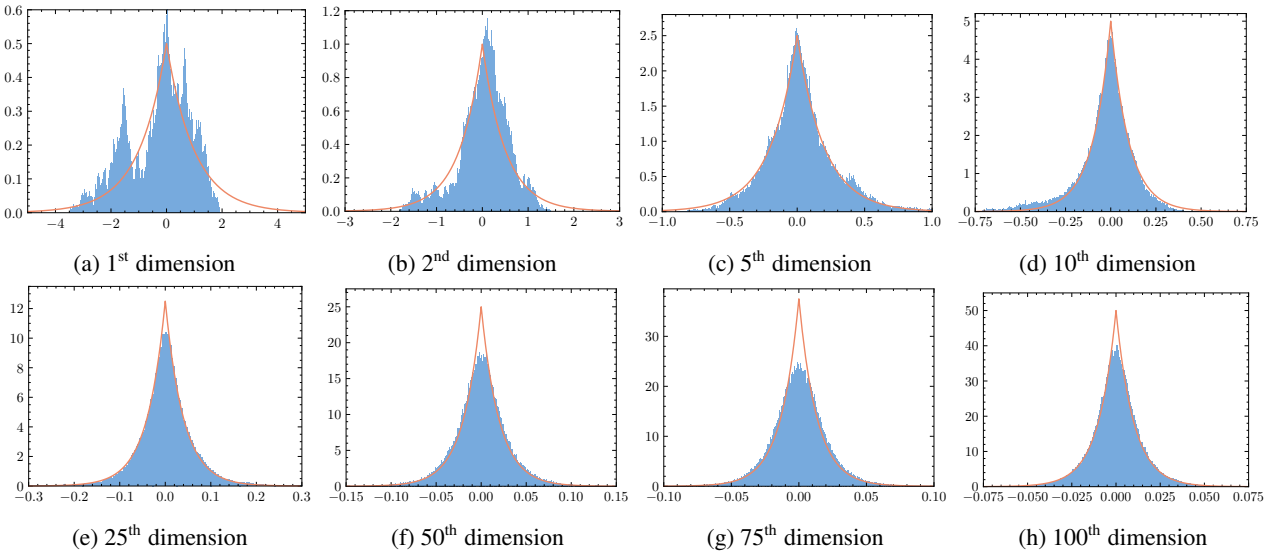
We first give more plots of different dimensions for the experiment in Section 5.



(a) 1st dimension (b) 2nd dimension (c) 5th dimension (d) 10th dimension

(e) 25th dimension (f) 50th dimension (g) 75th dimension (h) 100th dimension

Figure D.1. Histograms of samples (blue) from MYULA and the true densities (orange).



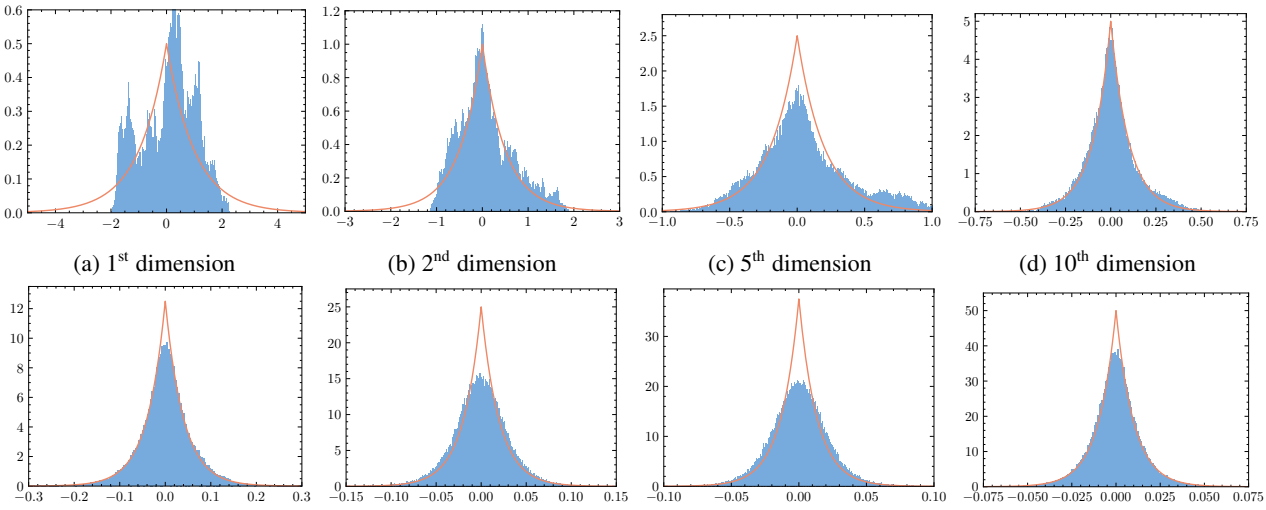
(a) 1st dimension (b) 2nd dimension (c) 5th dimension (d) 10th dimension

(e) 25th dimension (f) 50th dimension (g) 75th dimension (h) 100th dimension

Figure D.2. Histograms of samples (blue) from BMUMLA and the true densities (orange).

We also give the experimental results using a different (left) Bregman–Moreau envelope introduced in Appendix B.1, using the same step size $\gamma = 5 \times 10^{-6}$ in Figure D.3.

On the other hand, for practical purpose, we also perform the same set of experiments with another Bregman–Moreau envelope associated to the Legendre function $\psi(\boldsymbol{\theta}) = \sum_{i=1}^{d} \mathrm{e}^{\theta_i}$. This is chosen particularly because we can compute the corresponding closed form expressions of both of its associated left and right Bregman proximity operators. To do so, we compute the following left and right Bregman proximity operators associated to the exponential function.



(a) 1st dimension (b) 2nd dimension (c) 5th dimension (d) 10th dimension

(e) 25th dimension (f) 50th dimension (g) 75th dimension (h) 100th dimension

Figure D.3. Histograms of samples (blue) from LBMUMLA and the true densities (orange) with a different Bregman–Moreau envelope.

Proposition D.1. *The left Bregman proximity operator of* $\alpha|\cdot|$ *associated to the Legendre function* $\exp$ *for* $\alpha > 0$ *is*

$$\overleftarrow{\mathrm{P}}^{\exp}_{\lambda,\alpha|\cdot|}(\theta) = \begin{cases} \log(\mathrm{e}^{\theta} - \alpha\lambda) & \text{if } \theta > \log(1+\alpha\lambda), \\ \log(\mathrm{e}^{\theta} + \alpha\lambda) & \text{if } \theta < \log(1-\alpha\lambda), \\ 1 - \mathrm{e}^{\theta}(1+\theta) & \text{otherwise.} \end{cases}$$

The right Bregman proximity operator of $\alpha|\cdot|$ *associated to the Legendre function* $\exp$ *for* $\alpha > 0$ *is*

$$\overrightarrow{\mathrm{P}}^{\exp}_{\lambda,\alpha|\cdot|}(\theta) = \begin{cases} W(-\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \theta > \alpha\lambda, \\ W(\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \theta < -\alpha\lambda, \\ \mathrm{e}^{\theta} - (1+\theta) & \text{otherwise,} \end{cases}$$

where W *is the Lambert* W *function (Lambert, 1758; Corless et al., 1996), i.e., the inverse of* $\xi \mapsto \xi\mathrm{e}^{\xi}$ *on* $[0, +\infty[$.

Proof of Proposition D.1. According to Definition 2.5,

$$\overleftarrow{\mathrm{P}}^{\exp}_{\lambda,\alpha|\cdot|}(\theta) = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\lambda\alpha|\vartheta| + \mathrm{e}^{\vartheta} - \mathrm{e}^{\theta} - \mathrm{e}^{\theta}(\vartheta - \theta)\right\}.$$

First-order conditions give

$$\begin{cases} \alpha\lambda + \mathrm{e}^{\vartheta} - \mathrm{e}^{\theta} = 0 & \text{if } \vartheta > 0, \\ -\alpha\lambda + \mathrm{e}^{\vartheta} - \mathrm{e}^{\theta} = 0 & \text{if } \vartheta < 0, \end{cases}$$

which implies

$$\vartheta^{\star} = \begin{cases} \log(\mathrm{e}^{\theta} - \alpha\lambda) & \text{if } \vartheta^{\star} > 0, \\ \log(\mathrm{e}^{\theta} + \alpha\lambda) & \text{if } \vartheta^{\star} < 0 \end{cases} = \begin{cases} \log(\mathrm{e}^{\theta} - \alpha\lambda) & \text{if } \theta > \log(1+\alpha\lambda), \\ \log(\mathrm{e}^{\theta} + \alpha\lambda) & \text{if } \theta < \log(1-\alpha\lambda). \end{cases} \tag{D.1}$$

On the other hand, if $\vartheta = 0$, then

$$\operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\lambda\alpha|\vartheta| + \mathrm{e}^{\vartheta} - \mathrm{e}^{\theta} - \mathrm{e}^{\theta}(\vartheta - \theta)\right\} = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{1 - \mathrm{e}^{\theta} + \theta\mathrm{e}^{\theta}\right\} = 1 - \mathrm{e}^{\theta}(1+\theta), \tag{D.2}$$

which corresponds to the range $[\log(1-\alpha\lambda), \log(1+\alpha\lambda)]$ for θ. Combining (D.1) and (D.2) yields the first desired result.

Again, according to Definition 2.5,

$$\overrightarrow{\mathrm{P}}^{\exp}_{\lambda,\alpha|\cdot|}(\theta) = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\lambda\alpha|\vartheta| + \mathrm{e}^{\theta} - \mathrm{e}^{\vartheta} - \mathrm{e}^{\vartheta}(\theta-\vartheta)\right\}.$$

First-order conditions give

$$\begin{cases} \alpha\lambda - \mathrm{e}^{\vartheta}(\theta-\vartheta) = 0 & \text{if } \vartheta > 0, \\ -\alpha\lambda - \mathrm{e}^{\vartheta}(\theta-\vartheta) = 0 & \text{if } \vartheta < 0 \end{cases} \Leftrightarrow \begin{cases} (\vartheta-\theta)\mathrm{e}^{\vartheta-\theta} = -\alpha\lambda\mathrm{e}^{-\theta} & \text{if } \vartheta > 0, \\ (\vartheta-\theta)\mathrm{e}^{\vartheta-\theta} = \alpha\lambda\mathrm{e}^{-\theta} & \text{if } \vartheta < 0, \end{cases}$$

which implies

$$\begin{aligned} \vartheta^\star &= \begin{cases} W(-\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \vartheta^\star > 0 \text{ and } -\alpha\lambda\mathrm{e}^{-\theta} \geqslant -\mathrm{e}^{-1}, \\ W(\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \vartheta^\star < 0 \end{cases} \\ &= \begin{cases} W(-\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \theta > \alpha\lambda \text{ and } \theta \geqslant \log(\alpha\lambda) + 1, \\ W(\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \theta < -\alpha\lambda. \end{cases} \\ &= \begin{cases} W(-\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \theta > \alpha\lambda, \\ W(\alpha\lambda\mathrm{e}^{-\theta}) + \theta & \text{if } \theta < -\alpha\lambda. \end{cases} \end{aligned} \tag{D.3}$$

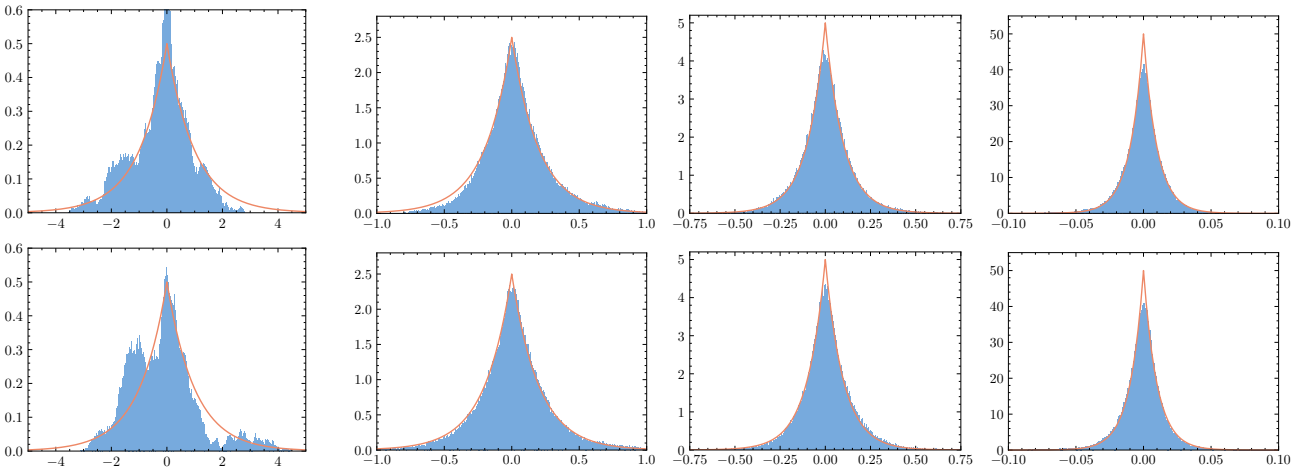
since $u \geqslant \log u + 1$ for any $u > 0$. Notice that the condition $-\alpha\lambda\mathrm{e}^{-\theta} \geqslant -\mathrm{e}^{-1}$ is required for the Lambert W function to be defined for a negative value.

On the other hand, if $\vartheta = 0$, then

$$\operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\lambda\alpha|\vartheta| + \mathrm{e}^{\theta} - \mathrm{e}^{\vartheta} - \mathrm{e}^{\vartheta}(\theta-\vartheta)\right\} = \operatorname*{argmin}_{\vartheta\in\mathbb{R}} \left\{\mathrm{e}^{\theta} - 1 - \theta\right\} = \mathrm{e}^{\theta} - (1+\theta), \tag{D.4}$$

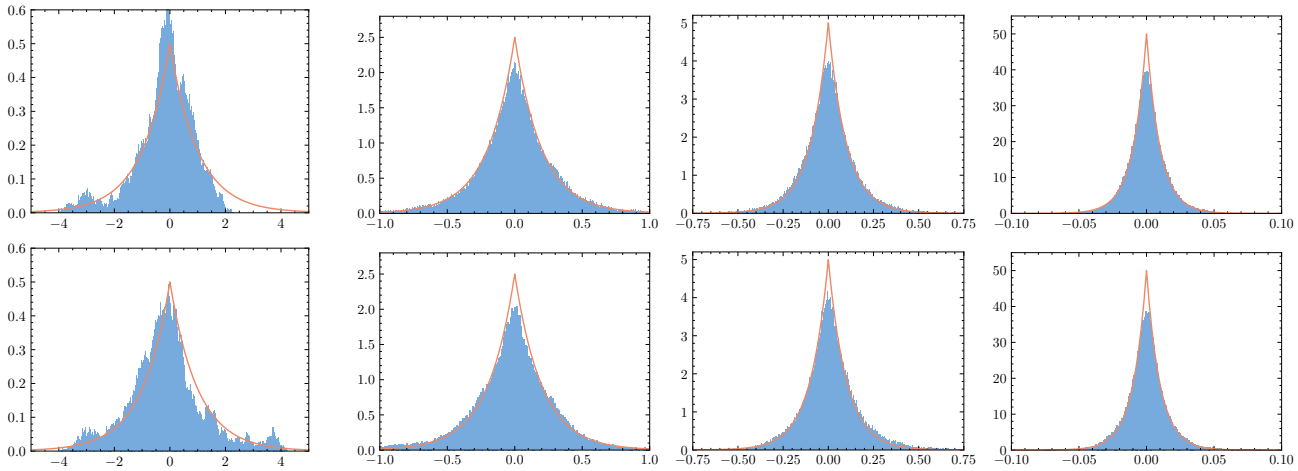
which corresponds to the range $[-\alpha\lambda, \alpha\lambda]$ for θ. Combining (D.3) and (D.4) yields the second desired result. □

The corresponding experiments are illustrated in Figure D.4. BMMMLA (Figure D.5) are also used in this setting. We observe that the right variants perform comparably to the left ones, both outperforming MYULA at the wide marginals (i.e., the lower dimensions).



(a) 1st dimension (b) 5th dimension (c) 10th dimension (d) 100th dimension

Figure D.4. Histograms of samples (in blue) from left BMUMLA (1st row), right BMUMLA (2nd row) and the true densities (in orange).



(a) 1st dimension (b) 5th dimension (c) 10th dimension (d) 100th dimension

Figure D.5. Histograms of samples (in blue) from left BMMMLA (1st row), right BMMMLA (2nd row) and the true densities (in orange).

D.2. Anisotropic Uniform Distribution

We consider the task of sampling from an anisotropic uniform distribution over the set $\mathcal{C} \coloneqq \prod_{i=1}^{d}[a_i, b_i]$, where $\boldsymbol{a} = (a_i)_{1\leqslant i\leqslant d}^{\top} \in \mathbb{R}^d$ and $\boldsymbol{b} = (b_i)_{1\leqslant i\leqslant d}^{\top} \in \mathbb{R}^d$. To perform this task using our proposed algorithm, we let $f = 0$ and $g = \iota_{\mathcal{C}}$. Note that the original mirror Langevin algorithm cannot apply to sampling uniform distributions, as mentioned in Li et al. (2022), as $f = 0$. However, by suitably choosing a Bregman–Moreau envelope, we can still perform approximate sampling (as opposed to exact sampling) using the BMMMLA.

Note that when $g = \iota_{\mathcal{C}}$ with $\mathcal{C} \subseteq \mathbb{R}^d$ being a closed convex set, the Bregman proximity operators of g are the Bregman projections (or projectors) onto $\mathcal{C}$, as illustrated in the following definition (Bauschke et al., 2018).

Definition D.2 (Bregman projections)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a closed convex set such that $\mathcal{X} \cap \mathcal{C} \neq \varnothing$, then $\overleftarrow{\mathrm{P}}_{\mathcal{C}}^{\varphi} \coloneqq \overleftarrow{\mathrm{P}}_{\iota_{\mathcal{C}}}^{\varphi}$ and $\overrightarrow{\mathrm{P}}_{\mathcal{C}}^{\varphi} \coloneqq \overrightarrow{\mathrm{P}}_{\iota_{\mathcal{C}}}^{\varphi}$ are the *left* and *right Bregman projections* onto $\mathcal{C}$ respectively.

For simplicity, we choose $\psi = \frac{1}{2}\|\cdot\|_2^2$. Then the Bregman projection onto $\mathcal{C}$ boils down to the Euclidean projection onto $\mathcal{C}$, which is given by

$$\overleftarrow{\mathrm{P}}_{\mathcal{C}}^{\varphi}(\boldsymbol{\theta}) = \overrightarrow{\mathrm{P}}_{\mathcal{C}}^{\varphi}(\boldsymbol{\theta}) = \mathrm{proj}_{\mathcal{C}}(\boldsymbol{\theta}) = (\min\{b_i, \max\{a_i, \theta_i\}\})_{1\leqslant i\leqslant d}^{\top}.$$

In the experiment, we consider the case where $a_i = -i$ and $b_i = i$ for all $i \in [\![d]\!]$, so that the target uniform distribution on $\mathcal{C} = [-1, 1] \times [-2, 2] \times \cdots \times [-d, d]$ is anisotropic, varying significantly across different dimensions. We use $\gamma = 0.01$, $\lambda = 1$ and $\boldsymbol{\beta} = (2\sqrt{d - i + 1})_{1\leqslant i\leqslant d}^{\top}$, and give the experimental results in Figures D.6 and D.7. We observe that BMUMLA outperforms MYULA at higher dimensions with wide marginals, where most samples lie in the desired ranges. Also note that all of Assumptions 3.1 to 3.5 and 3.8 hold. See Figure D.8 as a graphical verification of Assumption 3.8, with $\alpha = 2M_{\varphi_{\boldsymbol{\beta}}} + 0.1$ and $\beta_g = 250$.

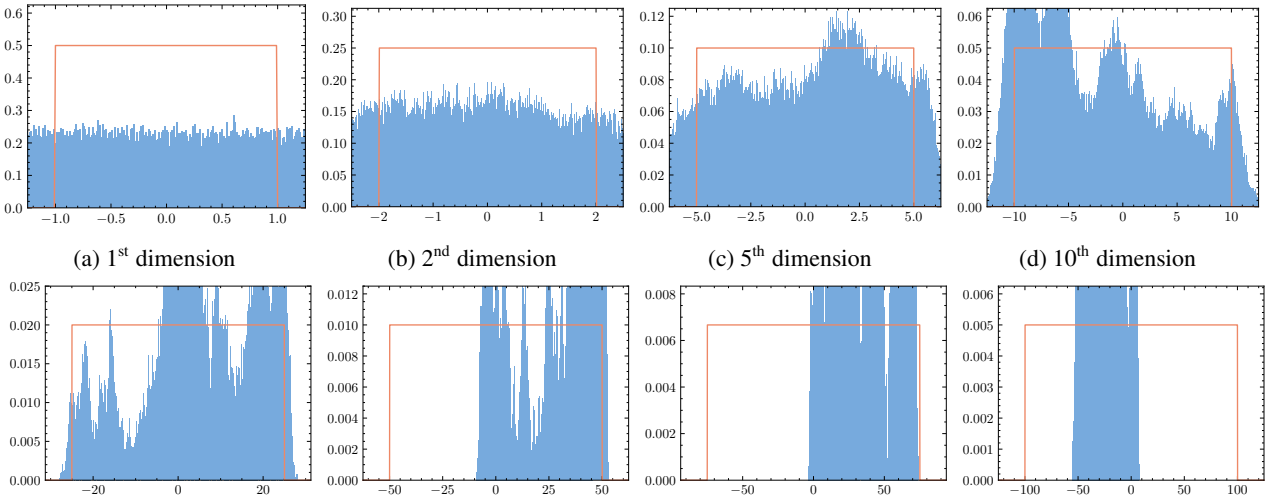
D.3. Bayesian Sparse Logistic Regression

We compare the performance of MYULA and BMUMLA in Bayesian sparse logistic regression. Suppose that we observe the samples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$, where $\boldsymbol{x}_n \in \mathbb{R}^d$ and $y_n \in \{0, 1\}$. In Bayesian logistic regression, the data are assumed to follow the model

$$y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}\left(\frac{\exp(\langle \boldsymbol{\theta}, \boldsymbol{x}_n\rangle)}{1 + \exp(\langle \boldsymbol{\theta}, \boldsymbol{x}_n\rangle)}\right), \tag{D.5}$$

for each $n \in [\![N]\!]$. The parameter $\boldsymbol{\theta} = (\theta_i)_{1\leqslant i\leqslant d}^{\top} \in \mathbb{R}^d$ is a random variable with a prior density p with respect to Lebesgue measure. Then, the posterior distribution of $\boldsymbol{\theta}$ takes the form

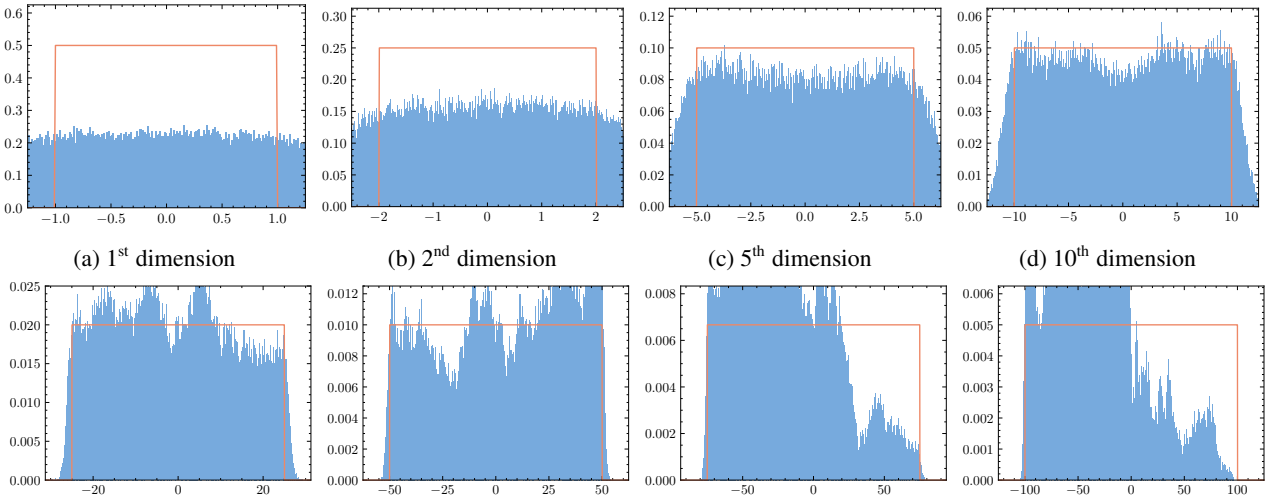
$$p(\boldsymbol{\theta} \mid \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}) \propto p(\boldsymbol{\theta}) \exp\left\{\sum_{n=1}^{N}(y_n\langle \boldsymbol{\theta}, \boldsymbol{x}_n\rangle - \log(1 + \exp(\langle \boldsymbol{\theta}, \boldsymbol{x}_n\rangle)))\right\}.$$



(a) 1st dimension (b) 2nd dimension (c) 5th dimension (d) 10th dimension

(e) 25th dimension (f) 50th dimension (g) 75th dimension (h) 100th dimension

Figure D.6. Histograms of samples (blue) from MYULA and the true densities (orange) for uniform distribution on $\mathcal{C}$.



(a) 1st dimension (b) 2nd dimension (c) 5th dimension (d) 10th dimension

(e) 25th dimension (f) 50th dimension (g) 75th dimension (h) 100th dimension

Figure D.7. Histograms of samples (blue) from BMUMLA and the true densities (orange) for uniform distribution on $\mathcal{C}$.

We are particularly concerned with the case with a prior in the form of a combination of an anisotropic Laplace distribution (which is sparsity-inducing) and a Gaussian distribution, where the unadjusted Langevin algorithm is no longer viable due to the nonsmoothness induced by the anisotropic Laplace distribution. In general, such a prior takes the form:

$$p(\boldsymbol{\theta}) \coloneqq p(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}_1, \alpha_2) \propto \exp\left\{ - \sum_{i=1}^{d} \alpha_{1,i} |\theta_i| - \frac{\alpha_2}{2} \sum_{i=1}^{d} \theta_i^2 \right\},$$

where $\boldsymbol{\alpha}_1 = (\alpha_{1,i})_{1 \leqslant i \leqslant d}^{\top} \in [0, +\infty[^d$ and $\alpha_2 \in [0, +\infty[$.

Then, the resulting posterior distribution has a potential of the following form:

$$U(\boldsymbol{\theta}) = \underbrace{\sum_{n=1}^{N} [\log(1 + \exp(\langle \boldsymbol{\theta}, \boldsymbol{x}_n \rangle)) - y_n \langle \boldsymbol{\theta}, \boldsymbol{x}_n \rangle] + \alpha_2 \|\boldsymbol{\theta}\|_2^2}_{=: f(\boldsymbol{\theta})} + \underbrace{\|\boldsymbol{\alpha}_1 \odot \boldsymbol{\theta}\|_1}_{=: g(\boldsymbol{\theta})}.$$

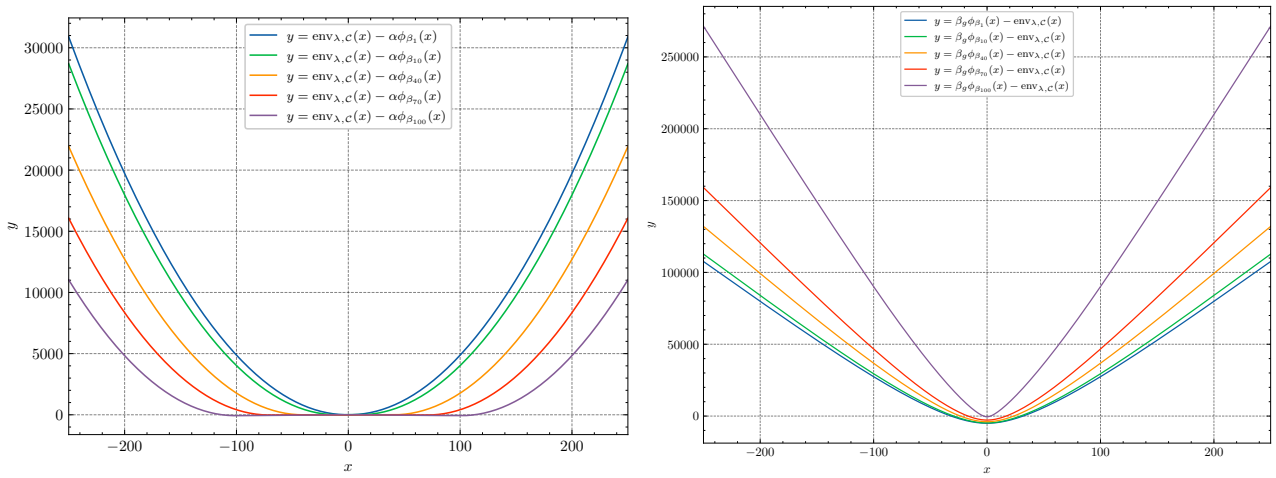


Figure D.8. Plots of $y = \mathrm{env}_{\lambda,\mathrm{e}}(x) - \alpha\phi_{\beta_i}(x)$ (left) and $y = \beta_g\phi_{\beta_i}(x) - \mathrm{env}_{\lambda,\mathrm{e}}(x)$ (right), for $i \in \{1, 10, 40, 70, 100\}$.

We take $d = 100$, $N = 1000$ and $\boldsymbol{\theta}^\star = (\mathbf{0}_{10}^\top, 0.1 \cdot \mathbf{1}_{10}^\top, 0.2 \cdot \mathbf{1}_{10}^\top, \ldots, 0.9 \cdot \mathbf{1}_{10}^\top)^\top \in \mathbb{R}^{100}$ as the ground truth. Then, each $x_{n,i}$ is generated from a standard Gaussian distribution and each y_n is sampled following (D.5) with $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$. In addition, we choose $\boldsymbol{\alpha}_1 = (10 \cdot \mathbf{1}_{10}^\top, 9 \cdot \mathbf{1}_{10}^\top, \ldots, 1 \cdot \mathbf{1}_{10}^\top)^\top$ and $\alpha_2 = 0.1$. Again, we use the hypentropy functions $\varphi_{\boldsymbol{\beta}}$ (for the mirror map) and $\psi_{\boldsymbol{\sigma}}$ (for the Bregman–Moreau envelope), with $\boldsymbol{\beta} = (2i^{1/4} \cdot \mathbf{1}_{10}^\top)_{1 \leqslant i \leqslant 10}^\top$ and $\boldsymbol{\sigma} = (\alpha_{1,i}^2)_{1 \leqslant i \leqslant d}^\top$. We also use a step size $\gamma = 5 \times 10^{-4}$ and a smoothing parameter $\lambda = 0.01$. Note that all of Assumptions 3.1 to 3.5 and 3.8 hold in this case. In particular, for Assumption 3.8, notice that f is indeed strongly convex.

We compare the performance of MYULA and BMUMLA by estimating the posterior means of $\boldsymbol{\theta}$ (as a whole or componentwise) and $\|\boldsymbol{\theta}\|_2^2/d$. We generate 30 samples (indexed by s) using each algorithm for 4000 iterations and average the samples to obtain estimates $\boldsymbol{\theta}_k = \frac{1}{30}\sum_{s=1}^{30} \boldsymbol{\theta}_{k,s}$ and $\|\boldsymbol{\theta}_k\|_2^2/d$ for the posterior means. From Figure D.9, we observe that the proposed left BMUMLA outperforms MYULA in the estimation of both posterior means.

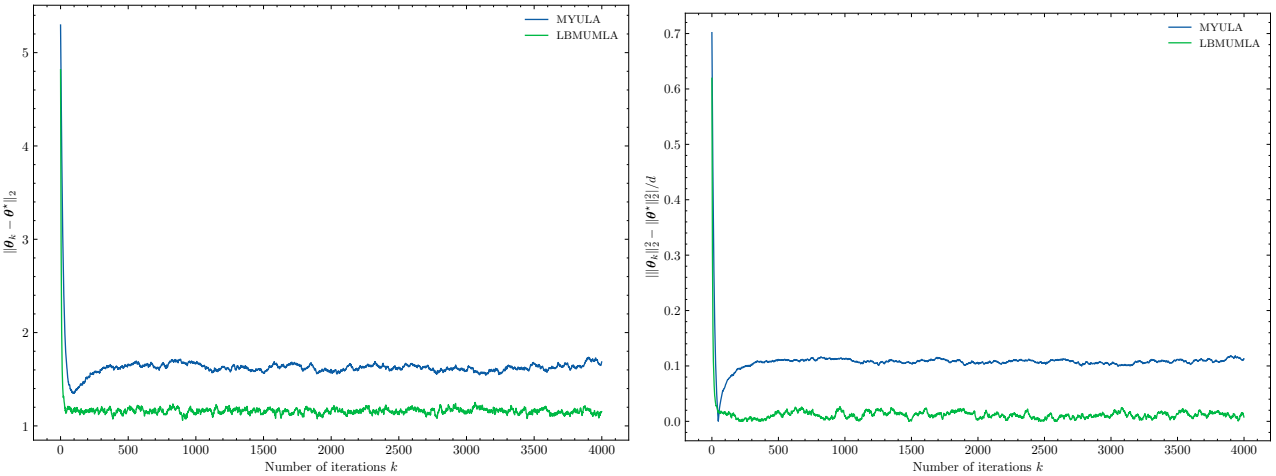
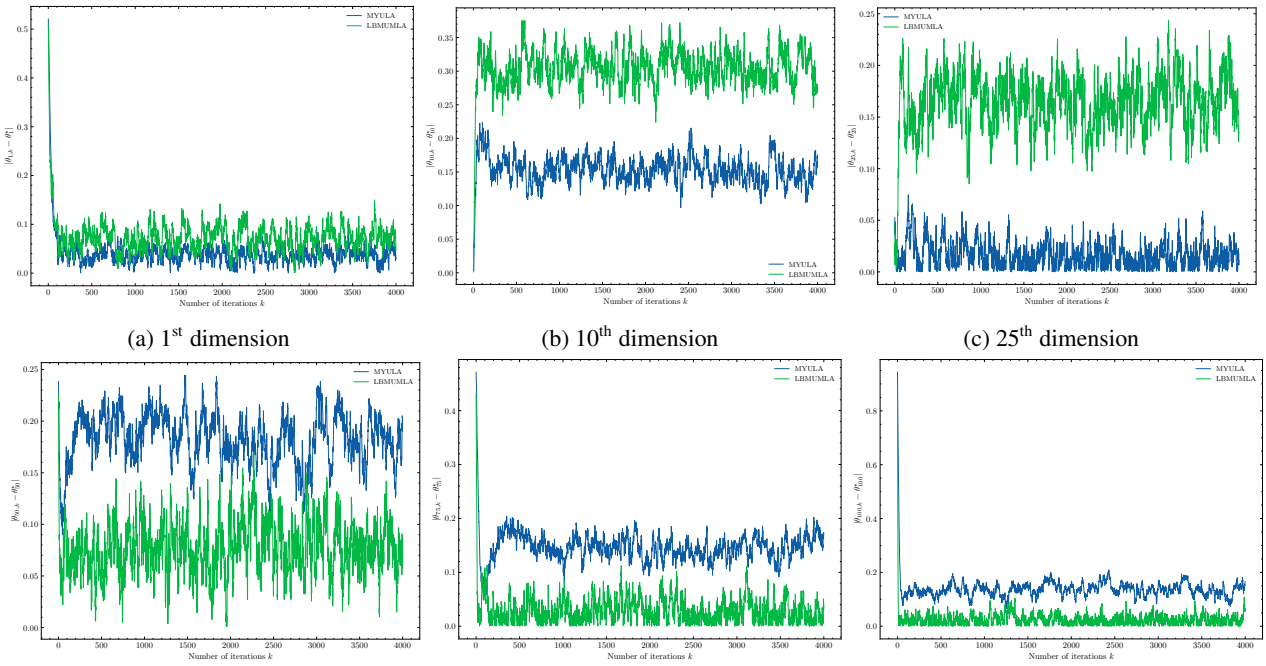


Figure D.9. Plots of estimation errors of the posterior means $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|_2$ (left) and $\big|\|\boldsymbol{\theta}_k\|_2^2 - \|\boldsymbol{\theta}^\star\|_2^2\big|/d$ (right).

We also plot the estimation errors of the posterior means of some components of $\boldsymbol{\theta}$. Figure D.10 reveals that MYULA gives smaller estimation errors than LBMUMLA at lower dimensions but high estimation errors at higher dimensions. However, we expect that the performance of BMUMLA would be further improved if $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ are more carefully picked or tuned, in order to fully adapt to the geometry of the posterior potential.



(a) 1st dimension (b) 10th dimension (c) 25th dimension

(d) 50th dimension (e) 75th dimension (f) 100th dimension

Figure D.10. Plots of estimation errors of the posterior means $|\theta_{k,i} - \theta_i^\star|$ for $i \in \{1, 25, 50, 75, 100\}$.