Primal Dual Interpretation of the Proximal Stochastic Gradient Langevin Algorithm

Adil Salim **Peter Richtárik**
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract

We consider the task of sampling with respect to a log concave probability distribution. The potential of the target distribution is assumed to be composite, *i.e.*, written as the sum of a smooth convex term, and a nonsmooth convex term possibly taking infinite values. The target distribution can be seen as a minimizer of the Kullback-Leibler divergence defined on the Wasserstein space (*i.e.*, the space of probability measures). In the first part of this paper, we establish a strong duality result for this minimization problem. In the second part of this paper, we use the duality gap arising from the first part to study the complexity of the Proximal Stochastic Gradient Langevin Algorithm (PSGLA), which can be seen as a generalization of the Projected Langevin Algorithm. Our approach relies on viewing PSGLA as a primal dual algorithm and covers many cases where the target distribution is not fully supported. In particular, we show that if the potential is strongly convex, the complexity of PSGLA is $\mathcal{O}(1/\varepsilon^2)$ in terms of the 2-Wasserstein distance. In contrast, the complexity of the Projected Langevin Algorithm is $\mathcal{O}(1/\varepsilon^{12})$ in terms of total variation when the potential is convex.

1 Introduction

Sampling from a target distribution is a fundamental task in machine learning. Consider the Euclidean space $\mathsf{X} = \mathbb{R}^d$ and a convex function $V : \mathsf{X} \to (-\infty, +\infty]$. Assuming that $\exp(-V)$ has a positive finite integral w.r.t. the Lebesgue measure Leb, we consider the task of sampling from the distribution $\mu^\star$ whose density is proportional to $\exp(-V(x))$ (we shall write $\mu^\star \propto \exp(-V)$).

If V is smooth, Langevin algorithm produces a sequence of iterates (x^k) asymptotically distributed according to a distribution close to $\mu^\star$. Langevin algorithm performs iterations of the form

$$x^{k+1} = x^k - \gamma \nabla V(x^k) + \sqrt{2\gamma} W^{k+1}, \tag{1}$$

where $\gamma > 0$ and $(W^k)_k$ is a sequence of i.i.d. standard Gaussian vectors in X. Each iteration of (1) can be seen as a gradient descent step for V, where the gradient of V is perturbed by a Gaussian vector. Hence, the iterations of Langevin algorithm look like those of the stochastic gradient algorithm; however the noise in Langevin algorithm is scaled by $\sqrt{\gamma}$ instead of γ. Nonasymptotic bounds for Langevin algorithm have been established in [17, 20]. Moreover, Langevin algorithm can be interpreted as an inexact gradient descent method to minimize the Kullback-Leibler (KL) divergence w.r.t. $\mu^\star$ in the space of probability measures [1, 5, 14, 19, 23, 33].

In many applications, the function V is naturally written as the sum of a smooth and a nonsmooth term. In Bayesian statistics for example, $\mu^\star$ typically represents some posterior distribution. In this case, V is the sum of the $\log$-likelihood (which is itself a sum over the data points) and the possibly nonsmooth potential of the prior distribution [19, 21, 32], which plays the role of a regularizer. In some other applications in Bayesian learning, the support of $\mu^\star$ is not the whole space X [9, 10] (*i.e.*, V can take the value $+\infty$). In order to cover these applications, we consider the case where V is

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

written as

$$V(x) := \mathbb{E}_\xi(f(x,\xi)) + G(x), \tag{2}$$

where ξ is a random variable, $f(\cdot, s) : \mathsf{X} \to \mathbb{R}$ for every $s \in \Xi$, $F(x) = \mathbb{E}_\xi(f(x,\xi))$ is smooth and convex and $G : \mathsf{X} \to (-\infty, +\infty]$ is nonsmooth and convex. We assume to have access to the stochastic gradient $\nabla_x f(x,\xi)$ (where ξ is a random variable with values in Ξ) and to the proximity operator $\mathrm{prox}_{\gamma G}$ of G. The template (2) covers many log concave densities [10, 13, 19, 21]. In optimization, the minimization of V can be efficiently tackled by the proximal stochastic gradient algorithm [3]. Inspired by this optimization algorithm, the Proximal Stochastic Gradient Langevin Algorithm (PSGLA) [19] is the method performing proximal stochastic gradient Langevin steps of the form

$$x^{k+1} = \mathrm{prox}_{\gamma G}\left(x^k - \gamma \nabla_x f(x^k, \xi^{k+1}) + \sqrt{2\gamma} W^{k+1}\right), \tag{3}$$

where $\gamma > 0$, (W^k) is a sequence of i.i.d. standard Gaussian random vectors in X, and (ξ^k) is a sequence of i.i.d. copies of ξ. Remarkably, the iterates x^k of PSGLA remain in the domain of G, *i.e.*, the support of $\mu^\star$, a property that is useful in many contexts. When G is Lipschitz continuous, the support of $\mu^\star$ is X and PSGLA can be interpreted as an inexact proximal gradient descent method for minimizing KL, with convergence rates proven in terms of the KL divergence [19]. However, for general G, the KL divergence can take infinite values along PSGLA. Therefore, a new approach is needed.

1.1 Related works

First, various instances of the PSGLA algorithm have already been considered. The only instance allowing $G(x)$ to be infinite (*i.e.*, the support of $\mu^\star$ not to be X) is the Projected Langevin Algorithm [10], which corresponds to our setting in the special case with $G = \iota_C$ (*i.e.*, the indicator function of a convex body[1] C), and $\nabla f(\cdot, s) \equiv \nabla F$ for every s (*i.e.*, the full gradient of F). In this case, $\mathrm{prox}_{\gamma G}$ is the orthogonal projection onto C and $\mu^\star$ is supported by C. Bubeck et al [10] provide complexity results in terms of sufficient number of iterations to achieve ε accuracy in terms of the Total Variation between the target distribution $\mu^\star$ and the current iterate distribution. Assuming that F is convex and smooth, the complexity of the Projected Langevin Algorithm is $\mathcal{O}(1/\varepsilon^{12})$[2], and if $F \equiv 0$, the complexity is improved to $\mathcal{O}(1/\varepsilon^8)$.

Other instances of PSGLA were proposed in the case where G is Lipschitz continuous or smooth (and hence finite). Wibisono [33] considered the case with $F = G$ and $\nabla f(\cdot, s) \equiv \nabla F$, proposing the Symmetrized Langevin Algorithm (SLA), and showed that the current iterate distribution converges linearly in Wasserstein distance to the invariant measure of the SLA, if F is strongly convex and smooth. Durmus et al [19] considered the case where G is Lipschitz continuous, and showed that the complexity of PSGLA is $\mathcal{O}(1/\varepsilon^2)$ in terms of the KL divergence and $\mathcal{O}(1/\varepsilon^4)$ in terms of the Total Variation distance if F is convex and smooth. If F is strongly convex, the complexity is $\mathcal{O}(1/\varepsilon^2)$ in Wasserstein distance and $\mathcal{O}(1/\varepsilon)$ in KL divergence. Bernton [5] studied a setting similar to [19] and derived a similar result for the Proximal Langevin Algorithm (*i.e.*, PSGLA without the gradient step) in the strongly convex case. The Proximal Langevin Algorithm was also studied in a recent paper of Wibisono [34], where a rapid convergence result was proven in the case where G is nonconvex but satisfies further smoothness and geometric assumptions.

Second, the task of sampling w.r.t. $\mu^\star$, where G is nonsmooth and possibly takes infinite values, using Langevin algorithm, has also been considered. When F is strongly convex and G an indicator function of a bounded convex set, the existence of an algorithm achieving $\mathcal{O}(1/\varepsilon^2)$ in Wasserstein and Total Variation distances was proven by Hsieh et al [22, Theorem 3]. However, an actual algorithm is only given in a specific, although nonconvex, case. Besides, MYULA (Moreau-Yosida Unadjusted Langevin Algorithm) [9, 21] can tackle the task of sampling from $\mu^\star$ efficiently. MYULA is equivalent to Langevin algorithm (1) applied to sampling from $\mu^\lambda \propto \exp(-F - G^\lambda)$, where G^λ is the Moreau-Yosida approximation of G [2]. By choosing the smoothing parameter $\lambda > 0$ appropriately, and making assumptions that allow to control the distance between μ^λ and $\mu^\star$ (*e.g.*, G Lipschitz or $G = \iota_C$), complexity results for MYULA were established in [9, 21]. For example, if G is the indicator function of a convex body, Brosse et al [9] show that the complexity

[1] A convex body is a compact convex set with a nonempty interior.

[2] Our big O notation ignores logarithm factors.

of MYULA is $\mathcal{O}(1/\varepsilon^6)$ in terms of the Total Variation distance (resp. 1-Wasserstein distance) if F is convex and smooth (resp., if F is strongly convex and smooth), provided that the algorithm is initialized from a minimizer of V. Similarly to PSGLA, MYULA involves one proximal step and one gradient step per iteration. However, the support of the smoothed distribution μ^λ is always X (even if $\mu^\star$ is not fully supported), and therefore the iterates of MYULA do not remain in the support of the target distribution $\mu^\star$, contrary to PSGLA.

Finally, the task of sampling w.r.t. $\mu^\star$, where V is not smooth but finite, has also been considered. The Perturbed Langevin Algorithm proposed by Chatterji et al [12] allows to sample from $\mu^\star$ in the case when G satisfies a weak form of smoothness (generalizing both Lipschitz continuity and smoothness) and without accessing its proximity operator. Finally, if G is Lipschitz continuous, the Stochastic Proximal Langevin Algorithm proposed by Salim et al [27] and Schechtman et al [28] allows to sample from $\mu^\star$ using cheap stochastic proximity operators only.

1.2 Contributions

In summary, PSGLA has complexity $\mathcal{O}(1/\varepsilon^2)$ in 2-Wasserstein distance if F is strongly convex [19] and G is Lipschitz. The only instance of PSGLA allowing G to be infinite is the Projected Langevin Algorithm. It has complexity $\mathcal{O}(1/\varepsilon^{12})$ in Total Variation [10][3] and only applies to the case where G is the indicator of a convex body. In the latter case, another Langevin algorithm called MYULA has complexity $\mathcal{O}(1/\varepsilon^6)$ in 1-Wasserstein distance [9], but allows the iterates to leave the support of $\mu^\star$. Besides, still in the case where G is an indicator function, there exists a Langevin algorithm achieving $\mathcal{O}(1/\varepsilon^2)$ rate in the Wasserstein distance [22].

In this paper, we consider other (i.e., new) cases where G can take infinite values. More precisely, we consider a general nonsmooth convex function G and we assume that $\exp(-V)$ has a mild Sobolev regularity. We develop new mathematical tools (e.g., a Lagrangian for the minimization of KL), that have their own interest, to obtain our complexity results. Our main result is to show that, surprisingly, PSGLA still has the complexity $\mathcal{O}(1/\varepsilon^2)$ in 2-Wasserstein distance if F is strongly convex, although G can take infinite values. We also show that, if F is just convex, PSGLA has the complexity $\mathcal{O}(1/\varepsilon^2)$ in terms of a newly defined duality gap, which can be seen as the notion that replaces KL, since KL can be infinite.

Our approach follows the line of works [5, 14, 19, 23, 25, 31, 33, 34] that formulate the task of sampling form $\mu^\star$ as the problem of minimizing the KL divergence w.r.t $\mu^\star$. In summary, our contributions are the following.

• In the first part of the paper, we reformulate the task of sampling from $\mu^\star$ as the resolution of a monotone inclusion defined on the space of probability measures. We subsequently use this reformulation to define a duality gap for the minimization of the KL divergence, and show that strong duality holds.

• In the second part of this paper, we use this reformulation to represent PSGLA as a primal dual stochastic Forward Backward algorithm involving monotone operators.

• This new representation of PSGLA, along with the strong duality result from the first part, allows us to prove new complexity results for PSGLA that extend and improve the state of the art.

• Finally, we conduct some numerical experiments for sampling from a distribution supported by a set of matrices (see appendix).

In the first part we combine tools from optimization duality [16] and optimal transport [1] and in the second part we combine tools from the analysis of the Langevin algorithm [19], and the analysis of primal dual optimization algorithms [11, 18].

The remainder is organized as follows. In Section 2 we provide some background knowledge on convex analysis and optimal transport. In Section 3 we develop a primal dual optimality theory for the task of sampling from $\mu^\star$. In Section 4 we give a new representation of PSGLA using monotone operators. We use it to state our main complexity result on PSGLA in Section 5. Numerical experiments and all proofs are postponed to the appendix. Therein, we also provide further intuitions on PSGLA, namely the connection between gradient descent and Langevin algorithm [19] and the

[3] This result also holds if F is not strongly convex.

connection between primal dual optimization and our approach. Finally, an extension of PSGLA for handling a third (stochastic, Lipschitz continuous and proximable) term in the definition of the potential V (2) is provided at the end of the appendix.

2 Background

Throughout this paper, we use the conventions $\exp(-\infty) = 0$ and $1/0 = +\infty$.

2.1 Convex analysis

In this section, we recall some facts from convex analysis. These facts will be used in the proofs without mention. For more details, the reader is referred to [4].

2.1.1 Convex optimization

By $\Gamma_0(\mathsf{X})$ we denote the set of proper, convex, lower semicontinuous functions $\mathsf{X} \to (-\infty, +\infty]$. A function $F \in \Gamma_0(\mathsf{X})$ is L-smooth if F is differentiable and its gradient ∇F is L-Lipschitz continuous. Consider $G \in \Gamma_0(\mathsf{X})$ and denote $\mathrm{dom}(G) := \{x \in \mathsf{X} \ : \ G(x) < \infty\}$ its domain. Given $x \in \mathsf{X}$, a subgradient of G at x is any vector $y \in \mathsf{X}$ satisfying

$$G(x) + \langle y, x' - x\rangle \leq G(x'), \tag{4}$$

for every $x' \in \mathsf{X}$. If the set $\partial G(x)$ of subgradients of G at x is not empty, then there exists a unique element of $\partial G(x)$ with minimal norm. This particular subgradient is denoted $\partial^0 G(x)$. The set valued map $\partial G(\cdot)$ is called the subdifferential. The proximity operator of G, denoted prox_G, is defined by

$$\mathrm{prox}_G(x) := \operatorname*{arg\,min}_{x' \in \mathsf{X}} \left\{ G(x') + \tfrac{1}{2}\|x - x'\|^2 \right\}. \tag{5}$$

By $\iota_C(\cdot)$ we denote the indicator function of set C given by $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = +\infty$ if $x \notin C$. If $G = \iota_C$, where C is a closed convex set, then prox_G is the orthogonal projection onto C. Moreover, $\mathrm{prox}_G(x)$ is the only solution x' to the inclusion $x \in x' + \partial G(x')$. The Fenchel transform of G is the function $G^* \in \Gamma_0(\mathsf{X})$ defined by $G^*(y) := \sup_{x \in \mathsf{X}} \{\langle y, x\rangle - G(x)\}$. Several properties relate G to its Fenchel transform G^*. First, the Fenchel transform of G^* is G. Then, the subdifferential ∂G^* is characterized by the relation $x \in \partial G^*(y) \Leftrightarrow y \in \partial G(x)$. Finally, G^* is λ-strongly convex if and only if G is $1/\lambda$-smooth.

2.1.2 Maximal monotone operators

A set valued function $A : \mathsf{X} \rightrightarrows \mathsf{X}$ is *monotone* if $\langle y - y', x - x'\rangle \geq 0$ whenever $y \in A(x)$ and $y' \in A(x')$. The *inverse* of A, denoted A^{-1}, is defined by the relation $x \in A^{-1}(y) \Leftrightarrow y \in A(x)$, and the *set of zeros* of A is $Z(A) := A^{-1}(0)$. If A is monotone, A is *maximal* if its *resolvent*, *i.e.*, the map $J_A : x \mapsto (I + A)^{-1}(x)$, is single valued. If $G \in \Gamma_0(\mathsf{X})$, then ∂G is a maximal monotone operator and $J_{\partial G} = \mathrm{prox}_G$. Moreover, $Z(\partial G) = \arg\min G$ and $(\partial G)^{-1} = \partial G^*$. If S is a skew symmetric matrix on X, the operator $x \mapsto Sx$ is maximal monotone. Finally, the sum $\partial G + S$ is also a maximal monotone operator. Many problems in optimization can be cast as the problem of finding a zero x of the sum of two maximal monotone operators $0 \in (A + B)(x)$ [16]. For instance, $Z(\nabla F + \partial G) = \arg\min F + G$. To solve this problem, the Forward Backward algorithm is given by the iteration $x^{k+1} = J_{P^{-1}A}(x^k - P^{-1}B(x^k))$, where P is a symmetric positive definite matrix ($P \in \mathbb{R}^{d\times d}_{++}$),[4] and B is single valued. Using the definition of the resolvent, the Forward Backward algorithm can equivalently be written as

$$P(x^{k+1/2} - x^k) = -\gamma B(x^k), \quad P(x^{k+1} - x^{k+1/2}) \in -\gamma A(x^{k+1}). \tag{6}$$

2.2 Optimal transport

In this section, we recall some facts from optimal transport theory. These facts will be used in the proofs without mention. For more details, the reader is referred to Ambrosio et al [1].

[4] The operators $P^{-1}A$ and $P^{-1}B$ are not monotone in general, however they are monotone under the inner product induced by P.

2.2.1 Wasserstein distance

By $\mathscr{B}(\mathsf{X})$ we denote the σ-field of Lesbesgue measurable subsets of X, and by $\mathcal{P}_2(\mathsf{X})$ the set of probability measures μ over $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$ with finite second moment $\int \|x\|^2 d\mu(x) < \infty$. Denote $\mathrm{supp}(\mu)$ the support of μ. The identity map I belongs to the Hilbert space $L^2(\mu; \mathsf{X})$ of μ-square integrable random vectors in X. We denote $\langle \cdot, \cdot \rangle_\mu$ (resp. $\| \cdot \|_\mu$) the inner product (resp. the norm) in this space. Given $T : \mathsf{X} \to \mathsf{Z}$, where Z is some Euclidean space, the *pushforward measure* of μ by T, also called the *image measure*, is defined by $T\#\mu(A) := \mu(T^{-1}(A))$ for every $A \in \mathscr{B}(\mathsf{Z})$. Consider $\mu, \nu \in \mathcal{P}_2(\mathsf{X})$. A *coupling* υ between μ and ν (we shall write $\upsilon \in \Gamma(\mu, \nu)$) is a probability measure over $(\mathsf{X}^2, \mathscr{B}(\mathsf{X}^2))$ such that $x^\star \# \upsilon = \mu$, where $x^\star : (x, y) \mapsto x$, and $y^\star \# \upsilon = \nu$, where $y^\star : (x, y) \mapsto y$. In other words, (X, Y) is a random variable such that the distribution of X is μ (we shall write $X \sim \mu$) and $Y \sim \nu$ if and only if the distribution of (X, Y) is a coupling. The (2-)*Wasserstein distance* is then defined by

$$W^2(\mu, \nu) := \inf_{\upsilon \in \Gamma(\mu,\nu)} \int \|x - y\|^2 d\upsilon(x, y). \tag{7}$$

Let $\mathcal{P}_2^r(\mathsf{X})$ be the set of elements $\mu \in \mathcal{P}_2(\mathsf{X})$ such that μ is absolutely continuous w.r.t. Leb (we shall write $\mu \ll \mathrm{Leb}$). Brenier's theorem asserts that if $\mu \in \mathcal{P}_2^r(\mathsf{X})$, then the $\inf$ defining $W^2(\mu, \nu)$ is actually a $\min$ achieved by a unique minimizer υ. Moreover, there exists a uniquely determined μ-almost everywhere (a.e.) map $T_\mu^\nu : \mathsf{X} \to \mathsf{X}$ such that $\upsilon = (I, T_\mu^\nu)\#\mu$, where $(I, T_\mu^\nu) : x \mapsto (x, T_\mu^\nu(x))$. In this case, T_μ^ν is called the *optimal pushforward* from μ to ν and satisfies

$$W^2(\mu, \nu) = \int \|x - T_\mu^\nu(x)\|^2 d\mu(x). \tag{8}$$

2.2.2 Geodesically convex functionals

We shall consider several functionals defined on the space $\mathcal{P}_2(\mathsf{X})$. For every $\mu \in \mathcal{P}_2^r(\mathsf{X})$ with density denoted $\mu(x)$ w.r.t. Leb, the *entropy* is defined by

$$\mathcal{H}(\mu) := \int \log(\mu(x)) d\mu(x), \tag{9}$$

and if $\mu \notin \mathcal{P}_2^r(\mathsf{X})$, then $\mathcal{H}(\mu) := +\infty$. Given $V \in \Gamma_0(\mathsf{X})$, the *potential energy* is defined for every $\mu \in \mathcal{P}_2(\mathsf{X})$ by

$$\mathcal{E}_V(\mu) := \int V(x) d\mu(x). \tag{10}$$

Finally, if $\mu' \in \mathcal{P}_2(\mathsf{X})$ such that $\mu \ll \mu'$, the *Kullback-Leibler (KL) divergence* is defined by

$$\mathrm{KL}(\mu|\mu') := \int \log\left(\tfrac{d\mu}{d\mu'}(x)\right) d\mu(x), \tag{11}$$

where $\frac{d\mu}{d\mu'}$ denotes the density of μ w.r.t. μ', and $\mathrm{KL}(\mu|\mu') := +\infty$ if μ is not absolutely continuous w.r.t. μ'. The functionals $\mathcal{H}$, $\mathcal{E}_V$ and $\mathrm{KL}(\cdot|\mu^\star)$ satisfy a form of convexity over $\mathcal{P}_2(\mathsf{X})$ called *geodesic convexity*. If $\mathcal{F} : \mathcal{P}_2(\mathsf{X}) \to (-\infty, +\infty]$ is geodesically convex, then for every $\mu \in \mathcal{P}_2^r(\mathsf{X})$, $\mu' \in \mathcal{P}_2(\mathsf{X})$, and $\alpha \in [0, 1]$, $\mathcal{F}\left((\alpha T_\mu^{\mu'} + (1 - \alpha) I)\#\mu\right) \leq \alpha \mathcal{F}(\mu') + (1 - \alpha)\mathcal{F}(\mu)$. Given $\mu \in \mathcal{P}_2^r(\mathsf{X})$, a *(Wasserstein) subgradient* of $\mathcal{F}$ at μ is a random variable $Y \in L^2(\mu; \mathsf{X})$ such that for every $\mu' \in \mathcal{P}_2(\mathsf{X})$,

$$\mathcal{F}(\mu) + \langle Y, T_\mu^{\mu'} - I \rangle_\mu \leq \mathcal{F}(\mu'). \tag{12}$$

Moreover, if Y' is a subgradient of $\mathcal{F}$ at μ', then the following monotonicity property holds

$$\langle Y' \circ T_\mu^{\mu'} - Y, T_\mu^{\mu'} - I \rangle_\mu \geq 0. \tag{13}$$

If the set $\boldsymbol{\partial}\mathcal{F}(\mu) \subset L^2(\mu; \mathsf{X})$ of subgradients of $\mathcal{F}$ at μ is not empty, then there exists a unique element of $\boldsymbol{\partial}\mathcal{F}(\mu)$ with minimal norm. This particular subgradient is denoted $\boldsymbol{\partial}^0\mathcal{F}(\mu)$. However, the set $\boldsymbol{\partial}\mathcal{F}(\mu)$ might be empty. A typical condition for nonemptiness requires the density $\mu(x)$ to have some Sobolev regularity. For every open set $\Omega \subset \mathsf{X}$, we denote $S^{1,1}(\Omega)$ the Sobolev space of Leb-integrable functions $u : \Omega \to \mathbb{R}$ admitting a Leb-integrable weak gradient $\nabla u : \Omega \to \mathsf{X}$. We say that $u \in S^{1,1}_{\mathrm{loc}}(\Omega)$ if $u \in S^{1,1}(K)$ for every bounded open set $K \subset \Omega$. Obviously, $S^{1,1}(\Omega) \subset S^{1,1}_{\mathrm{loc}}(\Omega)$.

2.3 Assumptions on F and G

Consider $F : \mathsf{X} \to \mathbb{R}$ and $G : \mathsf{X} \to (-\infty, +\infty]$. We make the following assumptions.

Assumption 1. The function F is convex and L-smooth. Moreover, $G \in \Gamma_0(\mathsf{X})$.

Note that $V := F + G \in \Gamma_0(\mathsf{X})$. We denote λ_F (resp. λ_{G^*}) the strong convexity parameter of F (resp. G^*), equal to zero if F (resp. G^*) is not strongly convex.

Assumption 2. The integral $\int \exp(-V) d\,\mathrm{Leb}$ is positive and finite.

Assumption 2 is needed to define the target distribution $\mu^\star \propto \exp(-V)$, and implies that $\mathrm{int}(D) \neq \emptyset$, where $D := \mathrm{dom}(V)$.

Lemma 1. If Assumptions 1 and 2 hold, then

$$\int |V(x)| \exp(-V(x)) dx < \infty, \quad \text{and} \quad \int \|x\|^2 \exp(-V(x)) dx < \infty.$$

Lemma 1 implies that $\mu^\star \in \mathcal{P}_2(\mathsf{X})$ and using Assumption 1, $\|\nabla F\| \in L^2(\mu^\star; \mathbb{R})$. Since $G \in \Gamma_0(\mathsf{X})$, G is differentiable Leb-a.e. (almost everywhere) on $\mathrm{int}(D)$, see [24, Theorem 25.5].

Assumption 3. The integral $\int_{\mathrm{int}(D)} \|\nabla G\|^2 \exp(-V) d\,\mathrm{Leb}$ is finite.

Assumption 3 is equivalent to requiring $\|\nabla G\| \in L^2(\mu^\star; \mathbb{R})$, see below. Moreover, we assume the following regularity property for the function $\exp(-V)$.

Assumption 4. The function $\exp(-V)$ belongs to the space $S^{1,1}_{\mathrm{loc}}(\mathsf{X})$.

Assumption 4 is a necessary condition for $\boldsymbol{\partial}\mathcal{H}(\mu^\star) \neq \emptyset$, see below. This assumption precludes $\mu^\star$ from being a uniform distribution. However, Assumption 4 is quite general, *e.g.*, $\exp(-V)$ need not be continuous or positive (see the numerical experiment section). Finally, we assume that the stochastic gradients of F have a bounded variance. Consider an abstract measurable space $(\Xi, \mathscr{G})$, and a random variable ξ with values in $(\Xi, \mathscr{G})$.

Assumption 5. For every $x \in \mathsf{X}$, $f(x, \xi)$ is integrable and $F(x) = \mathbb{E}_\xi(f(x, \xi))$. Moreover, there exists $\sigma_F \geq 0$ such that for every $x \in \mathsf{X}$, $\mathbb{V}_\xi(\|\nabla f(x, \xi)\|) \leq \sigma_F^2$, where $\mathbb{V}$ denotes the variance.

The last assumption implies that the stochastic gradients are unbiased: $\mathbb{E}_\xi(\nabla f(x, \xi)) = \nabla F(x)$ for every $x \in \mathsf{X}$.

3 Primal dual optimality in Wasserstein space

Let $\mathcal{F} : \mathcal{P}_2(\mathsf{X}) \to (-\infty, +\infty]$ be defined by

$$\mathcal{F}(\mu) := \mathcal{H}(\mu) + \mathcal{E}_V(\mu) = \mathcal{H}(\mu) + \mathcal{E}_F(\mu) + \mathcal{E}_G(\mu). \tag{14}$$

Using Lemma 1, $\mathcal{H}(\mu^\star)$ and $\mathcal{E}_V(\mu^\star)$ are finite real numbers. Moreover, using [19, Lemma 1.b], for every $\mu \in \mathcal{P}_2(\mathsf{X})$ such that $\mathcal{E}_V(\mu) < \infty$, we have the identity

$$\mathcal{F}(\mu) - \mathcal{F}(\mu^\star) = \mathrm{KL}(\mu | \mu^\star). \tag{15}$$

Equation (15) says that $\mu^\star$ is the unique minimizer of $\mathcal{F}$: $\mu^\star = \arg\min \mathcal{F}$.

3.1 Subdifferential calculus

The following result is a consequence of [1, Theorem 10.4.13].

Theorem 2. Let $\mu \propto \rho$ be an element of $\mathrm{dom}(\mathcal{F})$. Then, $\mathrm{supp}(\mu) \subset \overline{D}$ and $\mu(\overline{D} \setminus \mathrm{int}(D)) = 0$. Moreover, $\boldsymbol{\partial}\mathcal{F}(\mu) \neq \emptyset$ if and only if $\rho \in S^{1,1}_{\mathrm{loc}}(\mathrm{int}(D))$ and there exists $w \in L^2(\mu)$ such that

$$w(x)\rho(x) = \nabla\rho(x) + \rho(x)\nabla V(x), \tag{16}$$

for μ-a.e. x. In this case, $w = \boldsymbol{\partial}^0\mathcal{F}(\mu)$.

If Assumptions 1 and 2 hold, then $\mathcal{F}(\mu^\star) < \infty$ using Lemma 1. Then, Theorem 2 implies that $\mu^\star(\mathrm{int}(D)) = 1$. Therefore, using [24, Theorem 25.5], G and V are $\mu^\star$-a.s. differentiable.

Moreover, applying Theorem 2 with $V \equiv 0$, we can replace $\mathcal{F}$ by $\mathcal{H}$ and D by X. We obtain that $\boldsymbol{\partial}\mathcal{H}(\mu) \neq \emptyset$ if and only if $\rho \in S^{1,1}_{\mathrm{loc}}(\mathsf{X})$ and $w\rho = \nabla\rho$ for some $w \in L^2(\mu;\mathsf{X})$. Now, we set $\mu = \mu^\star$ and $\rho = \exp(-V)$. Using Assumption 4 and $w = -\nabla V$, we obtain that $\boldsymbol{\partial}^0\mathcal{H}(\mu^\star) = -\nabla V$ $\mu^\star$-a.e. Therefore, using that ∇G is well defined $\mu^\star$-a.e., $\mu^\star$ satisfies

$$0 = \nabla F(x) + \boldsymbol{\partial}^0\mathcal{H}(\mu^\star)(x) + \nabla G(x), \text{ for } \mu^\star - \text{a.e. } x. \tag{17}$$

Equation (17) can be seen as the first order optimality conditions associated with the minimization of the functional $\mathcal{F}$. Consider the "dual" variable $Y^\star : x \mapsto \nabla G(x)$ defined $\mu^\star$ a.e. Using Assumption 3 and $\mu^\star(\mathrm{int}(D)) = 1$, $Y^\star \in L^2(\mu^\star;\mathsf{X})$. We can express the first order optimality condition (17) as $0 = \nabla F(x) + \boldsymbol{\partial}^0\mathcal{H}(\mu^\star)(x) + Y^\star(x)$, $\mu^\star$ a.e. Besides, $Y^\star(x) \in \partial G(x)$, therefore $0 \in -x + \partial G^*(Y^\star(x))$ using $\partial G^* = (\partial G)^{-1}$. Denote $\nu^\star := Y^\star \# \mu^\star \in \mathcal{P}_2(\mathsf{X})$ and $\pi^\star := (I, Y^\star)\#\mu^\star \in \mathcal{P}_2(\mathsf{X}^2)$. The relationship between $\mu^\star$ and $Y^\star$ can be summarized as

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \nabla F(x) + \boldsymbol{\partial}^0\mathcal{H}(\mu^\star)(x) & +y \\ -x & +\partial G^*(y) \end{bmatrix} \text{ for } \pi^\star \text{ a.e. } (x,y). \tag{18}$$

In the sequel, we fix the probability space $(\Omega, \mathscr{F}, \mathbb{P}) = (\mathsf{X}^2, \mathcal{B}(\mathsf{X}^2), \pi^\star)$, denote $\mathbb{E}$ the mathematical expectation and L^2 the space $L^2(\Omega, \mathscr{F}, \mathbb{P}; \mathsf{X})$. The expression "almost surely" (a.s.) will be understood w.r.t. $\mathbb{P}$. Recall that $x^\star$ is the map $(x,y) \mapsto x$ and $y^\star : (x,y) \mapsto y$. Using Assumption 3, $x^\star, y^\star \in L^2$, $x^\star \sim \mu^\star$, $y^\star \sim \nu^\star$, $(x^\star, y^\star) \sim \pi^\star$ and $y^\star = \nabla G(x^\star)$ a.s.

3.2 Lagrangian function and duality gap

We introduce the following Lagrangian function defined for every $\mu \in \mathcal{P}_2(\mathsf{X})$ and $y \in L^2$ by

$$\mathscr{L}(\mu, y) := \mathcal{E}_F(\mu) + \mathcal{H}(\mu) - \mathcal{E}_{G^*}(\nu) + \mathbb{E}\langle x, y\rangle, \tag{19}$$

where $x = T^\mu_{\mu^\star}(x^\star)$. This Lagrangian is similar to the one used in Euclidean optimization; see the appendix. We also define the duality gap by

$$\mathscr{D}(\mu, y) := \mathscr{L}(\mu, y^\star) - \mathscr{L}(\mu^\star, y). \tag{20}$$

The next theorem, which is of independent interest, can be interpreted as a strong duality result for the Lagrangian function $\mathscr{L}$, see [24, Lemma 36.2].

Theorem 3 (Strong duality)**.** Let Assumptions 1–4 hold true. Then, for every $\mu \in \mathcal{P}_2(\mathsf{X}), y \in L^2$, $\mathscr{D}(\mu, y) \geq 0$ and $\mathscr{L}(\mu, y) \leq \mathcal{F}(\mu)$. Moreover, $(\mu^\star, y^\star)$ is a saddle point of $\mathscr{L}$ with saddle value $\mathcal{F}(\mu^\star)$, *i.e.*,

$$\mathscr{L}(\mu^\star, y) \leq \mathcal{F}(\mu^\star) = \mathscr{L}(\mu^\star, y^\star) \leq \mathscr{L}(\mu, y^\star). \tag{21}$$

Finally, $\mathscr{L}(\mu^\star, y) = \mathcal{F}(\mu^\star)$ if and only if $y = y^\star$, and, if F is strictly convex, $\mathcal{F}(\mu^\star) = \mathscr{L}(\mu, y^\star)$ if and only if $\mu = \mu^\star$.

The proof of Theorem 3 relies on using (18) to write the duality gap as the sum of the Bregman divergences of F, G^* and $\mathcal{H}$. We shall use the nonnegativity of the duality gap to derive convergence bounds for PSGLA.

4 Forward Backward representation of PSGLA

In this section, we present our viewpoint on PSGLA (3). More precisely, we represent PSGLA as a (stochastic) Forward Backward algorithm involving (stochastic) monotone operators which are not necessarily subdifferentials.

Intuition. Let $\pi \in \mathcal{P}_2(\mathsf{X}^2)$ and consider $A, B(\pi) \in L^2(\pi; \mathsf{X}^2)$ the set valued maps

$$A : (x,y) \mapsto \begin{bmatrix} y \\ -x & +\partial G^*(y) \end{bmatrix}, \quad B(\pi) : (x,y) \mapsto \begin{bmatrix} \nabla F(x) + \boldsymbol{\partial}\mathcal{H}(\mu)(x) \\ 0 \end{bmatrix}, \tag{22}$$

where $\mu = x^\star \# \pi$. The maps $\pi \mapsto A$ and $\pi \mapsto B(\pi)$ satisfy a monotonicity property similar to (13) (note that A is a maximal monotone operator as the sum of $S : (x,y) \mapsto (y, -x)$ and the subdifferential of the $\Gamma_0(\mathsf{X}^2)$ function $(x,y) \mapsto G^*(y)$). Inclusion (18) can be rewritten as

$$0 \in (A + B(\pi^\star))(x,y), \text{ for } \pi^\star \text{ a.e. } (x,y). \tag{23}$$

Rigorous Forward Backward representation. The "monotone" inclusion (23) intuitively suggests the following stochastic Forward Backward algorithm for obtaining samples from $\pi^\star$ (and hence from $\mu^\star$ by marginalizing):

$$P\begin{bmatrix} x^{k+1/2}-x^k \\ y^{k+1/2}-y^k \end{bmatrix} = -\gamma \begin{bmatrix} \nabla f(x^k,\xi^{k+1}) - \sqrt{\frac{2}{\gamma}} W^{k+1} \\ 0 \end{bmatrix} \tag{24}$$

$$P\begin{bmatrix} x^{k+1}-x^{k+1/2} \\ y^{k+1}-y^{k+1/2} \end{bmatrix} \in -\gamma A(x^{k+1},y^{k+1}). \tag{25}$$

Above, $P \in \mathbb{R}^{d\times d}_{++}$ is an appropriately chosen matrix. Indeed, Algorithm (24)-(25) looks like a stochastic Forward-Backward algorithm [6, 7, 15, 26] where the gradient is perturbed by a Gaussian vector, as in the Langevin algorithm (1). In Algorithm (24)-(25), we cannot set P to be the identity map of X^2 because the inclusion (25) is intractable in this case. We take $P : (x,y) \mapsto x$, *i.e.*, with our notations, $P = x^\star$. Although the matrix P is only semi-definite positive, the next lemma shows that Algorithm (24)-(25) is still well defined. More precisely, the next lemma shows that $x^{k+1} = \mathrm{prox}_{\gamma G}(x^{k+1/2})$ (by taking $z = (x^{k+1/2}, y^{k+1/2})$ in the lemma) and hence the resulting algorithm (24)-(25) is PSGLA. Based on the representation (24)-(25) of PSGLA, the next lemma also provides an important inequality used later in the proof of Theorem 5.

Lemma 4. Let $z = (x,y), z' = (x',y') \in \mathsf{X}^2$. Then $P(z'-z) \in -\gamma A(z')$ if and only if $x' = \mathrm{prox}_{\gamma G}(x)$ and $y' = \mathrm{prox}_{G^*/\gamma}(x/\gamma)$. Moreover, if $G \in \Gamma_0(\mathsf{X})$ is $1/\lambda_{G^*}$-smooth, then

$$\begin{aligned}\|x'-x^\star\|^2 \leq & \|x-x^\star\|^2 - 2\gamma\left(G^*(y') - G^*(y^\star) - \langle y', x^\star\rangle + \langle y^\star, x\rangle\right) \\ & - \gamma(\lambda_{G^*}+\gamma)\|y'-y^\star\|^2 + \gamma^2\|y^\star\|^2.\end{aligned} \tag{26}$$

5 Main results

We now provide our main result on PSGLA (3). For $r \in \mathbb{N}/2$, denote μ^r (resp. ν^r) the distribution of x^r (resp. y^r), defined in the previous section.

Theorem 5. Let Assumptions 1, 2, 3 and 5 hold true. If F is λ_F-strongly convex and G is $1/\lambda_{G^*}$-smooth, then for every $\gamma \leq 1/L$,

$$\begin{aligned}W^2(\mu^{k+1},\mu^\star) \leq & (1-\gamma\lambda_F)W^2(\mu^k,\mu^\star) - \gamma(\lambda_{G^*}+\gamma)W^2(\nu^{k+1},\nu^\star) \\ & - 2\gamma\left(\mathscr{L}(\mu^{k+1/2},y^\star) - \mathscr{L}(\mu^\star, y_\star^{k+1})\right) + \gamma^2 C,\end{aligned} \tag{27}$$

where $C := \int_{\mathrm{int}(D)} \|\nabla G(x)\|^2 d\mu^\star(x) + 2(Ld+\sigma_F^2)$ and $y_\star^{k+1} := \mathrm{prox}_{G^*/\gamma}(x_\star^{k+1/2}/\gamma) \sim \nu^{k+1}$, where $x_\star^{k+1/2} := T_{\mu^\star}^{\mu^{k+1/2}}(x^\star)$.

The proof of Theorem 5 relies on using Lemma 4 along with [19, Lemma 30]. Inspecting the proof of Theorem 5, one can see that any $\bar\mu, \bar y$ can replace $\mu^\star, y^\star$[5]. The situation is similar to primal dual algorithms in optimization [11, 18] and Evolution Variational Inequalities in optimal transport [1].

The next corollary is obtained by using $\mathscr{D}(\mu^{k+1/2}, y_\star^{k+1}) \geq 0$ (Theorem 3) and iterating (27).

Corollary 6. Let Assumptions 1–5 hold true. If $\gamma \leq 1/L$, then

$$\min_{j\in\{0,\dots,k-1\}} \mathscr{D}(\mu^{j+1/2}, y_\star^{j+1}) \leq \tfrac{1}{2\gamma k} W^2(\mu^0,\mu^\star) + \tfrac{\gamma}{2}C, \tag{28}$$

$$\min_{j\in\{1,\dots,k\}} W^2(\nu^j,\nu^\star) \leq \tfrac{1}{\gamma(\lambda_{G^*}+\gamma)k} W^2(\mu^0,\mu^\star) + \tfrac{\gamma}{\lambda_{G^*}+\gamma}C. \tag{29}$$

Finally, if $\lambda_F > 0$, then

$$W^2(\mu^k,\mu^\star) \leq (1-\gamma\lambda_F)^k W^2(\mu^0,\mu^\star) + \tfrac{\gamma}{\lambda_F}C. \tag{30}$$

If G is Lipschitz continuous (in particular if $G \equiv 0$), then our Assumptions hold true. Moreover, inequality (28) recovers [19, Corollary 18] but with the duality gap instead of the KL

[5]The proof does not rely on specific properties of the latter like being primal dual optimal.

divergence. Obtaining a result in terms of KL divergence is hopeless for PSGLA in general because the KL divergence is infinite; see the appendix. Connecting the convergence of the duality gap to zero to known modes of convergence is left for future work. Besides, obtaining an inequality like (29) that holds when F is just convex is rather not standard in the literature on Langevin algorithm, see [25, 35]. Corollary 6 implies the following complexity results. Given $\varepsilon > 0$, choosing $\gamma = \min(1/L, \varepsilon/C)$ and $k \geq \max(L/\varepsilon, C/\varepsilon^2) W^2(\mu^0, \mu^\star)$ in inequality (28) leads to $\min_{j\in\{0,\dots,k-1\}} \mathscr{D}(\mu^{j+1/2}, y_\star^{j+1}) \leq \varepsilon$. If $\lambda_{G^*} > 0$ (*i.e.*, if G is smooth), choosing $\gamma = \min(1/L, \frac{\lambda_{G^*}\varepsilon}{2C})$ and $k \geq \max(\frac{2L}{\lambda_{G^*}\varepsilon}, \frac{4C}{\lambda_{G^*}^2\varepsilon^2}) W^2(\mu^0, \mu^\star)$ in inequality (29) leads to $\min_{j\in\{1,\dots,k\}} W^2(\nu^j, \nu^\star) \leq \varepsilon$. Finally, if $\lambda_F > 0$ (*i.e.*, if F is strongly convex), choosing $\gamma = \min(1/L, \frac{\lambda_F\varepsilon}{2C})$ and $k \geq \frac{1}{\gamma\lambda_F} \log(2W^2(\mu^0, \mu^\star)/\varepsilon)$ *i.e.*,

$$k \geq \max\left(\frac{L}{\lambda_F}, \frac{2C}{\lambda_F^2\varepsilon}\right) \log\left(\frac{2W^2(\mu^0,\mu^\star)}{\varepsilon}\right), \quad C = \int_{\mathrm{int}(D)} \|\nabla G(x)\|^2 d\mu^\star(x) + 2(Ld + \sigma_F^2) \tag{31}$$

in inequality (30), leads to $W^2(\mu^k, \mu^\star) \leq \varepsilon$. [6] In the case where G is M-Lipschitz continuous, the complexity (31) improves [19, Corollary 22] since $\int_{\mathrm{int}(D)} \|\nabla G(x)\|^2 d\mu^\star(x) \leq M^2$.

6 Conclusion

We made a step towards theoretical understanding the properties of the Langevin algorithm in the case where the target distribution is not smooth and not fully supported. This case is known to be difficult to analyze and has many applications [9, 10, 21]. Our analysis improves and extends the state of the art.

Moreover, our approach is new. We developed a primal dual theory for a minimization problem over the Wasserstein space, which is of independent interest. A broader duality theory for minimization problems in the Wasserstein space would be of practical and theoretical interest.

7 Acknowledgement

We thank Laurent Condat for introducing us to the primal dual view of the proximal gradient algorithm in Hilbert spaces.

8 Broader impact

Our work contributes to the understanding of a sampling algorithm used in statistics. Our main results are of theoretical nature (convergence rates). Therefore, we do not see any immediate societal impact of our results.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[2] Y. F Atchadé. A Moreau-Yosida approximation scheme for a class of high-dimensional posterior distributions. *arXiv preprint arXiv:1505.07072*, 2015.

[3] Y. F Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(1):310–342, 2017.

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

[5] E. Bernton. Langevin Monte Carlo and JKO splitting. In *Conference on Learning Theory*, pages 1777–1798, 2018.

[6] P. Bianchi and W. Hachem. Dynamical behavior of a stochastic Forward-Backward algorithm using random monotone operators. *Journal of Optimization Theory and Applications*, 171(1):90–120, 2016.

[6] The dependence in d of the factor $W^2(\mu^0, \mu^\star)$ can be explicited under further assumptions, see [19, 25].

[7] P. Bianchi, W. Hachem, and A. Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *Journal of Convex Analysis*, 26(2):397–436, 2019.

[8] S. Brazitikos, A. Giannopoulos, P. Valettas, and B.-H. Vritsiou. *Geometry of isotropic convex bodies*, volume 196. American Mathematical Soc., 2014.

[9] N. Brosse, A. Durmus, E. Moulines, and M. Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Conference on Learning Theory*, pages 319–342, 2017.

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

[11] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[12] N. S Chatterji, J. Diakonikolas, M. I Jordan, and P. L Bartlett. Langevin monte carlo without smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 1716–1726, 2020.

[13] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs. A variational formulation for frame-based inverse problems. *Inverse Problems*, 23(4):1495–1518, June 2007.

[14] X. Cheng and P. L Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211, 2018.

[15] P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure and Applied Functional Analysis*, 1(1):13–37, 2016.

[16] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi. Proximal splitting algorithms: Overrelax them all! *arXiv preprint arXiv:1912.00137*, 2019.

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

[18] Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

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

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

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

[22] Y.-P. Hsieh, A. Kavis, P. Rolland, and V. Cevher. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887, 2018.

[23] Y.-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. L Bartlett, and M. I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

[24] R. T Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1970.

[25] P. Rolland, A. Eftekhari, A. Kavis, and V. Cevher. Double-loop unadjusted langevin algorithm. In *International Conference on Machine Learning*, pages 4326–4334, 2020.

[26] L. Rosasco, S. Villa, and B. C Vũ. Stochastic forward–backward splitting for monotone inclusions. *Journal of Optimization Theory and Applications*, 169(2):388–406, 2016.

[27] A. Salim, D. Kovalev, and P. Richtárik. Stochastic proximal langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems*, pages 6649–6661, 2019.

[28] S. Schechtman, A. Salim, and P. Bianchi. Passty Langevin. In *Conference on Machine Learning (CAp)*, 2019.

[29] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[30] P. Toulis, T. Horel, and E. M Airoldi. The proximal Robbins-Monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020. to appear.

[31] S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8092–8104, 2019.

[32] M. Welling and Y. W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

[33] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, page 2093–3027, 2018.

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

[35] D. Zou, P. Xu, and Q. Gu. Stochastic variance-reduced Hamilton Monte Carlo methods. In *International Conference on Machine Learning*, pages 6028–6037, 2018.

Appendix

Contents

A Numerical experiments

In this section, we illustrate our results on PSGLA through numerical experiments.

Sampling *a posteriori*. We consider a statistical framework where i.i.d. random vectors (data) $D_1, \dots, D_n$ with distribution $\mathbb{P}_{x^\star}$ are observed. We adopt a Bayesian strategy where we assume the distribution $\mathbb{P}_{x^\star}$ to be indexed by a random vector $x^\star$ with values in X. Denote $\mathcal{L}(\cdot, x^\star)$ the density of $\mathbb{P}_{x^\star}$ (a.k.a. the likelihood function) w.r.t. some reference measure. Given a prior distribution for $x^\star$ with density π w.r.t. Leb, our goal is construct samples $x^1, \dots, x^k$ from the posterior distribution

$$\mu^\star(x|D_1, \dots, D_n) \propto \pi(x) \prod_{i=1}^{n} \mathcal{L}(D_i, x), \tag{32}$$

in order *e.g.* to estimate the mean *a posteriori* via Monte Carlo approximations,

$$m^\star := \int x \mu^\star(x|D_1, \dots, D_n) d\,\mathrm{Leb}(x) \simeq \frac{1}{k} \sum_{j=1}^{k} x^j. \tag{33}$$

Wishart distribution. In the experiments, the Euclidean space X is a space of $d \times d$ symmetric matrices and π is the Wishart distribution defined by

$$\pi(x) \propto |\det(x)|^{\frac{\nu-d-1}{2}} \exp\left(-\frac{\mathrm{tr}(V^{-1}x)}{2}\right) \mathbf{1}_{\mathbb{R}^{d\times d}_{++}}(x), \tag{34}$$

where $\nu > d-1$ and $V \in \mathbb{R}^{d\times d}_{++}$ are parameters of the distribution. Note that $\pi(x) = 0$ if x is not a positive definite matrix. The mean of the Wishart distribution is equal to νV. The Wishart distribution is widely used in Random matrix theory and applications, see [29]. Indeed, the Wishart distribution is a conjugate prior to the Gaussian likelihood. More precisely, assume that for every $D \in \mathbb{R}^d$ and $x \in \mathbb{R}^{d\times d}_{++}$,

$$\mathcal{L}(D, x) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{1}{2} D^T x D\right) \sqrt{\det(x)}, \tag{35}$$

is the density of a centered Gaussian distribution with precision matrix (*i.e.*, inverse variance-covariance matrix) x. Then, if π is Wishart with parameters ν and V (*i.e.*, π is given by (34)), then the posterior distribution $\mu^\star(\cdot|D_1, \dots, D_n)$ (32) is Wishart with parameters $\nu' = n + \nu$ and $V' = \left(I + \sum_{i=1}^n D_i D_i^T\right)^{-1}$:

$$\mu^\star(x|D_1, \dots, D_n) \propto |\det(x)|^{\frac{(\nu+n)-d-1}{2}} \exp\left(-\frac{\mathrm{tr}\left((V^{-1} + \sum_{i=1}^n D_i D_i^T)x\right)}{2}\right) \mathbf{1}_{\mathbb{R}^{d\times d}_{++}}(x). \tag{36}$$

Moreover, the mean of the posterior distribution is equal to

$$m^\star = (n+\nu) \left(I + \sum_{i=1}^{n} D_i D_i^T\right)^{-1}. \tag{37}$$

Setup. We consider two *a posterori* sampling problems.

First, we consider the task of learning the mean of the data. More precisely, $\mathbb{P}_{x^\star}$ is a Gaussian distribution over $\mathbb{R}$ with mean $x^\star$ and unit variance. We use the Wishart distribution π with $V = I$ as the prior distribution. Note that in this one dimensional case, the Wishart distribution π boils down to a Gamma distribution over $\mathbb{R}$.

In other words, $\pi(x) \propto \exp(-G(x))$ and $\mu^\star(x|D_1, \dots, D_n) \propto \exp(-G(x) - \sum_{i=1}^n f_i(x))$ where the functions $f_i, G \in \Gamma_0(\mathsf{X})$ are defined by

$$G(x) := -\frac{\nu-d-1}{2} \log|x| + \frac{x}{2} + \iota_{(0,+\infty)}(x),$$
$$f_i(x) := \frac{|x - D_i|^2}{2},$$

for every $i \in \{1, \dots, n\}$. The data points D_i are generated randomly using a Gaussian distribution. Note that f_i is smooth and strongly convex. Moreover, G is nonsmooth and the proximity operator of G has a closed form thanks to recent results.[7] We consider $d = 1$ in order to be able to represent the numerical results with histograms, see Figures 1-2.

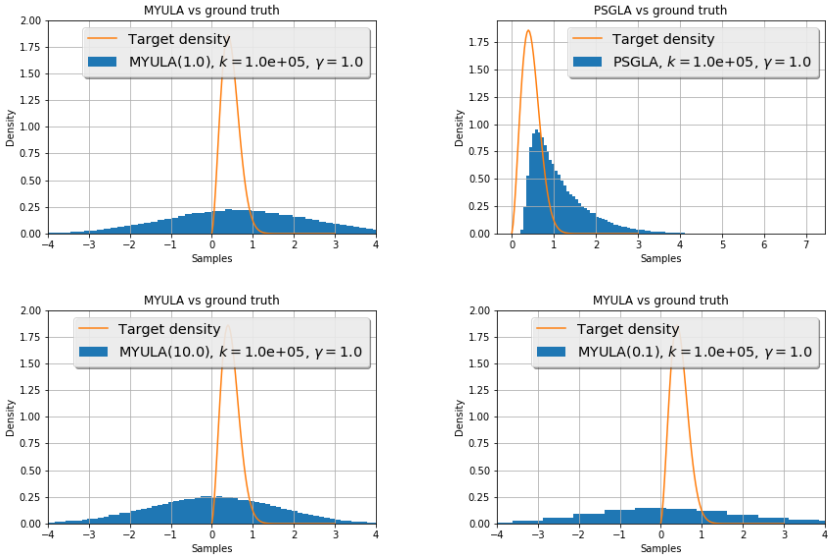

Figure 1: Histograms drawn by the k iterates of PSGLA and MYULA(λ), for various values of λ, compared to the target distribution $\mu^\star$. Case $d = 1, \gamma = 1.0$.

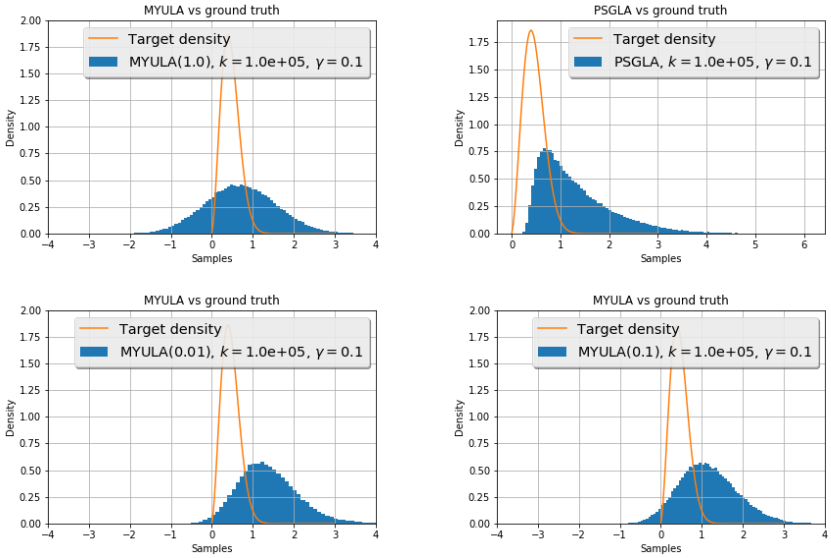

Figure 2: Histograms drawn by the k iterates of PSGLA and MYULA(λ), for various values of λ, compared to the target distribution $\mu^\star$. Case $d = 1, \gamma = 0.1$.

Then, we consider the task of learning the precision matrix of the data. More precisely, $\mathbb{P}_{x^\star}$ is a centered Gaussian distribution over $\mathbb{R}^d$ with precision matrix $x^\star$. We use the Wishart distribution π with $V = I$ as the prior distribution. Since the prior distribution π (34) is conjugate to the likelihood function $\mathcal{L}(\cdot, x^\star)$ (35), the posterior distribution $\mu^\star$ is given by (36). Thus, we can use $\mu^\star$ as a ground

[7] see `www.proximity-operator.net`

truth. In other words, $\mu^\star(x|D_1,\dots,D_n) \propto \exp(-G(x)-F(x))$ where the functions $F, G \in \Gamma_0(\mathsf{X})$ are defined by

$$G(x) := -\frac{(\nu+n)-d-1}{2}\log|\det(x)| + \frac{\mathrm{tr}(x)}{2} + \iota_{\mathbb{R}^{d\times d}_{++}}(x),$$
$$F(x) := \sum_{i=1}^{n} \frac{\mathrm{tr}(D_i D_i^T x)}{2}.$$

The data points D_i are generated randomly using a Gaussian distribution. Note that F is smooth and convex, hence Assumptions 1-2 are satisfied. Moreover, Assumptions 3-4 are satisfied *e.g.* if $n+\nu > d+3$. Finally, G is nonsmooth and the proximity operator of G has a closed form [4, Corollary 24.65]. We consider several values of d: $d=1, d=10$ and $d=100$. The number of entries of the iterates is d^2 and, since the matrices are symmetric, the dimension of the sampling problem is slightly larger than $d^2/2$. Since the mean of $\mu^\star$ is known, we use it as a ground truth and perform a mean *a posteriori* estimation using the estimators (33) constructed by PSGLA and MYULA. The convergence of the estimators is illustrated in Figures 3 and 4 for the cases $d=10$ and $d=100$. Moreover, in order to visualize better the multidimensional results of Figures 3 and 4, we consider the *same* sampling problem with $d=1$ and plot histograms to represent the results, see Figure 5.

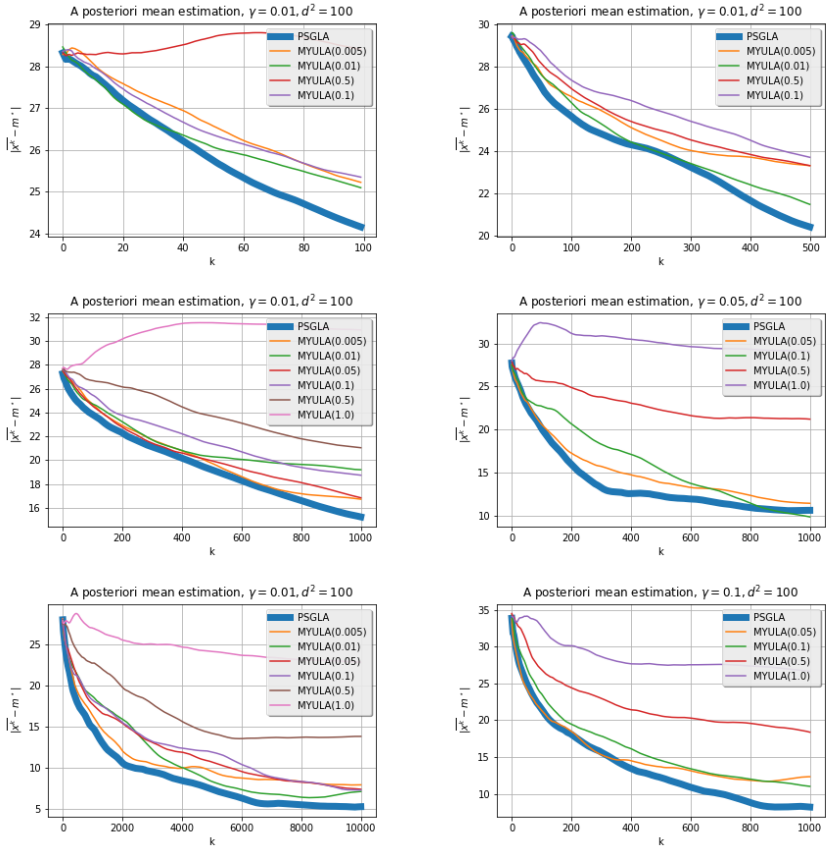

Figure 3: Frobenius distance between $m^\star$ and the mean *a posteriori* estimators (ergodic means) constructed by PSGLA and MYULA(λ), for various values of λ, γ as a function of k in the case $d=10$.

Algorithms. We compare PSGLA to various versions of MYULA [21], parametrized by the smoothing parameter $\lambda > 0$. We use the same learning rate γ for the algorithms[8]. We denote these algorithms MYULA(λ). Both MYULA and PSGLA compute one proximity operator $\mathrm{prox}_{\gamma G}$ and

[8] In general, a Langevin algorithm becomes more precise and slower as $\gamma \to 0$.

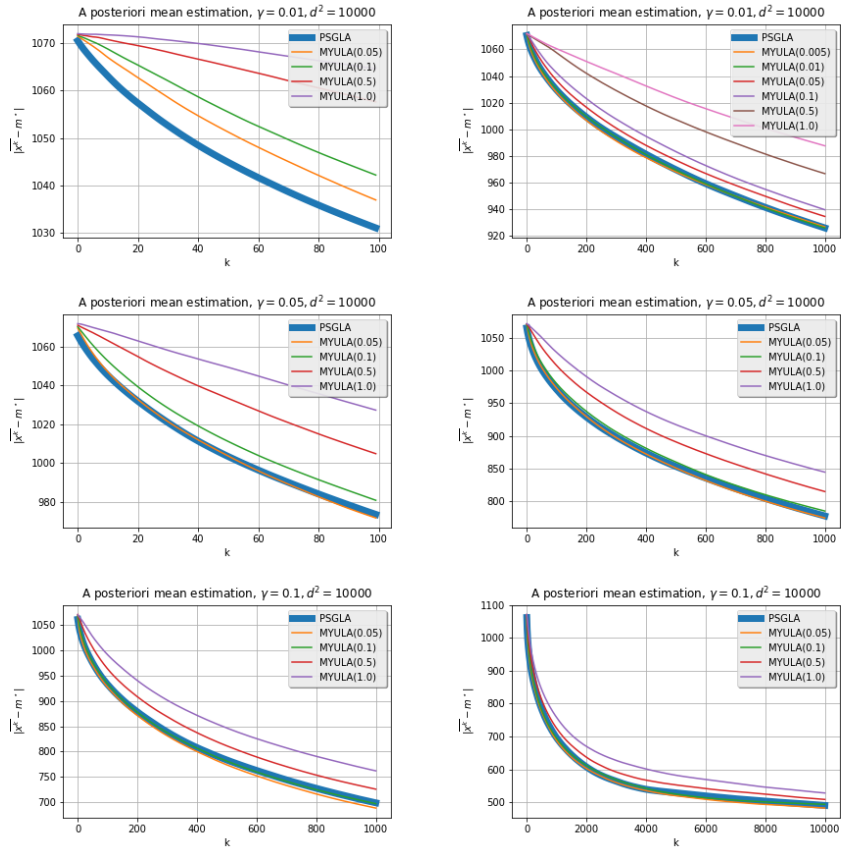

Figure 4: Frobenius distance between $m^\star$ and the mean *a posteriori* estimators (ergodic means) constructed by PSGLA and MYULA(λ), for various values of λ, γ as a function of k in the case $d = 100$.

one gradient per iteration[9]. They both require to sample one Gaussian random variable over the space of symmetric matrices at each iteration.

Observations. Using Langevin algorithm to sample from distributions which are not fully supported is known to be a difficult task [9, 10, 21]. In Figures 1-2 and 5, we see that the shape of the histograms drawn by PSGLA are closer to the target distribution than the shape of the benchmarks histograms. This means that PSGLA converges faster than the benchmarks for this sampling task. This behavior was expected: PSGLA does not introduce extra bias by introducing a smoothing parameter λ.

More importantly, we see that the iterates of PSGLA are always feasible i.e., they lie in the support of the target distribution, contrary to the benchmarks.

Finally, PSGLA is a proximal method, whereas the benchmarks are instances of the standard Langevin algorithm (applied to the smoothed problem depending on λ). In stochastic optimization, proximal methods are known to be more stable than gradient methods [30]. This phenomenon is observed here. For instance, the range of step sizes allowed by the benchmarks is controlled by the smoothing parameter λ, see [21]. Using a step size too large for the benchmarks leads to a numerical instability that does not occur for PSGLA.

Figures 3 and 4 are multidimensional extensions of Figure 5. Each figure in 3 and 4 corresponds to a new run. We plotted the convergence of the mean *a posteriori* estimators (*i.e.*, ergodic means) constructed by MYULA(λ) and PSGLA.

[9]We use a slight extension of MYULA allowing to use a stochastic gradient.

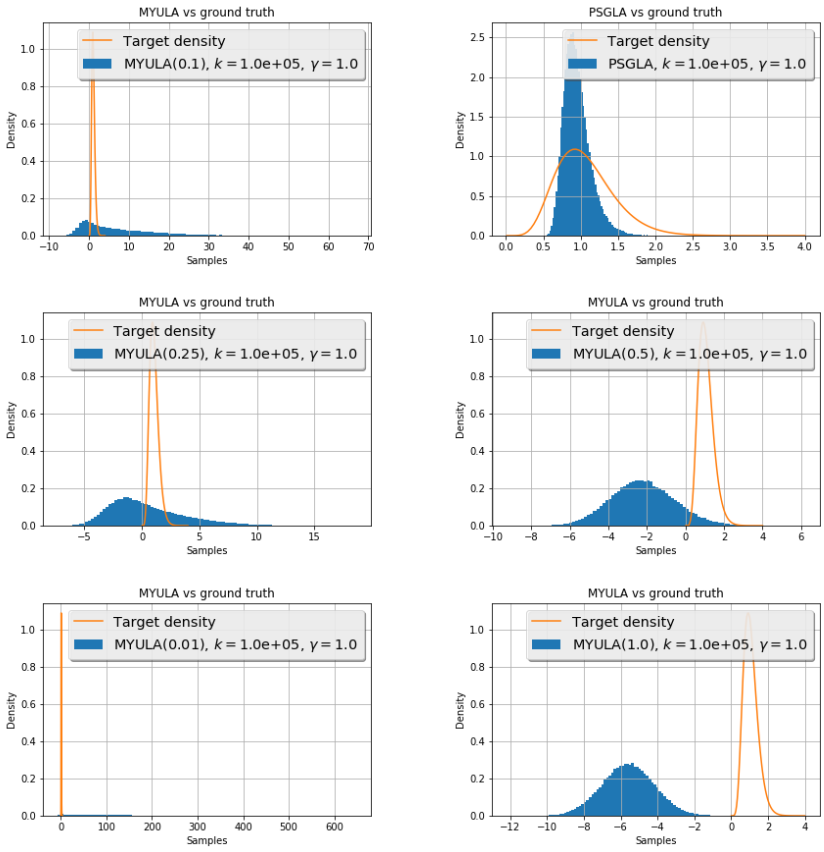

Figure 5: Histograms drawn by the k iterates of PSGLA and MYULA(λ), for various values of λ, compared to the target distribution $\mu^\star$. Case $d = 1$, $\gamma = 1.0$.

In general, we see that PSGLA is as good as the MYULA(λ) for the best value of λ, while conserving the feasibility of the iterates, and without having to select the value of λ.

B Postponed proofs

B.1 Proof of Lemma 1

Using [8, Lemma 2.2.1], there exist $A, B > 0$, $\exp(-V(x)) \leq A\exp(-B\|x\|) \leq A$. The last inequality implies $\int \|x\|^2 \exp(-V(x))dx < \infty$. Moreover, $-V(x) \leq \log(A) \leq C := \max(0, \log(A))$. Using that $u \mapsto u\exp(-u)$ is nonincreasing on $[1, +\infty)$.

$$\begin{aligned} V(x)\exp(-V(x)) &= V(x)\exp(-V(x))\mathbf{1}_{V(x)\leq\|x\|^2+1} + V(x)\exp(-V(x))\mathbf{1}_{V(x)>\|x\|^2+1} \\ &\leq \left(\|x\|^2+1\right)\exp(-V(x)) + \left(\|x\|^2+1\right)\exp\left(-(\|x\|^2+1)\right). \end{aligned}$$

Using $|V(x)| = V(x)\mathbf{1}_{V(x)\geq 0} - V(x)\mathbf{1}_{V(x)<0} \leq C + V(x)$,

$$|V(x)|\exp(-V(x)) \leq \left(\|x\|^2+1+C\right)A\exp(-B\|x\|) + \left(\|x\|^2+1\right)\exp\left(-(\|x\|^2+1)\right).$$

We conclude using that the r.h.s. is integrable.

B.2 Proof of Theorem 3

The proof is divided in six parts, each part proving one claim. Denote $x = T^{\mu}_{\mu^\star}(x^\star)$.

Part I. First,

$$\mathscr{L}(\mu, y^\star) = \mathcal{E}_F(\mu) + \mathcal{H}(\mu) - \mathcal{E}_{G^*}(\nu^\star) + \mathbb{E}\langle x, y^\star\rangle, \tag{38}$$

and

$$\mathscr{L}(\mu^\star, y) = \mathcal{E}_F(\mu^\star) + \mathcal{H}(\mu^\star) - \mathcal{E}_{G^*}(\nu) + \mathbb{E}\langle x^\star, y\rangle. \tag{39}$$

Therefore, the duality gap can be rewritten

$$\begin{aligned}\mathscr{D}(\mu, y) =&\mathcal{E}_F(\mu) - \mathcal{E}_F(\mu^\star) + \mathcal{H}(\mu) - \mathcal{H}(\mu^\star) + \mathcal{E}_{G^*}(\nu) - \mathcal{E}_{G^*}(\nu^\star)\\ &+ \mathbb{E}\langle x, y^\star\rangle - \mathbb{E}\langle x^\star, y\rangle\\ =&\mathcal{E}_F(\mu) - \mathcal{E}_F(\mu^\star) + \mathcal{H}(\mu) - \mathcal{H}(\mu^\star) + \mathcal{E}_{G^*}(\nu) - \mathcal{E}_{G^*}(\nu^\star)\\ &+ \mathbb{E}\langle x - x^\star, y^\star\rangle - \mathbb{E}\langle x^\star, y - y^\star\rangle.\end{aligned}$$

Using (18), $y^\star = -\nabla F(x^\star) - \boldsymbol{\partial}^0\mathcal{H}(\mu^\star)(x^\star)$ and $x^\star \in \partial G^*(y^\star)$.

$$\begin{aligned}\mathscr{D}(\mu, y) =&\mathcal{E}_F(\mu) - \mathcal{E}_F(\mu^\star) - \mathbb{E}\langle\nabla F(x^\star), x - x^\star\rangle\\ &+ \mathcal{H}(\mu) - \mathcal{H}(\mu^\star) - \mathbb{E}\langle\boldsymbol{\partial}^0\mathcal{H}(\mu^\star)(x^\star), x - x^\star\rangle\\ &+ \mathcal{E}_{G^*}(\nu) - \mathcal{E}_{G^*}(\nu^\star) - \mathbb{E}\langle x^\star, y - y^\star\rangle\\ =&\mathbb{E}F(x) - \mathbb{E}F(x^\star) - \mathbb{E}\langle\nabla F(x^\star), x - x^\star\rangle\\ &+ \mathcal{H}(\mu) - \mathcal{H}(\mu^\star) - \langle\boldsymbol{\partial}^0\mathcal{H}(\mu^\star), T^\mu_{\mu^\star} - I\rangle_{\mu^\star}\\ &+ \mathbb{E}G^*(y) - \mathbb{E}G^*(y^\star) - \mathbb{E}\langle x^\star, y - y^\star\rangle,\end{aligned} \tag{40}$$

where the last equality comes from the transfer theorem. We get $\mathscr{D}(\mu, y) \geq 0$ using the convexity of F, the convexity of G^* and the geodesic convexity of $\mathcal{H}$ (inequality (12)).

Part II. Since $\mathscr{D}(\mu, y^\star) \geq 0$ and $\mathscr{D}(\mu^\star, y) \geq 0$,

$$\mathscr{L}(\mu^\star, y) \leq \mathscr{L}(\mu^\star, y^\star) \leq \mathscr{L}(\mu, y^\star). \tag{41}$$

Part III. Then,

$$\begin{aligned}\mathscr{L}(\mu, y) &= \mathcal{E}_F(\mu) + \mathcal{H}(\mu) + \mathbb{E}\left(\langle x, y\rangle - G^*(y)\right) &(42)\\ &\leq \mathcal{E}_F(\mu) + \mathcal{H}(\mu) + \mathbb{E}\left(\sup\langle x, \cdot\rangle - G^*\right) &(43)\\ &\leq \mathcal{E}_F(\mu) + \mathcal{H}(\mu) + \mathbb{E}G(x) &(44)\\ &= \mathcal{F}(\mu),\end{aligned}$$

using [4, Proposition 13.15].

Part IV. Using [4, Proposition 16.10], $\sup\langle x, \cdot\rangle - G^* = G(x)$, and $\langle x, y\rangle - G^*(y) = G(x)$ if and only if $y \in \partial G(x)$ (or $x \in \partial G^*(y)$). Taking $\mu = \mu^\star$ and $y = y^\star$ in (42), we have $x^\star = x \in \partial G^*(y^\star)$ and therefore $\mathscr{L}(\mu^\star, y^\star) = \mathcal{F}(\mu^\star)$.

Part V. Assume that $\mathscr{L}(\mu^\star, \bar{y}) = \mathcal{F}(\mu^\star)$. We shall prove that $\bar{y} = y^\star$ a.s. Since $\mathcal{F}(\mu^\star) \leq \mathscr{L}(\mu^\star, \bar{y})$, inequality (43) becomes an equality when $\mu = \mu^\star$ and $y = \bar{y}$. Therefore, $\langle x^\star, \bar{y}\rangle - G^*(\bar{y}) = \sup\langle x^\star, \cdot\rangle - G^* = G(x^\star)$ a.s., which implies $\bar{y} \in \partial G(x^\star) = \{\nabla G(x^\star)\} = \{y^\star\}$ a.s.

Part VI. Assume that $\mathscr{L}(\mu^\star, y^\star) = \mathscr{L}(\bar{\mu}, y^\star)$ and that F is strictly convex. We shall prove that $\mu^\star = \bar{\mu}$. We have $\mathscr{D}(\bar{\mu}, y^\star) = 0$, and, using (40),

$$\mathscr{D}(\bar{\mu}, y^\star) = \mathbb{E}F(\bar{x}) - \mathbb{E}F(x^\star) - \mathbb{E}\langle\nabla F(x^\star), \bar{x} - x^\star\rangle + \mathcal{H}(\bar{\mu}) - \mathcal{H}(\mu^\star) - \langle\boldsymbol{\partial}^0\mathcal{H}(\mu^\star), T^{\bar{\mu}}_{\mu^\star} - I\rangle_{\mu^\star}, \tag{45}$$

where $\bar{x} = T^{\bar{\mu}}_{\mu^\star}(x^\star)$. Using the convexity of F and the geodesic convexity of $\mathcal{H}$, $F(\bar{x}) - F(x^\star) - \langle\nabla F(x^\star), \bar{x} - x^\star\rangle = 0$ a.s. Using the strict convexity of F, $\bar{x} = x^\star$ a.s., therefore $\bar{\mu} = \mu^\star$.

B.3 Proof of Lemma 4

If $x' = x - \gamma y'$ and $0 \in -x' + \partial G^*(y')$, we have $y' \in \partial G(x')$ and $x' = \operatorname{prox}_{\gamma G}(x)$. Moreover, $0 \in -x + \gamma y' + \partial G^*(y')$ implies $y' = \operatorname{prox}_{G^*/\gamma}(x/\gamma)$. One can easily check that this is an

equivalence. Denote $\|\cdot\|_P$ the semi-norm induced by P on X^2 defined by $\|z\|_P = \|x\|$ for every $z = (x, y) \in \mathsf{X}^2$, and $\langle \cdot, \cdot \rangle_P$ the semi-inner product associated. We have

$$\|z' - z^\star\|_P^2 = \|z - z^\star\|_P^2 + 2\langle z' - z, z' - z^\star \rangle_P - \|z' - z\|_P^2. \tag{46}$$

We now identify the terms. First, $\|z' - z^\star\|_P^2 = \|x' - x^\star\|^2$, $\|z - z^\star\|_P^2 = \|x - x^\star\|^2$ and $\|z' - z\|_P^2 = \|x' - x\|^2 = \gamma^2\|y'\|^2$. Second, using $P(z' - z) \in -\gamma A(z')$, and the definition of A, there exists $g^* \in G^*(y')$ such that

$$\langle z' - z, z' - z^\star \rangle_P = \langle P(z' - z), z' - z^\star \rangle = -\gamma\langle y', x' - x^\star \rangle + \gamma\langle x', y' - y^\star \rangle - \gamma\langle g^*, y' - y^\star \rangle. \tag{47}$$

Hence

$$\langle z' - z, z' - z^\star \rangle_P \leq \gamma\langle y', x^\star \rangle - \gamma\langle x', y^\star \rangle - \gamma\left(G^*(y') - G^*(y^\star) + \frac{\lambda_{G^*}}{2}\|y' - y^\star\|^2 \right), \tag{48}$$

using the strong convexity of G^*. Plugging into (46),

$$\|x' - x^\star\|^2 \leq \|x - x^\star\|^2 - 2\gamma\left(G^*(y') - G^*(y^\star) + \frac{\lambda_{G^*}}{2}\|y' - y^\star\|^2 - \langle y', x^\star \rangle + \langle y^\star, x' \rangle \right) - \gamma^2\|y'\|^2. \tag{49}$$

Using $x' = x - \gamma y'$,

$$-2\gamma\langle y^\star, x' \rangle - \gamma^2\|y'\|^2 = -2\gamma\langle y^\star, x \rangle - 2\gamma\langle y^\star, -\gamma y' \rangle - \gamma^2\|y'\|^2 = -2\gamma\langle y^\star, x \rangle - \gamma^2\|y' - y^\star\|^2 + \gamma^2\|y^\star\|^2. \tag{50}$$

Plugging into (49) concludes the proof.

B.4 Proof of Theorem 5

We first recall a standard inequality of the stochastic gradient Langevin algorithm, see *e.g.* [19, Lemma 30].

Lemma 7 ([19]). Let Assumptions 1, 2 and 5 hold true. Then, if F is λ_F-strongly convex, for every $\gamma \leq 1/L$,

$$\begin{aligned} W^2(\mu^{k+1/2}, \mu^\star) \leq & (1 - \gamma\lambda_F) W^2(\mu^k, \mu^\star) + 2\gamma^2(Ld + \sigma_F^2) \\ & - 2\gamma\left(\mathcal{E}_F(\mu^{k+1/2}) + \mathcal{H}(\mu^{k+1/2}) - \mathcal{E}_F(\mu^\star) - \mathcal{H}(\mu^\star) \right). \end{aligned} \tag{51}$$

We now prove Theorem 5. The main tool for the proof is Lemma 4. Replace x by $x_\star^{k+1/2} \sim \mu^{k+1/2}$ in (26). Then $y' = y_\star^{k+1} \sim \nu^{k+1}$ and $\operatorname{prox}_{\gamma G}(x_\star^{k+1/2}) \sim \mu^{k+1}$. Therefore,

$$W^2(\mu^{k+1}, \mu^\star) \leq \mathbb{E}(\|\operatorname{prox}_{\gamma G}(x_\star^{k+1/2}) - x^\star\|^2), \quad W^2(\nu^{k+1}, \nu^\star) \leq \mathbb{E}(\|y_\star^{k+1} - y^\star\|^2).$$

Consequently, taking expectation in (26) we get

$$\begin{aligned} W^2(\mu^{k+1}, \mu^\star) \leq & W^2(\mu^{k+1/2}, \mu^\star) - \gamma(\lambda_{G^*} + \gamma) W^2(\nu^{k+1}, \nu^\star) + \gamma^2 \int \|y\|^2 d\nu^\star(y) \\ & - 2\gamma\left(\mathcal{E}_{G^*}(\nu^{k+1}) - \mathcal{E}_G^*(\nu^\star) - \mathbb{E}\langle y_\star^{k+1}, x^\star \rangle + \mathbb{E}\langle y^\star, x_\star^{k+1/2} \rangle \right). \end{aligned}$$

Combining with Lemma 7, we get the result.

B.5 Proof of Corollary 6

From Theorem 5,

$$\gamma(\lambda_{G^*} + \gamma) W^2(\nu^{j+1}, \nu^\star) + 2\gamma\mathscr{D}(\mu^{j+1/2}, y_\star^{j+1}) \leq W^2(\mu^j, \mu^\star) - W^2(\mu^{j+1}, \mu^\star) + \gamma^2 C. \tag{52}$$

Summing over $j \in \{0, \dots, k-1\}$,

$$\gamma(\lambda_{G^*} + \gamma) \sum_{j=0}^{k-1} W^2(\nu^{j+1}, \nu^\star) + 2\gamma \sum_{j=0}^{k-1} \mathscr{D}(\mu^{j+1/2}, y_\star^{j+1}) \tag{53}$$

$$\leq W^2(\mu^0, \mu^\star) - W^2(\mu^k, \mu^\star) + k\gamma^2 C. \tag{54}$$

Therefore,

$$\gamma(\lambda_{G^*}+\gamma)k \min_{j\in\{0,\dots,k-1\}} W^2(\nu^{j+1},\nu^\star) + 2\gamma k \min_{j\in\{0,\dots,k-1\}} \mathscr{D}(\mu^{j+1/2}, y_\star^{j+1}) \tag{55}$$

$$\leq W^2(\mu^0,\mu^\star) + k\gamma^2 C, \tag{56}$$

which implies

$$\min_{j\in\{0,\dots,k-1\}} \mathscr{D}(\mu^{j+1/2}, y_\star^{j+1}) \leq \frac{1}{2\gamma k} W^2(\mu^0,\mu^\star) + \gamma\frac{C}{2}, \tag{57}$$

and,

$$\min_{j\in\{1,\dots,k\}} W^2(\nu^j,\nu^\star) \leq \frac{1}{\gamma(\lambda_{G^*}+\gamma)k} W^2(\mu^0,\mu^\star) + \frac{\gamma}{\lambda_{G^*}+\gamma} C. \tag{58}$$

Moreover, if $\lambda_F > 0$, Theorem 5 implies

$$W^2(\mu^{k+1},\mu^\star) \leq (1-\gamma\lambda_F) W^2(\mu^k,\mu^\star) + \gamma^2 C. \tag{59}$$

Iterating, we obtain

$$W^2(\mu^k,\mu^\star) \leq (1-\gamma\lambda_F)^k W^2(\mu^0,\mu^\star) + \gamma\frac{C}{\lambda_F}. \tag{60}$$

C Further intuition on PSGLA

C.1 Stochastic gradient descent interpretation of the stochastic gradient Langevin algorithm

As mentionned in the introduction, Langevin algorithm can be interpreted as a gradient descent algorithm in the space $\mathcal{P}_2(\mathsf{X})$ to minimize $\mathrm{KL}(\cdot|\mu^\star)$, see *e.g.* [19]. More precisely, consider the case where $G \equiv 0$ and denote μ^k the distribution of x^k. Then PSGLA boils down to the stochastic gradient Langevin algorithm (*i.e.*, PSGLA without proximal step) and satisfy the following inequality (Lemma 7)

$$W^2(\mu^{k+1},\mu^\star) \leq (1-\gamma\lambda_F) W^2(\mu^k,\mu^\star) - 2\gamma\left(\mathcal{F}(\mu^{k+1}) - \mathcal{F}(\mu^\star)\right) + 2\gamma^2(Ld+\sigma_F^2), \tag{61}$$

if F is L-smooth, λ_F-strongly convex and $\gamma \leq 1/L$. The last inequality is similar to a standard inequality used in the analysis of SGD. More precisely, the analysis of SGD often relies on an inequality similar to (61), by replacing the Wasserstein distance by the Euclidean distance and $\mathcal{F}$ by the objective function to be minimized by SGD (note that $Ld+\sigma_F^2$ is a constant). Therefore, unrolling the recursion (61) (which is the standard way to obtain convergence rates for SGD) leads to the complexity $\mathcal{O}(1/\varepsilon^2)$ in terms of objective gap $\mathcal{F}(\mu) - \mathcal{F}(\mu^\star)$. Using (15), recall that the objective gap is the KL divergence.

In this paper, we considered the case $G \neq 0$. One can obtain an inequality similar to (61) for PSGLA if G is Lipschitz continuous, see [19, 27]. However, for a general $G \in \Gamma_0(\mathsf{X})$, it is hopeless. Indeed, $\mathcal{F}(\mu^{k+1}) - \mathcal{F}(\mu^\star) = +\infty$ in general because $\mathcal{H}(\mu^{k+1}) = +\infty$ since μ^{k+1} is not absolutely continuous w.r.t. Leb (*e.g.* when the proximal step is a projection). Moreover, $\mathcal{F}(\mu^{k+1/2}) - \mathcal{F}(\mu^\star) = +\infty$ in general because $\mathcal{E}_G(\mu^{k+1/2}) = +\infty$ since $\mu^{k+1/2}$ is not supported by $\mathrm{dom}(G)$ ($\mathrm{supp}(\mu^{k+1/2}) = \mathsf{X}$ because of the Gaussian noise). Therefore, one cannot obtain a rate in terms of KL divergence (*i.e.*, objective gap) for PSGLA in general, since the KL divergence is equal to $+\infty$.

On order to overcome this difficulty, we assumed 4 and adopted a primal dual interpretation of PSGLA where PSGLA is seen as a Forward Backward algorithm involving monotone operators. We obtained an inequality similar to (61), but with the duality gap instead of the objective gap, and we proved that the duality gap is nonnegative.

C.2 Primal dual interpretation of the proximal gradient algorithm

The approach of this paper can also be used to interpret the proximal gradient algorithm as a primal dual algorithm.

Consider the minimization problem

$$\min_{x\in\mathsf{X}} F(x) + G(x). \tag{62}$$

To solve Problem (62), the proximal gradient algorithm is written

$$x^{k+1} = \mathrm{prox}_{\gamma G}\left(x^k - \gamma \nabla F(x^k)\right). \tag{63}$$

The proximal gradient algorithm can be seen as a primal dual algorithm for Problem (62) [24].

Indeed, a solution $x^\star$ to Problem (62) satisfies $0 \in \nabla F(x^\star) + \partial G(x^\star)$. Consider the dual variable $y^\star \in \partial G(x^\star)$ such that $0 = \nabla F(x^\star) + y^\star$. Since, $y^\star \in \partial G(x^\star)$, $0 \in -x^\star + \partial G^*(y^\star)$ using $\partial G^* = (\partial G)^{-1}$. Finally,

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \nabla F(x^\star) & +y^\star \\ -x^\star & +\partial G^*(y^\star) \end{bmatrix}. \tag{64}$$

Consider the set valued maps

$$B : (x, y) \mapsto \begin{bmatrix} \nabla F(x) \\ 0 \end{bmatrix},$$

and

$$A : (x, y) \mapsto \begin{bmatrix} y \\ -x \quad +\partial G^*(y) \end{bmatrix},$$

where we used vector notation. The maps A and B are maximal monotone operators (note that B is the gradient of $(x, y) \mapsto F(x)$ and A was used in Section 4). Inclusion (64) can be rewritten as

$$0 \in (A + B)(x^\star, y^\star). \tag{65}$$

In order to solve (65), one can apply the Forward Backward algorithm

$$P(x^{k+1/2} - x^k) = -\gamma B(x^k), \quad P(x^{k+1} - x^{k+1/2}) \in -\gamma A(x^{k+1}), \tag{66}$$

for a well chosen $P \in \mathbb{R}^{d\times d}_{++}$. As above, we take $P : (x, y) \mapsto x$. Although the matrix P is only semi-definite positive, we showed in Lemma (4) that $x^{k+1} = \mathrm{prox}_{\gamma G}(x^{k+1/2})$. Hence, the primal dual Forward Backward algorithm (66) is equivalent to the proximal gradient algorithm (63).

Moreover, the proof technique used for Theorem 5 can be adapted to analyze the proximal gradient algorithm as a primal dual algorithm. The complexity result obtained for the proximal gradient algorithm with this approach is suboptimal. However, the derivation of this complexity result sheds some light on PSGLA.

First, using the (strong) convexity of F,

$$\begin{aligned}
\|x^{k+1/2} - x^\star\|^2 &= \|x^k - x^\star\|^2 + \gamma^2\|\nabla F(x^k)\|^2 - 2\gamma\langle \nabla F(x^k), x^k - x^\star\rangle \\
&\leq (1 - \gamma\lambda_F)\|x^k - x^\star\|^2 + \gamma^2\|\nabla F(x^k)\|^2 - 2\gamma\left(F(x^k) - F(x^\star)\right) \\
&\leq (1 - \gamma\lambda_F)\|x^k - x^\star\|^2 + \gamma^2\|\nabla F(x^k)\|^2 - 2\gamma\left(F(x^{k+1/2}) - F(x^\star)\right) \\
&\quad - 2\gamma\left(F(x^k) - F(x^{k+1/2})\right).
\end{aligned}$$

Using the smoothness of F,

$$F(x^{k+1/2}) - F(x^k) \leq \langle \nabla F(x^k), x^{k+1/2} - x^k\rangle + \frac{L}{2}\|x^{k+1/2} - x^k\|^2 = -\gamma\left(1 - \frac{\gamma L}{2}\right)\|\nabla F(x^k)\|^2.$$

Therefore,

$$\|x^{k+1/2} - x^\star\|^2 \leq \|x^k - x^\star\|^2 - \gamma^2\left(1 - \gamma L\right)\|\nabla F(x^k)\|^2 - 2\gamma\left(F(x^{k+1/2}) - F(x^\star)\right). \tag{67}$$

Inequality (67) is analogue to Lemma 7. Moreover, using Lemma 4,

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 \leq &\|x^{k+1/2} - x^\star\|^2 \\
&- 2\gamma\left(G^*(y^{k+1}) - G^*(y^\star) - \langle y^{k+1}, x^\star\rangle + \langle y^\star, x^{k+1/2}\rangle\right) \\
&- \gamma(\lambda_{G^*} + \gamma)\|y^{k+1} - y^\star\|^2 + \gamma^2\|y^\star\|^2,
\end{aligned}$$

where $y^{k+1} = \mathrm{prox}_{G^*/\gamma}(x^k/\gamma)$. Summing the two last inequality, and using $\gamma \leq 1/L$,

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 \leq &(1 - \gamma\lambda_F)\|x^k - x^\star\|^2 - \gamma(\lambda_{G^*} + \gamma)\|y^{k+1} - y^\star\|^2 \\
&- 2\gamma\left(\mathscr{L}(x^{k+1/2}, y^\star) - \mathscr{L}(x^\star, y^{k+1})\right) + \gamma^2\|y^\star\|^2,
\end{aligned} \tag{68}$$

where $\mathscr{L}(x, y) = F(x) - G^*(y) + \langle x, y\rangle$ is the Lagrangian function and $\mathscr{L}(x^{k+1/2}, y^\star) - \mathscr{L}(x^\star, y^{k+1})$ is the duality gap. The last inequality is similar to Theorem 5.

Remark 1. With slight modifications of the derivations above, one can get the better result

$$\|x^{k+1} - x^\star\|^2 \leq (1 - \gamma\lambda_F)\|x^k - x^\star\|^2 - 2\gamma\left(\mathscr{L}(x^{k+1}, y^\star) - \mathscr{L}(x^\star, y^{k+1})\right).$$

However, the proof technique would not adapt to Langevin algorithm.

Remark 2. Similarly to the result of Theorem 5, $x^\star, y^\star$ can be replaced by any $\bar{x}, \bar{y}$. The proof technique does not use specific properties of $x^\star, y^\star$, as being primal dual optimal. Primal dual optimality of $x^\star, y^\star$ is only needed to prove that the duality gap is nonnegative.

D Generalization to a stochastic three operators splitting

In order to cover more applications, for instance involving *several Lipschitz proximable terms* in the potential, we quickly generalize the results of Section 5. Our primal dual framework can be plugged to the results of [27] instead of [19], leading to an extension of Section 5.

Consider the task of sampling from $\mu^\star \propto \exp(-V)$, where

$$V(x) = \mathbb{E}(f(x,\xi)) + \mathbb{E}(r(x,\xi)) + G(x). \tag{69}$$

We assume the following.

Assumption 6. For every $x \in \mathsf{X}$, $r(x,\xi)$ is integrable and $R(x) := \mathbb{E}_\xi(r(x,\xi))$. Moreover, $r(\cdot,\xi) \in \Gamma_0(\mathsf{X})$ a.s. Finally, there exists $M \geq 0$ such that for every $x \in \mathsf{X}$, $\mathbb{E}_\xi(\|\partial^0 r(x,\xi)\|^2) \leq M^2$.

Assumption 6 holds *e.g.* if $r(x,\xi)$ is $\ell(\xi)$-Lipschitz continuous and $\mathbb{E}_\xi(\ell^2(\xi)) < \infty$, since $\mathbb{E}_\xi(\|\nabla^0 r(x,\xi)\|^2) \leq \mathbb{E}_\xi(\ell^2(\xi))$. By replacing F by $F + R$ in Section 3, Theorem 3 still hold with the Lagrangian function

$$\mathscr{L}(\mu, y) := \mathcal{E}_F(\mu) + \mathcal{E}_R(\mu) + \mathcal{H}(\mu) - \mathcal{E}_{G^*}(\nu) + \mathbb{E}\langle x, y\rangle, \tag{70}$$

where $x = T^\mu_{\mu^\star}(x^\star)$. Note that $\mathrm{dom}(R) = \mathsf{X}$, hence R is differentiable a.e. using [24, Theorem 25.5]. In order to sample from $\mu^\star$, the stochastic proximal Langevin algorithm (SPLA) [27] is written

$$x^{k+1} = \mathrm{prox}_{\gamma G}\left(\mathrm{prox}_{\gamma r(\cdot,\xi)}\left(x^k - \gamma\nabla_x f(x^k, \xi^{k+1}) + \sqrt{2\gamma}W^{k+1}\right)\right). \tag{71}$$

SPLA recovers PSGLA by taking $R \equiv 0$. Moreover, SPLA is analyzed in [27] only in the case where G also satisfies Assumption 6 (and hence $\mathrm{dom}(G) = \mathsf{X}$) which is stronger than assuming 4. In this section, we denote

$$x^{k+1/2} := \mathrm{prox}_{\gamma r(\cdot,\xi)}\left(x^k - \gamma\nabla_x f(x^k, \xi^{k+1}) + \sqrt{2\gamma}W^{k+1}\right), \tag{72}$$

and $\mu^{k+1/2}$ its distribution. We shall prove the following extension of Theorem 5 assuming that G satisfies the Assumption 4. This extension of Theorem 5 leads to an extension of Corollary 6 providing complexity results for SPLA similar to PSGLA.

Theorem 8. Let Assumptions 1, 2, 3, 5 and 6 hold true. If F is λ_F-strongly convex and G is $1/\lambda_{G^*}$-smooth, then for every $\gamma \leq 1/L$,

$$\begin{aligned} W^2(\mu^{k+1}, \mu^\star) \leq &(1 - \gamma\lambda_F)W^2(\mu^k, \mu^\star) - \gamma(\lambda_{G^*} + \gamma)W^2(\nu^{k+1}, \nu^\star) \\ &- 2\gamma\left(\mathscr{L}(\mu^{k+1/2}, y^\star) - \mathscr{L}(\mu^\star, y_\star^{k+1})\right) + \gamma^2 C, \end{aligned} \tag{73}$$

where $C := \int_{\mathrm{int}(D)} \|\nabla G(x)\|^2 d\mu^\star(x) + 2(Ld + \sigma_F^2 + M^2)$ and $y_\star^{k+1} := \mathrm{prox}_{G^*/\gamma}(x_\star^{k+1/2}/\gamma) \sim \nu^{k+1}$, where $x_\star^{k+1/2} := T^{\mu^{k+1/2}}_{\mu^\star}(x^\star)$.

Before proving Theorem 8, we recall the following consequence of [27, Theorem 1], which generalizes Lemma 7.

Lemma 9 ([27])**.** Let Assumptions 1, 2, 5 and 6 hold true. Then, if F is λ_F-strongly convex, for every $\gamma \leq 1/L$,

$$\begin{aligned} W^2(\mu^{k+1/2}, \mu^\star) \leq &(1 - \gamma\lambda_F)W^2(\mu^k, \mu^\star) + 2\gamma^2(Ld + \sigma_F^2 + M^2) \\ &- 2\gamma\left(\mathcal{E}_F(\mu^{k+1/2}) + \mathcal{E}_R(\mu^{k+1/2}) + \mathcal{H}(\mu^{k+1/2}) - \mathcal{E}_F(\mu^\star) - \mathcal{E}_R(\mu^\star) - \mathcal{H}(\mu^\star)\right). \end{aligned} \tag{74}$$

Proof. Apply [27, Theorem 1] by taking $G_2 \equiv \ldots \equiv G_n \equiv 0$ and noting that the KL term in [27, Equation 3] is equal to $\left(\mathcal{E}_F(\mu^{k+1/2}) + \mathcal{E}_R(\mu^{k+1/2}) + \mathcal{H}(\mu^{k+1/2}) - \mathcal{E}_F(\mu^\star) - \mathcal{E}_R(\mu^\star) - \mathcal{H}(\mu^\star)\right)$ using our notations, see Equation (15). □

We now prove Theorem 8, similarly to Theorem 5.

The main tool for the proof is Lemma 4. Replace x by $x_\star^{k+1/2} \sim \mu^{k+1/2}$ in (26). Then $y' = y_\star^{k+1} \sim \nu^{k+1}$ and $\operatorname{prox}_{\gamma G}(x_\star^{k+1/2}) \sim \mu^{k+1}$. Therefore,

$$W^2(\mu^{k+1}, \mu^\star) \leq \mathbb{E}(\|\operatorname{prox}_{\gamma G}(x_\star^{k+1/2}) - x^\star\|^2), \quad W^2(\nu^{k+1}, \nu^\star) \leq \mathbb{E}(\|y_\star^{k+1} - y^\star\|^2).$$

Consequently, taking expectation in (26) we get

$$\begin{aligned} W^2(\mu^{k+1}, \mu^\star) \leq & W^2(\mu^{k+1/2}, \mu^\star) - \gamma(\lambda_{G^*} + \gamma) W^2(\nu^{k+1}, \nu^\star) + \gamma^2 \int \|y\|^2 d\nu^\star(y) \\ & - 2\gamma \left(\mathcal{E}_{G^*}(\nu^{k+1}) - \mathcal{E}_G^*(\nu^\star) - \mathbb{E}\langle y_\star^{k+1}, x^\star\rangle + \mathbb{E}\langle y^\star, x_\star^{k+1/2}\rangle\right). \end{aligned}$$

Combining with Lemma 9, we get the result.