Stochastic Proximal Langevin Algorithm: Potential Splitting and Nonasymptotic Rates

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

Abstract

We propose a new algorithm—Stochastic Proximal Langevin Algorithm (SPLA)—for sampling from a log concave distribution. Our method is a generalization of the Langevin algorithm to potentials expressed as the sum of one stochastic smooth term and multiple stochastic nonsmooth terms. In each iteration, our splitting technique only requires access to a stochastic gradient of the smooth term and a stochastic proximal operator for each of the nonsmooth terms. We establish nonasymptotic sublinear and linear convergence rates under convexity and strong convexity of the smooth term, respectively, expressed in terms of the KL divergence and Wasserstein distance. We illustrate the efficiency of our sampling technique through numerical simulations on a Bayesian learning task.

1   Introduction

Many applications in the field of Bayesian machine learning require to sample from a probability distribution $\mu^\star$ with density $\mu^\star(x)$, $x \in \mathbb{R}^d$. Due to their scalability, Monte Carlo Markov Chain (MCMC) methods such as Langevin Monte Carlo [48] or Hamiltonian Monte Carlo [28] are popular algorithms to solve such problems. Monte Carlo methods typically generate a sequence of random variables $(x^k)_{k \geq 0}$ with the property that the distribution of x^k approaches $\mu^\star$ as k grows.

While the theory of MCMC algorithms has remained mainly asymptotic, in recent years the exploration of non-asymptotic properties of such algorithms has led to a renaissance in the field [14, 26, 39, 15, 16, 19, 22, 12, 53, 10, 52]. In particular, if $\mu^\star(x) \propto \exp(-U(x))$, where U is a smooth convex function, [14, 19] provide explicit convergence rates for the *Langevin algorithm (LA)*

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

where $\gamma > 0$ and $(W^k)_{k \geq 0}$ is a sequence of i.i.d. standard Gaussian random variables. The function U, also called the *potential*, enters the algorithm through its gradient.

In optimization, the problem $\min U$ where U is composite, *i.e.* is a sum of nonsmooth terms which must be handled separately, has many instances, see [17, Section 2]. These optimization problems can be seen as a Maximum A Posteriori (MAP) computation of some Bayesian model. Sampling a posterori in these models allows for a better Bayesian inference [20]. In these cases, the task of sampling a posterori takes the form of sampling from the target distribution $\mu^\star$, where U has a composite form.

In this work we study the setting where the potential U is the sum of a single smooth and a potentially large number of nonsmooth convex functions. In particular, we consider the problem

$$\text{Sample from} \quad \mu^\star(x) \propto \exp(-U(x)), \quad \text{where} \quad U(x) := F(x) + \sum_{i=1}^{n} G_i(x), \qquad (1)$$

---

[*]Also affiliated with Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

Table 1: Complexity results obtained in Corollaries 2, 3 and 4 of our main result (Theorem 1).

F	Stepsize γ	Rate	Theorem
convex	$\mathcal{O}(\varepsilon)$	$\mathrm{KL}(\mu_{\hat{x}^k} \mid \mu^\star) \leq \frac{1}{2\gamma(k+1)} W^2(\mu_{x^0}, \mu^\star) + \mathcal{O}(\gamma)$	Cor 2
α-strongly convex	$\mathcal{O}(\varepsilon\alpha)$	$W^2(\mu_{x^k}, \mu^\star) \leq (1-\gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) + \mathcal{O}\left(\frac{\gamma}{\alpha}\right)$	Cor 3
α-strongly convex	$\mathcal{O}(\varepsilon\alpha)$	$\mathrm{KL}(\mu_{\tilde{x}^k} \mid \mu^\star) \leq \alpha(1-\gamma\alpha)^{k+1} W^2(\mu_{x^0}, \mu^\star) + \mathcal{O}(\gamma)$	Cor 4

where $F : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function and $G_1, \ldots, G_n : \mathbb{R}^d \to \mathbb{R}$ are (possibly nonsmooth) convex functions. The additive model for U offers ample flexibility as typically there are multiple decompositions of U in the form (1).

2 Contributions

We now briefly comment some of the key contributions of this work.

⋄ **A splitting technique for Langevin algorithm.** We propose a new variant of LA for solving (1), which we call *Stochastic Proximal Langevin Algorithm (SPLA)*. We assume that F and G can be written as expectations over some simpler functions $f(\cdot, \xi)$ and $g_i(\cdot, \xi)$

$$F(x) = \mathbb{E}_\xi(f(x, \xi)), \quad \text{and} \quad G_i(x) = \mathbb{E}_\xi(g_i(x, \xi)). \tag{2}$$

SPLA (see Algorithm 1 in Section 4) only requires accesses to the gradient of $f(\cdot, \xi)$ and to proximity operators of the functions $g_i(\cdot, \xi)$. SPLA can be seen as a Langevin version of the stochastic Passty algorithm [30, 36]. To the best of our knowledge, this is the first time a splitting technique that involves multiple (stochastic) proximity operators is used in a Langevin algorithm.

Remarks: Current forms of LA tackle problem (1) using stochastic subgradients [18]. If $n = 1$ and G_1 is proximable (i.e., the learner has access to the full proximity operator of G_1), it has recently been proposed to use proximity operators instead of (sub)gradients [20, 18], as it is done in the optimization literature [29, 2]. Indeed, in this case, the proximal stochastic gradient method is an efficient method to minimize U. If $n > 1$, and the functions G_i are proximable (but not U), the minimization of U is usually tackled using the *operator splitting* framework: the (stochastic) three-operator splitting [51, 17] or (stochastic) primal dual algorithms [13, 46, 9, 35]. These algorithms involve the computation of (stochastic) gradients and (full) proximity operators and enjoy numerical stability properties. However, proximity operators are sometimes difficult to implement. In this case, *stochastic proximity operators* are cheaper[2] than full proximity operators and numerically more stable than stochastic subgradients to handle nonsmooth terms [32, 31, 4, 5, 6] but also smooth [43] terms. In this paper, we bring together the advantages of operator splitting and stochastic proximity operators for sampling purposes.

⋄ **Theory.** We perform a nonasymptotic convergence analysis of SPLA. Our main result, Theorem 1, gives a tractable recursion involving the Kullback-Leibler divergence and Wasserstein distance (when U is strongly convex) between $\mu^\star$ and the distribution of certain samples generated by our method. We use this result to show that the KL divergence is lower than ε after $\mathcal{O}(1/\varepsilon^2)$ iterations if the constant stepsize $\gamma = \mathcal{O}(\varepsilon)$ is used (Corollary 2). Assuming F is α-strongly convex, we show that the Wasserstein distance and (resp. the KL divergence) decrease exponentially, up to an oscillation region of size $\mathcal{O}(\gamma/\alpha)$ (resp. $\mathcal{O}(\gamma)$) as shown in Corollary 3 (resp. Corollary 4). If we wish to push the Wasserstein distance below ε (resp. the KL divergence below $\alpha\varepsilon$), this could be achieved by setting $\gamma = \mathcal{O}(\varepsilon\alpha)$, and it would be sufficient to take $\mathcal{O}(1/\varepsilon \log 1/\varepsilon)$ iterations. These results are summarized in Table 1. The obtained convergence rates match the previous known results obtained in simpler settings [18]. Note that convergence rates of optimization methods involving multiple stochastic proximity operators haven't been established yet.

---

[2]See www.proximity-operator.net

Remarks: Our proof technique is inspired by [38], which is itself based on [18]. In [38], the authors consider the $n = 1$ case, and assume that the smooth function F is proximable. In [18], a proximal stochastic (sub)gradient Langevin algorithm is studied. In this paper, convergence rates are established by showing that the probability distributions of the iterates shadow some discretized gradient flow defined on a measure space. Hence, our work is a contribution to recent efforts to understand Langevin algorithm as an optimization algorithm in a space of probability measures [26, 49, 3].

$\diamond$ **Online setting.** In online settings, U is unknown but revealed across time. Our approach provides a reasonable algorithm for such situations, especially in cases when the information revealed about U is stationary in time. In particular, this includes online Bayesian learning with structured priors or nonsmooth log likelihood [50, 23, 40, 47]. In this context, the learner is required to sample from some posterior distribution $\mu^\star$ that takes the form (1) where $F, G_1, \ldots, G_n$ are intractable. However, these functions can be cheaply sampled, or are revealed across time through i.i.d. streaming data.

$\diamond$ **Simulations.** We illustrate the promise of our approach numerically by performing experiments with SPLA. We first apply SPLA to a stochastic and nonsmooth toy model with a ground truth. Then, we consider the problem of *sampling from the posterior distribution in the Graph Trend Filtering* context [47]. For this nonsmooth large scale simulation problem, SPLA is performing better than the state of the art method that uses stochastic subgradients instead of stochastic proximity operators. Indeed, in the optimization litterature [2], proximity operators are already known to be more stable than subgradients.

3 Technical Preliminaries

In this section, we recall certain notions from convex analysis and probability theory, which are keys to the developments in this paper, state our main assumptions, and introduce needed notations.

3.1 Subdifferential, minimal section and proximity operator

Given a convex function $g : \mathbb{R}^d \to \mathbb{R}$, its *subdifferential* at x, $\partial g(x)$, is the set

$$\partial g(x) := \left\{ d \in \mathbb{R}^d \ : \ g(x) + \langle d, y - x \rangle \leq g(y) \right\}.$$

Since $\partial g(x)$ is a nonempty closed convex set [2], the projection of 0 onto $\partial g(x)$—the least norm element in the set $\partial g(x)$—is well defined, and we call this element $\nabla^0 g(x)$. The function $\nabla^0 g : \mathbb{R}^d \to \mathbb{R}^d$ is called the *minimal section* of ∂g. The *proximity operator* associated with g is the mapping $\mathrm{prox}_g : \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$\mathrm{prox}_g(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \tfrac{1}{2} \|x - y\|^2 + g(y) \right\}.$$

Due to its implicit definition, prox_g can be hard to evaluate.

3.2 Stochastic structure of F and G_i: integrability, smoothness and convexity

Here we detail the assumptions behind the stochastic structure (2) of the functions $F = \mathbb{E}_\xi(f(x, \xi))$ and $G_i = \mathbb{E}_\xi(g_i(x, \xi))$ defining the potential U. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and denote $\mathbb{E}$ the mathematical expectation and $\mathbb{V}$ the variance. Consider ξ a random variable from Ω to another probability space $(\Xi, \mathscr{G})$ with distribution μ.

Assumption 1 (Integrability). The functions $f : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ and $g_i : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$, $i = 1, \ldots, n$, are μ-integrable for every $x \in \mathbb{R}^d$.

Furthermore, we will make the following convexity and smoothness assumptions.

Assumption 2 (Convexity and differentiability). The function $f(\cdot, s)$ is convex and differentiable for every $s \in \Xi$. The functions $g_i(\cdot, s)$ are convex for every $i \in \{1, 2, \ldots, n\}$.

The gradient of $f(\cdot, s)$ is denoted $\nabla f(\cdot, s)$, the subdifferential of $g_i(\cdot, s)$ is denoted $\partial g_i(\cdot, s)$ and its minimal section is denoted $\nabla^0 g_i(\cdot, s)$. Under Assumption 2, it is known that F is convex and differentiable and that $\nabla F(x) = \mathbb{E}_\xi(\nabla f(x, \xi))$ [34]. Next, we assume that F is smooth and α-strongly convex. However, we allow $\alpha = 0$ if F is not strongly convex. We will only assume that $\alpha > 0$ in Corollaries 3 and 4.

Assumption 3 (Convexity and smoothness of F)**.** The gradient of F is L-Lipschitz continuous, where $L \geq 0$. Moreover, F is α-strongly convex, where $\alpha \geq 0$.

Under Assumption 2, the second part of the above holds for $\alpha = 0$. Finally, we will introduce two noise conditions on the stochastic (sub)gradients of $f(\cdot, s)$ and $g_i(\cdot, s)$.

Assumption 4 (Bounded variance of $\nabla f(x, \cdot)$)**.** There exists $\sigma_F \geq 0$, such that $\mathbb{V}_\xi(\|\nabla f(x, \xi)\|) \leq \sigma_F^2$ for every $x \in \mathbb{R}^d$.

Assumption 5 (Bounded second moment of $\nabla^0 g_i(x, \cdot)$)**.** For every $i \in \{1, 2, \ldots, n\}$, there exists $L_{G_i} \geq 0$ such that $\mathbb{E}_\xi(\|\nabla^0 g_i(x, \xi)\|^2) \leq L_{G_i}^2$ for every $x \in \mathbb{R}^d$.

Note that if $g_i(\cdot, s)$ is $\ell_i(s)$-Lipschitz continuous for every $s \in \Xi$, and if $\ell_i(s)$ is μ-square integrable, then Assumption 5 holds.

3.3 KL divergence, entropy and potential energy

Recall from (1) that $U := F + \sum_{i=1}^n G_i$ and assume that $\int \exp(-U(x))dx < \infty$. Our goal is to sample from the unique distribution $\mu^\star$ over $\mathbb{R}^d$ with density $\mu^\star(x)$ (w.r.t. the Lebesgue measure denoted $\mathcal{L}$) proportional to $\exp(-U(x))$, for which we write $\mu^\star(x) \propto \exp(-U(x))$. The closeness between the samples of our algorithm and the target distribution $\mu^\star$ will be evaluated in terms of information theoretic and optimal transport theoretic quantities.

Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-field of $\mathbb{R}^d$. Given two nonnegative measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we write $\mu \ll \nu$ if μ is absolutely continuous w.r.t. ν, and denote $\frac{d\mu}{d\nu}$ its density. The *Kullback-Leibler (KL) divergence* between μ and ν, $\mathrm{KL}(\mu \mid \nu)$, quantifies the closeness between μ and ν. If $\mu \ll \nu$, then the KL divergence is defined by

$$\mathrm{KL}(\mu \mid \nu) := \int \log\left(\frac{d\mu}{d\nu}(x)\right) d\mu(x),$$

and otherwise we set $\mathrm{KL}(\mu \mid \nu) = +\infty$. Up to an additive constant, $\mathrm{KL}(\cdot \mid \mu^\star)$ can be seen as the sum of two terms [37]: the *entropy* $\mathcal{H}(\mu)$ and the *potential energy* $\mathcal{E}_U(\mu)$. The entropy of μ is given by $\mathcal{H}(\mu) := \mathrm{KL}(\mu \mid \mathcal{L})$, and the potential energy of μ is defined by $\mathcal{E}_U(\mu) := \int U d\mu(x)$.

3.4 Wasserstein distance

Although the KL divergence is equal to zero if and only if $\mu = \nu$, it is not a mathematical distance (metric). The *Wasserstein distance*, defined below, metricizes the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures over $\mathbb{R}^d$ with a finite second moment. Consider $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. A *transference plan* of (μ, ν) is a probability measure υ over $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ with marginals μ, ν: for every $A \in \mathcal{B}(\mathbb{R}^d)$, $\upsilon(A \times \mathbb{R}^d) = \mu(A)$ and $\upsilon(\mathbb{R}^d \times A) = \nu(A)$. In particular, the product measure $\mu \otimes \nu$ is a transference plan. We denote $\Gamma(\mu, \nu)$ the set of transference plans. A *coupling* of (μ, ν) is a random variable (X, Y) over some probability space with values in $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ (i.e., X and Y are random variables with values in $\mathbb{R}^d$) such that the distribution of X is μ and the distribution of Y is ν. In other words, (X, Y) is a coupling of μ, ν if the distribution of (X, Y) is a transference plan of μ, ν. The Wasserstein distance of order 2 between μ and ν is defined by

$$W^2(\mu, \nu) := \inf\left\{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\upsilon(x, y), \quad \upsilon \in \Gamma(\mu, \nu)\right\}.$$

One can see that $W^2(\mu, \nu) = \inf \mathbb{E}(\|X - Y\|^2)$, where the inf is taken over all couplings (X, Y) of μ, ν defined on some probability space with expectation $\mathbb{E}$.

4 The SPLA Algorithm and its Convergence Rates

4.1 The algorithm

To solve the sampling problem (1), our Stochastic Proximal Langevin Algorithm (SPLA) generates a sequence of random variables $(x^k)_{k \geq 0}$ from $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ defined as follows

$$
\begin{aligned}
z^k &= x^k - \gamma \nabla f(x^k, \xi^k) \\
y_0^k &= z^k + \sqrt{2\gamma} W^k \\
y_i^k &= \text{prox}_{\gamma g_i(\cdot, \xi^k)}(y_{i-1}^k) \quad \text{for} \quad i = 1, \ldots, n \\
x^{k+1} &= y_n^k,
\end{aligned}
$$

where $(W^k)_{k \geq 0}$ is a sequence of i.i.d. standard Gaussian random variables, $(\xi^k)_{k \geq 0}$ is a sequence of i.i.d. copies of ξ and $\gamma > 0$ is a positive step size. Our SPLA method is formalized as Algorithm 1; its steps are explained therein.

Algorithm 1 Stochastic Proximal Langevin Algorithm (SPLA)

Initialize: $x^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
    Sample random ξ^k $\triangleright$ used for stoch. approximation: $F \approx f(\cdot, \xi^k)$ and $G_i \approx g_i(\cdot, \xi^k)$
    $z^k = x^k - \gamma \nabla f(x^k, \xi^k)$ $\triangleright$ a stochastic gradient descent step in F
    Sample random W^k $\triangleright$ a standard Gaussian vector in $\mathbb{R}^d$
    $y_0^k = z^k + \sqrt{2\gamma} W^k$ $\triangleright$ a Langevin step w.r.t. F
    for $i = 1, \ldots, n$ **do**
        $y_i^k = \text{prox}_{\gamma g_i(\cdot, \xi^k)}(y_{i-1}^k)$ $\triangleright$ prox step to handle the term $G_i(\cdot) = \mathbb{E}_\xi g_i(\cdot, \xi)$
    end for
    $x^{k+1} = y_n^k$ $\triangleright$ the final SPLA step, accounting for F and $G_1, G_2, \ldots, G_n$
end for

4.2 Main theorem

We now state our main results in terms of Kullback-Leibler divergence and Wasserstein distance. We denote μ_x the distribution of every random variable x defined on $(\Omega, \mathscr{F}, \mathbb{P})$.

Theorem 1. Let Assumptions 1–5 hold and assume that $\gamma \leq 1/L$. There exists $C \geq 0$ such that,

$$
2\gamma \, \text{KL}(\mu_{y_0^k} \mid \mu^\star) \leq (1 - \gamma\alpha) W^2(\mu_{x^k}, \mu^\star) - W^2(\mu_{x^{k+1}}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld + C). \tag{3}
$$

The constant C can be expressed as a linear combination of $L_{G_1}^2, \ldots, L_{G_n}^2$ with integer coefficients. Moreover, if $n = 2$, then $C := 2(L_{G_1}^2 + L_{G_2}^2)$. More generally, if for every $i \in \{2, \ldots, n\}$, $g_i(\cdot, \xi)$ admits almost surely the representation $g_i(\cdot, \xi) = \tilde{g}_i(\cdot, \xi_i)$ where $\xi_2, \ldots, \xi_n$ are independent random variables, then $C := n \sum_{i=1}^n L_{G_i}^2$.

Proof. A full proof can be found in the Supplementary material. We only sketch the main steps here. For every μ-integrable function $g : \mathbb{R}^d \to \mathbb{R}$, we denote $\mathcal{E}_g(\mu) = \int g \mathrm{d}\mu$. Moreover, we denote $\mathcal{F} = \mathcal{E}_U + \mathcal{H}$. First, using [18, Lemma 1], $\mu^\star \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathcal{E}_U(\mu^\star), \mathcal{H}(\mu^\star) < \infty$ and if $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$
\text{KL}(\mu \mid \mu^\star) = \mathcal{E}_U(\mu) + \mathcal{H}(\mu) - (\mathcal{E}_U(\mu^\star) + \mathcal{H}(\mu^\star)) = \mathcal{F}(\mu) - \mathcal{F}(\mu^\star),
$$

provided that $\mathcal{E}_U(\mu) < \infty$. Then, we decompose $\mathcal{E}_U(\mu) = \mathcal{E}_F(\mu) + \mathcal{E}_G(\mu)$ where $G = \sum_{i=1}^n G_i$. Using [18] again, we can establish the inequality

$$
2\gamma \left[\mathcal{H}(\mu_{y_0^k}) - \mathcal{H}(\mu^\star) \right] \leq W^2(\mu_{z^k}, \mu^\star) - W^2(\mu_{y_0^k}, \mu^\star). \tag{4}
$$

Then, if $\gamma \leq 1/L$ we obtain, for every random variable a with distribution $\mu^\star$,

$$
\mathbb{E} \left[\| z^k - a \|^2 \right] \leq (1 - \gamma\alpha) \mathbb{E} \left[\| x^k - a \|^2 \right] + 2\gamma \left[\mathcal{E}_F(\mu^\star) - \mathcal{E}_F(\mu_{z^k}) \right] + 2\gamma^2 \sigma_F^2, \tag{5}
$$

using standard computations regarding the stochastic gradient descent algorithm. Using the smoothness of F and the definition of the Wasserstein distance, this implies

$$2\gamma\left[\mathcal{E}_F(\mu_{y_0^k}) - \mathcal{E}_F(\mu^\star)\right] \leq (1-\gamma\alpha)W^2(\mu_{x^k},\mu^\star) - W^2(\mu_{z^k},\mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld).$$

It remains to establish $2\gamma\left[\mathcal{E}_G(\mu_{y_0^k}) - \mathcal{E}_G(\mu^\star)\right] \leq W^2(\mu_{y_0^k},\mu^\star) - W^2(\mu_{x^{k+1}},\mu^\star) + \gamma^2 C$, which is the main technical challenge of the proof. This is done using the frameworks of Yosida approximation of random subdifferentials and Moreau regularizations of random convex functions [2]. Equation (3) is obtained by summing the obtained inequalities. $\qquad\square$

4.3 Link with Wasserstein Gradient Flows

Equation (3) is reminiscent of the fact that the SPLA shadows the gradient flow of $\mathrm{KL}(\cdot \mid \mu^\star)$ in the metric space $(\mathcal{P}_2(\mathbb{R}^d), W)$. To see this, first consider the gradient flow associated to F. By definition, it is the flow of the differential equation [7]

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{x}(t) = -\nabla F(\mathsf{x}(t)), \quad t > 0. \tag{6}$$

The function x can alternatively be defined as a solution of the variational inequalities

$$2\left(F(\mathsf{x}(t)) - F(a)\right) \leq -\frac{\mathrm{d}}{\mathrm{d}t}\|\mathsf{x}(t) - a\|^2, \quad t > 0, \quad \forall a \in \mathbb{R}^d. \tag{7}$$

The iterates $(u^k)_{k\geq 0}$ of the stochastic gradient descent (SGD) algorithm applied to F can be seen as a (noisy) Euler discretization of (6) with a step size $\gamma > 0$. This idea has been used successfully in the stochastic approximation litterature [33, 24]. This analogy goes further since a fundamental inequality used to analyze SGD applied to F is ([27])

$$2\gamma\mathbb{E}\left(F(u^{k+1}) - F(a)\right) \leq \mathbb{E}\|u^k - a\|^2 - \mathbb{E}\|u^{k+1} - a\|^2 + \gamma^2 K, \quad k \geq 0,$$

where $K \geq 0$ is some constant, which can be seen as a discrete counterpart of (7). Note that this inequality is similar to (5) that is used in the proof of Theorem 1.

In the optimal transport theory, the point of view of (7) is used to define the gradient flow of a (geodesically) convex function $\mathcal{F}$ defined on $\mathcal{P}_2(\mathbb{R}^d)$ (see [37] or [1, Page 280]). Indeed, the gradient flow $(\nu_t)_{t\geq 0}$ of $\mathcal{F}$ in the space $(\mathcal{P}_2(\mathbb{R}^d), W)$ satisfies for every $t > 0, \mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$2\left(\mathcal{F}(\nu_t) - \mathcal{F}(\mu)\right) \leq -\frac{\mathrm{d}}{\mathrm{d}t}W^2(\nu_t,\mu), \tag{8}$$

which can be seen as a continuous time counterpart of Equation (3) by setting $\mathcal{F} = \mathrm{KL}(\cdot \mid \mu^\star)$. Furthermore, Equation (4) in the proof of Theorem 1 is also related to (8). It is obtained by applying Equation (8) with $\mathcal{F} = \mathcal{H}$ and $\nu_0 = \mu_{z^k}$ (see *e.g* [18, Lemma 5]).

4.4 Explicit convergence rates for convex and strongly convex F

Corollaries 2, 3 and 4 below are obtained by unrolling the recursion provided by Theorem 1. The results are summarized in Table 1.

Corollary 2 (Convex F)**.** Consider a sequence of independent random variables $(j_k)_{k\geq 0}$ such that $(j_k)_{k\geq 0}$ is independent of $(W^k)_k$ and $(\xi^k)_k$, and the distribution of j_k is uniform over $\{0,\ldots,k\}$. Denote $\hat{x}^k = y_0^{j_k}$. If $\gamma \leq 1/L$, then,

$$\mathrm{KL}(\mu_{\hat{x}^k} \mid \mu^\star) \leq \frac{1}{2\gamma(k+1)}W^2(\mu_{x^0},\mu^\star) + \frac{\gamma}{2}(2\sigma_F^2 + 2Ld + C).$$

Hence, given any $\varepsilon > 0$, choosing stepsize $\gamma = \min\left\{\frac{1}{L}, \frac{\varepsilon}{2\sigma_F^2 + 2Ld + C}\right\}$ and a number of iterations

$$k + 1 \geq \max\left\{\frac{L}{\varepsilon}, \frac{2\sigma_F^2 + 2Ld + C}{\varepsilon^2}\right\}W^2(\mu_{x^0},\mu^\star),$$

implies $\mathrm{KL}(\mu_{\hat{x}^k} \mid \mu^\star) \leq \varepsilon$.

Corollary 3 (Strongly convex F)**.** If $\alpha > 0$ and $\gamma \leq 1/L$, then,

$$W^2(\mu_{x^k},\mu^\star) \leq (1-\gamma\alpha)^k W^2(\mu_{x^0},\mu^\star) + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{\alpha}.$$

Hence, given any $\varepsilon > 0$, choosing stepsize $\gamma = \min\left\{\frac{1}{L}, \frac{\varepsilon\alpha}{2(2\sigma_F^2 + 2Ld + C)}\right\}$ and a number of iterations

$$k \geq \max\left\{\frac{L}{\alpha}, \frac{2(2\sigma_F^2 + 2Ld + C)}{\varepsilon\alpha^2}\right\} \log\left(\frac{2W^2(\mu_{x^0}, \mu^\star)}{\varepsilon}\right),$$

implies $W^2(\mu_{x^k}, \mu^\star) \leq \varepsilon$.

Corollary 4 (Strongly convex F). Consider a sequence of independent random variables $(j_k)_{k\geq 0}$ such that $(j_k)_k$ is independent of $(W^k)_k$ and $(\xi^k)_k$. Assume that the distribution of j_k is geometric over $\{0, \ldots, k\}$:

$$\mathbb{P}(j_k = r) \propto (1 - \gamma\alpha)^{-r}.$$

Denote $\tilde{x}^k = x^{j_k}$. If $\alpha > 0$ and $\gamma \leq 1/L$, then,

$$\mathrm{KL}(\mu_{\tilde{x}^k} \mid \mu^\star) \leq \frac{\alpha W^2(\mu_{x^0}, \mu^\star)}{2} \cdot \frac{(1 - \gamma\alpha)^{k+1}}{1 - (1 - \gamma\alpha)^{k+1}} + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{2}.$$

Hence, given any $\varepsilon > 0$, choosing stepsize $\gamma = \min\left\{\frac{1}{L}, \frac{\varepsilon\alpha}{2\sigma_F^2 + 2Ld + C}\right\}$ and a number of iterations

$$k \geq \max\left\{\frac{L}{\alpha}, \frac{2\sigma_F^2 + 2Ld + C}{\varepsilon\alpha^2}\right\} \log\left(2\max\left\{1, \frac{W^2(\mu_{x^0}, \mu^\star)}{\varepsilon}\right\}\right),$$

implies $\mathrm{KL}(\mu_{\tilde{x}^k} \mid \mu^\star) \leq \alpha\varepsilon$.

We can compare these bounds with the one of [18]. First, in the particular case $n = 1$ and $g_1(\cdot, s) \equiv G_1$, SPLA boils down to the algorithm of [18, Section 4.2], Corollary 2 matches exactly [18, Corollary 18] and Corollary 3 matches [18, Corollary 22]. To our knowledge, Corollary 4 has no counterpart in the litterature. We now focus on the case $F \equiv 0$ and $n = 1$ of SPLA, as it concentrates the innovations of our paper. In this case, $L = 0$ and $\sigma_F = 0$. Compared to the Stochastic Subgradient Langevin Algorithm (SSLA) [18, Section 4.1], Corollary 2 matches with [18, Corollary 14].

5 Numerical experiments

5.1 Simulations using a ground truth

We first concentrate on the case $F \equiv 0$ and $n = 1$. Let $U = |x| = \mathbb{E}_\xi(|x| + x\xi)$ ($g_1(x, s) = |x| + xs$), where ξ is standard Gaussian. The target $\mu^\star \propto \exp(-U)$ is a standard Laplace distribution in $\mathbb{R}$. In this case, $L = \alpha = \sigma_F = 0$ and $C = L_{G_1}^2 = 2$. We shall illustrate the bound on $\mathrm{KL}(\mu_{\hat{x}^k} \mid \mu^\star)$ (Corollary 2 for SPLA and [18, Corollary 14] for SSLA) for both algorithms using histograms. Note that the distribution $\mu_{\hat{x}^k}$ of $\hat{x}^k$ is a (deterministic) mixture of the μ_{x^j}: $\mu_{\hat{x}^k} = \frac{1}{k}\sum_{j=1}^{k}\mu_{x^j}$. Using Pinsker inequality, we can bound the total variation distance between $\mu_{\hat{x}^k}$ and $\mu^\star$ from the bound on KL, and illustrate this by histograms. In Figure 1, we take $\gamma = 10$ and do 10^5 iterations of both algorithms. Note that here the complexity of one iteration of SPLA or SSLA is the same. One can
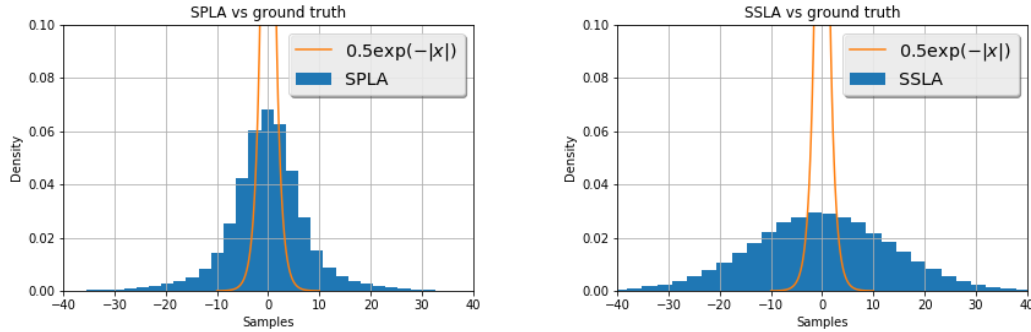

Figure 1: Comparison between histograms of SPLA and SSLA and the true density $0.5\exp(-|x|)$.

see that SPLA enjoys the well known advantages of stochastic proximal methods [43]: precision, numerical stability (less outliers), and robustness to step size.

Algorithm 2 SPLA for the Graph Trend Filtering

Initialize: $x^0 \in \mathbb{R}^V$
for $k = 0, 1, 2, \ldots$ **do**
$\quad z^k = x^k - \frac{\gamma}{\sigma^2}(x^k - Y)$
$\quad$ Sample random W^k $\qquad\qquad\qquad\qquad\qquad$ ▷ standard Gaussian vector in $\mathbb{R}^V$
$\quad y_0^k = z^k + \sqrt{2\gamma}W^k$
$\quad$ **for** $i = 1, \ldots, n$ **do**
$\quad\quad$ Sample uniform random edges e_i
$\quad\quad y_i^k = \text{prox}_{\gamma g_{e_i}}(y_{i-1}^k)$
$\quad$ **end for**
$\quad x^{k+1} = y_n^k$
end for

5.2 Application to Trend Filtering on Graphs

In this section we consider the following Bayesian point of view of *trend filtering on graphs* [42]. Consider a finite undirected graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. Denote d the cardinality of V and $|E|$ the cardinality of E. A realization of a random vector $Y \in \mathbb{R}^V$ is observed. In a Bayesian framework, the distribution of Y is parametrized by a vector $X \in \mathbb{R}^V$ which is itself random and whose distribution p is proportional to $\exp(-\lambda\,\text{TV}(x, G))$, where $\lambda > 0$ is a scaling parameter and where for every $x \in \mathbb{R}^V$

$$\text{TV}(x, G) = \sum_{i,j \in V, \{i,j\} \in E} |x(i) - x(j)|,$$

is the Total Variation regularization over G. The goal is to learn X after an observation of Y. The paper [47] consider the case where the distribution of Y given X (a.k.a the likelihood) is proportional to $\exp(-\frac{1}{2\sigma^2}\|X - y\|^2)$, where $\sigma \geq 0$ is another scaling parameter. In other words, the distribution of Y given X is $N(X, \sigma^2 I)$, a normal distribution centered at X with variance $\sigma^2 I$ (where I is the $d \times d$ identity matrix). Denoting

$$\pi(x \mid y) \propto \exp(-U(x)), \quad U(x) = \frac{1}{2\sigma^2}\|x - y\|^2 + \lambda\,\text{TV}(x, G),$$

the posterior distribution of X given Y, the maximum a posteriori estimator in this Bayesian framework is called the Graph Trend Filtering estimate [47]. It can be written

$$x^\star = \arg\max_{x \in \mathbb{R}^V} \pi(x \mid Y) = \arg\min_{x \in \mathbb{R}^V} \frac{1}{2\sigma^2}\|x - Y\|^2 + \lambda\,\text{TV}(x, G).$$

Although maximum a posteriori estimators carry some information, they are not able to capture uncertainty in the learned parameters. Samples a posteriori provide a better understanding of the posterior distribution and allow to compute other Bayesian estimates such as confidence intervals. This allows to avoid overfitting among other things. In our context, sampling a posteriori would require to sample from the target distribution $\mu^\star(x) = \pi(x \mid Y)$.

In the case where G is a 2D grid (which can be identified with an image), the proximity operator of $\text{TV}(\cdot, G)$ can be computed using a subroutine [8] and the proximal stochastic gradient Langevin algorithm can be used to sample from $\pi(\cdot \mid Y)$ [20, 18]. However, on a large/general graph, the proximity operator of $\text{TV}(\cdot, G)$ is hard to evaluate [41, 36]. Since $\text{TV}(\cdot, G)$ is written as a sum, we shall rather select a batch of random edges and compute the proximity operators over these randomly chosen edges. More precisely, we write the potential U defining $\pi(x \mid Y)$ in the form (1) by setting

$$U(x) = F(x) + \sum_{i=1}^n G_i(x), \quad F(x) = \frac{1}{2\sigma^2}\|x - Y\|^2, \quad G_i(x) = \lambda\frac{|E|}{n}\mathbb{E}_{e_i}\left(|x(v_i) - x(w_i)|\right),$$

where for every $i \in \{1, \ldots, n\}$, $e_i = \{v_i, w_i\} \in E$ is an uniform random edge and the e_i are independent. For every edge $e = \{v, w\} \in E$, (where v, w are vertices) denote $g_e(x) = \lambda\frac{|E|}{n}|x(v) - x(w)|$ and note that $G_i(x) = \mathbb{E}_{e_i}(g_{e_i}(x))$. The parameter n can be seen as a batch parameter: $\sum_{i=1}^n g_{e_i}(x)$ is an unbiaised approximation of $TV(x, G)$. Also note that we set $f(\cdot, s) \equiv F$. The SPLA applied to sample from $\pi(\cdot \mid Y)$ is presented as Algorithm 2. In our simulations, the SPLA is compared to two different versions of the Langevin algorithm. In the

Stochastic Subgradient Langevin Algorithm (SSLA) [18], stochastic subgradients of g_{e_i} are used instead of stochastic proximity operators. In the Proximal Langevin Algorithm (ProxLA) [18], the full proximity operator of $\sum_{i=1}^{n} G_i$ is computed using a subroutine. As mentioned in [36, 47], we use the gradient algorithm for the dual problem. The plots in Figure 2 provide simulations of the algorithms on our machine (using one thread of a 2,800 MHz CPU and 256GB RAM). Additional numerical experiments are available in the Appendix. Four real life graphs from the dataset [25] are considered : the Facebook graph (4,039 nodes and 88,234 edges, extracted from the Facebook social network), the Youtube graph (1,134,890 nodes and 2,987,624 edges, extracted from the social network included in the Youtube website), the Amazon graph (the 334,863 nodes represent products linked by and 925,872 edges) and the DBLP graph (a co-authorship network of 317,080 nodes and 1,049,866 edges). On the larger graphs, we only simulate SPLA and SSLA since the computation of a full proximity operator becomes prohibitive. Numerical experiments over the Amazon and the DBLP graphs are available in the Supplementary material.
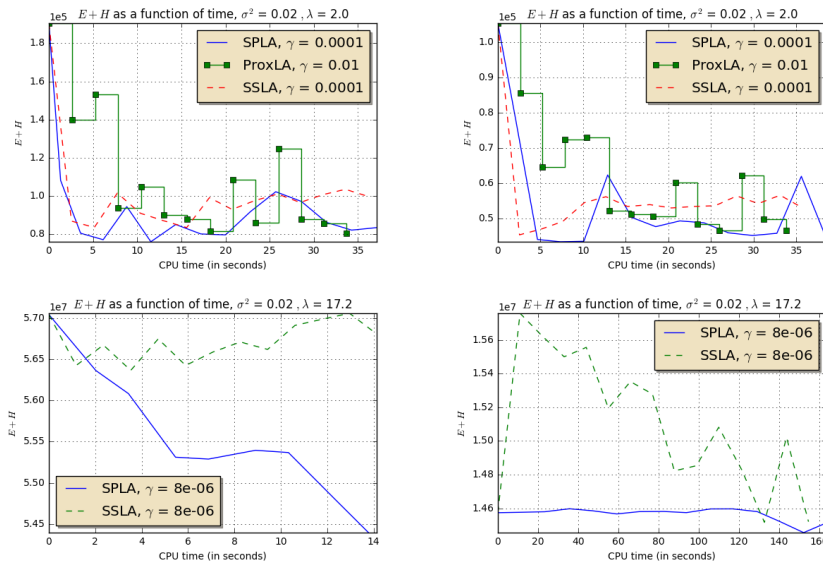

Figure 2: **Top row:** The functional $\mathcal{F} = \mathcal{H} + \mathcal{E}_U$ as a function of CPU time for the three algorithms over the Facebook graph. Left: $Y \sim N(0, I)$. Right: $Y \sim N(0, I)$ and then half of the coordinates of Y are put to zero. **Bottom row:** The functional $\mathcal{F} = \mathcal{H} + \mathcal{E}_U$ as a function of CPU time for the two algorithms over the Youtube graph. Left: $Y \sim N(0, I)$. Right: $Y \sim N(0, I)$ and then half of the coordinates of Y are put to zero.

In our simulations, we represent the functional $\mathcal{F} = \mathcal{H} + \mathcal{E}_U$ as a function of CPU time while running the algorithms. The parameters λ and σ are chosen such that the log likelihood term and the Total Variation regularization term have the same weight. The functionals $\mathcal{H}$ and $\mathcal{E}_U$ are estimated using five random realizations of each iterate $\hat{x}^k$ ($\mathcal{H}$ is estimated using a kernel density estimator). The batch parameter n is equal to 400. We consider cases where Y has a standard gaussian distribution and cases where half of the components of Y are standard gaussians and half are equal to zero (this correspond to the graph inpainting task [11]). SPLA and SSLA are always simulated with the same step size.

As expected, the numerical experiments show the advantage of using stochastic proximity operators instead of stochastic subgradients. It is a standard fact that proximity operators are better than subgradients to tackle ℓ^1-norm terms [2]. Our figures show that stochastic proximity operators are numerically more stable than alternatives [43]. Our figures also show the advantage of stochastic methods (SSLA or SPLA) over deterministic ones for large scale problems. The SSLA and the SPLA provide iterates about one hundred times more frequently than ProxLA, and are faster in the first iterations.

9

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] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.

[3] E. Bernton. Langevin Monte Carlo and JKO splitting. *arXiv preprint arXiv:1802.08671*, 2018.

[4] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[5] 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] 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.

[7] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.

[8] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9:263–340, 2010.

[9] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[10] N. S Chatterji, N. Flammarion, Y.-A. Ma, P. L Bartlett, and M. I Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*, 2018.

[11] S. Chen, A. Sandryhaila, G. Lederman, Z. Wang, J. MF Moura, P. Rizzo, J. Bielak, J. H Garrett, and J. Kovačević. Signal inpainting on graphs via total variation minimization. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 8267–8271, 2014.

[12] X. Cheng, N. S Chatterji, P. L Bartlett, and M. I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[13] L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[14] 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.

[15] A. S Dalalyan and A. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[16] A. S Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

[17] D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017.

[18] 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.

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

[20] 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.

[21] E. Hazan, K. Y Levy, and S. Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *International conference on machine learning*, pages 1833–1841, 2016.

[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] S. Kotz, T. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

[24] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[25] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[26] 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.

[27] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[28] R. M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[29] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[30] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

[31] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization. *Journal of Machine Learning Research*, 18:1–42, 2018.

[32] P. Richtárik and M. Takáč. Stochastic reformulations of linear systems: algorithms and convergence theory. *arXiv:1706.01108*, 2017.

[33] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[34] R. T. Rockafellar and R. J.-B. Wets. On the interchange of subdifferentiation and conditional expectations for convex functionals. *Stochastics*, 7(3):173–182, 1982.

[35] L. Rosasco, S. Villa, and B. C. Vũ. Stochastic inertial primal-dual algorithms. *arXiv preprint arXiv:1507.00852*, 2015.

[36] A. Salim, P. Bianchi, and W. Hachem. Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs. *IEEE Transactions on Automatic Control*, 2019.

[37] F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[38] S. Schechtman, A. Salim, and P. Bianchi. Passty Langevin. In *CAp*, 2019.

[39] U. ŞimŠekli. Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov Chain Monte Carlo. In *International Conference on Machine Learning*, pages 3200–3209, 2017.

[40] P. Sollich. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine learning*, 46(1-3):21–52, 2002.

[41] W. Tansey and J. G Scott. A fast and flexible algorithm for the graph-fused lasso. *arXiv preprint arXiv:1505.06475*, 2015.

[42] R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

[43] P. Toulis, T. Horel, and E. M Airoldi. Stable Robbins-Monro approximations through stochastic proximal updates. *arXiv preprint arXiv:1510.00967*, 2015.

[44] T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[45] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[46] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Applied Mathematics*, 38(3):667–681, 2013.

[47] Y.-X. Wang, J. Sharpnack, A. J Smola, and R. J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

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

[49] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. *arXiv preprint arXiv:1802.08089*, 2018.

[50] X. Xu and M. Ghosh. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

[51] A. Yurtsever, B. C. Vu, and V. Cevher. Stochastic three-composite convex minimization. In *Advances in Neural Information Processing Systems*, pages 4329–4337, 2016.

[52] D. Zou, P. Xu, and Q. Gu. Subsampled stochastic variance-reduced gradient langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*, 2018.

[53] D. Zou, P. Xu, and Q. Gu. Sampling from non-log-concave distributions via variance-reduced gradient langevin dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2936–2945, 2019.

Appendix

Contents

A Lemmas Needed for the Proof of the Main Theorem

In order to prove Theorem 1, we will need to establish several lemmas. For every convex function $g : \mathbb{R}^d \to \mathbb{R}$, we denote $\mathcal{E}_g(\mu) = \int g \mathrm{d}\mu$. In the sequel, we assume that Assumptions 1–5 hold true.

We first recall [18, Lemma 1].

Lemma 5. The target distribution satisfies $\mu^\star \in \mathcal{P}_2(\mathbb{R}^d), \mathcal{E}_U(\mu^\star)$ and $\mathcal{H}(\mu^\star) < \infty$.

Moreover, if $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$\mathrm{KL}(\mu \mid \mu^\star) = \mathcal{E}_U(\mu) + \mathcal{H}(\mu) - (\mathcal{E}_U(\mu^\star) + \mathcal{H}(\mu^\star)) = \mathcal{F}(\mu) - \mathcal{F}(\mu^\star),$$

provided that $\mathcal{E}_U(\mu) < \infty$.

Lemma 6.

$$2\gamma \left[\mathcal{H}(\mu_{y_0^k}) - \mathcal{H}(\mu^\star) \right] \le W^2(\mu_{z^k}, \mu^\star) - W^2(\mu_{y_0^k}, \mu^\star). \tag{9}$$

Proof. This is an application of [18, Lemma 5] with $\mu \leftarrow \mu_{z^k}$ and $\pi \leftarrow \mu^\star$.

The proof is the roughly the same as proving the $\mathcal{O}(1/k)$ convergence rate of the gradient descent algorithm, but in continuous time and in the space $\mathcal{P}_2(\mathbb{R}^d)$. For the sake of completeness, we provide the main arguments of a different proof than [18], that makes easier the connection with Lyapunov techniques used in the analysis of gradient descent/gradient flows in Euclidean spaces.

Consider Brownian motion (B_t) and the rescaled Brownian Motion $(\sqrt{2}B_t)$ initialized with $\sqrt{2}B_0 \sim \mu_{z^k}$ and denote, for every $t \ge 0$, ν_t the distribution of $\sqrt{2}B_t$. Then $(\nu_t)_t$ is a gradient flow of $\mathcal{H}$ (see [37]). This implies (see [1, Page 280])

$$\forall t > 0, \quad 2 \left(\mathcal{H}(\nu_t) - \mathcal{H}(\mu^\star) \right) \le - \frac{\mathrm{d}}{\mathrm{d}t} W^2(\nu_t, \mu^\star). \tag{10}$$

This also implies [45, Page 711] that (the objective function) $\mathcal{H}$ is a Lyapunov function for (the gradient flow) $(\nu_t)_{t \ge 0}$:

$$\forall t > 0, \quad \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{H}(\nu_t) \le 0. \tag{11}$$

Now consider the function $\ell(t) = t \left(\mathcal{H}(\nu_t) - \mathcal{H}(\mu^\star) \right) + \frac{1}{2} W^2(\nu_t, \mu^\star)$. For every $t > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \ell(t) = \left(\mathcal{H}(\nu_t) - \mathcal{H}(\mu^\star) \right) + t \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{H}(\nu_t) - \left(\mathcal{H}(\nu_t) - \mathcal{H}(\mu^\star) \right) \le 0,$$

using the inequalities (A) and (A). In other words, ℓ is also a Lyapunov function. Therefore, for every $\varepsilon > 0, \ell(\gamma) \le \ell(\varepsilon)$, which implies[3] $\ell(\gamma) \le \ell(0)$ *i.e*,

$$\gamma \left(\mathcal{H}(\nu_\gamma) - \mathcal{H}(\mu^\star) \right) \le \frac{1}{2} W^2(\nu_0, \mu^\star) - \frac{1}{2} W^2(\nu_\gamma, \mu^\star).$$

It remains to note that $\nu_0 = \mu_{z^k}$ and $\nu_\gamma = \mu_{y_0^k}$. $\qquad\square$

Lemma 7.

$$2\gamma \left[\mathcal{E}_F(\mu_{y_0^k}) - \mathcal{E}_F(\mu_{z^k}) \right] \le 2\gamma^2 L d . \tag{12}$$

Proof. This is an application of [18, Lemma 3] with $\mu \leftarrow \mu_{z^k}$. We provide the proof for the sake of completeness. It can be related to smoothing techniques used in optimization [21].

First, using the convexity and the smoothness of F,

$$0 \le F(y_0^k) - F(z^k) + \left\langle \nabla F(z^k), z^k - y_0^k \right\rangle \le \frac{L}{2} \| y_0^k - z^k \|^2. \tag{13}$$

Note that $y_0^k - z^k = \sqrt{2\gamma} W^k$ is independent of z^k, that $\mathbb{E}(y_0^k - z^k) = 0$ and that $\mathbb{E}(\| y_0^k - z^k \|^2) = 2\gamma d$. Taking the expectation in (A) gives the result. $\qquad\square$

---

[3] Here, one has to prove the ℓ is continuous at 0, we skip this part of the proof

Lemma 8. Let $\gamma \leq 1/L$. Then

$$2\gamma \left[\mathcal{E}_F(\mu_{z^k}) - \mathcal{E}_F(\mu^\star)\right] \leq (1 - \gamma\alpha)W^2(\mu_{x^k}, \mu^\star) - W^2(\mu_{z^k}, \mu^\star) + 2\gamma^2\sigma_F^2. \tag{14}$$

Proof. We choose arbitrary $a \in \mathbb{R}^d$ and start with an upper-bound for $\left\|z^k - a\right\|^2$:

$$\left\|z^k - a\right\|^2 = \left\|x^k - a - \gamma\nabla f(x^k, \xi^k)\right\|^2$$
$$= \left\|x^k - a\right\|^2 + 2\gamma\left\langle\nabla f(x^k, \xi^k), a - x_k\right\rangle + \gamma^2\left\|\nabla f(x^k, \xi^k)\right\|^2.$$

By taking an expectation with respect to ξ^k we get

$$\mathbb{E}_{\xi^k}\left[\left\|z^k - a\right\|^2\right] = \left\|x^k - a\right\|^2 + 2\gamma\left\langle\nabla F(x^k), a - x^k\right\rangle + \gamma^2\left\|\nabla F(x^k)\right\|^2$$
$$+ \gamma^2\mathbb{E}_{\xi^k}\left[\left\|\nabla F(x^k) - \nabla f(x^k, \xi^k)\right\|^2\right]$$
$$\leq \left\|x^k - a\right\|^2 + 2\gamma\left(F(a) - F(x^k) - \frac{\alpha}{2}\left\|x^k - a\right\|^2\right) + \gamma^2\left\|\nabla F(x^k)\right\|^2 + \gamma^2\sigma_F^2$$

$$= (1 - \gamma\alpha)\left\|x^k - a\right\|^2 + 2\gamma\left(F(a) - F(x^k)\right) + \gamma^2\left\|\nabla F(x^k)\right\|^2 + \gamma^2\sigma_F^2. \tag{15}$$

Next, we use the L-smoothness of F and the fact that $\gamma \leq 1/L$:

$$\mathbb{E}_{\xi^k}F(z^k) \leq F(x^k) + \mathbb{E}_{\xi^k}\left[\left\langle\nabla F(x^k), z^k - x^k\right\rangle\right] + \frac{L}{2}\mathbb{E}_{\xi^k}\left[\left\|z^k - x^k\right\|^2\right]$$
$$= F(x^k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\left\|\nabla F(x^k)\right\|^2 + \frac{L}{2}\gamma^2\mathbb{E}_{\xi^k}\left[\left\|\nabla f(x^k, \xi^k)\right\|^2 - \left\|\nabla F(x^k)\right\|^2\right]$$
$$\leq F(x^k) - \frac{\gamma}{2}\left\|\nabla F(x^k)\right\|^2 + \frac{L}{2}\gamma^2\sigma_F^2.$$

Since $L\gamma \leq 1$,

$$\mathbb{E}_{\xi^k}F(z^k) \leq F(x^k) - \frac{\gamma}{2}\left\|\nabla F(x^k)\right\|^2 + \frac{\gamma}{2}\sigma_F^2,$$

which gives us an upper-bound for $\gamma^2\left\|\nabla F(x^k)\right\|^2$:

$$\gamma^2\left\|\nabla F(x^k)\right\|^2 \leq 2\gamma\left(F(x^k) - \mathbb{E}_{\xi^k}F(z^k)\right) + \gamma^2\sigma_F^2,$$

since $\gamma \leq 1/L$. Plugging this into (A) we obtain

$$\mathbb{E}_{\xi^k}\left[\left\|z^k - a\right\|^2\right] \leq (1 - \gamma\alpha)\left\|x^k - a\right\|^2 + 2\gamma\left(F(a) - \mathbb{E}_{\xi^k}F(z^k)\right) + 2\gamma^2\sigma_F^2. \tag{16}$$

Now, let $a \sim \mu^\star$, i.e. a is a random vector sampled from the distribution with density $\mu^\star$. By taking the full expectation in (A) we get

$$\mathbb{E}\left[\left\|z^k - a\right\|^2\right] \leq (1 - \gamma\alpha)\mathbb{E}\left[\left\|x^k - a\right\|^2\right] + 2\gamma\left[\mathcal{E}_F(\mu^\star) - \mathcal{E}_F(\mu_{z^k})\right] + 2\gamma^2\sigma_F^2.$$

Using the definition of Wasserstein distance we get

$$W^2(\mu_{z^k}, \mu^\star) \leq (1 - \gamma\alpha)\mathbb{E}\left[\left\|x^k - a\right\|^2\right] + 2\gamma\left[\mathcal{E}_F(\mu^\star) - \mathcal{E}_F(\mu_{z^k})\right] + 2\gamma^2\sigma_F^2$$

Note that in the last inequality, x^k can be replaced by any random variable with distribution μ_{x^k}. Taking the $\inf$ over all possible couplings of μ_{x^k} and $\mu^\star$ we get

$$W^2(\mu_{z^k}, \mu^\star) \leq (1 - \gamma\alpha)W^2(\mu_{x^k}, \mu^\star) + 2\gamma\left[\mathcal{E}_F(\mu^\star) - \mathcal{E}_F(\mu_{z^k})\right] + 2\gamma^2\sigma_F^2.$$

$\square$

Remark 1. We now recall standard facts from convex analysis that will be used without mention in the sequel. These results can be found in [2] or [7]. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The Moreau envelope g^γ of g is defined by

$$g^\gamma(x) = \min_{y \in \mathbb{R}^d} g(y) + \frac{1}{2\gamma}\left\|y - x\right\|^2,$$

and is a $1/\gamma$-smooth convex function such that $g^\gamma(x) \leq g(x)$ and $g^\gamma(x) \to_{\gamma \to 0} g(x)$ for every $x \in \mathbb{R}^d$. The proximity operator of g and the Moreau envelope are linked through their definitions

$$g^\gamma(x) = \frac{1}{2\gamma} \left\| \text{prox}_{\gamma g}(x) - x \right\|^2 + g(\text{prox}_{\gamma g}(x)),$$

but also through the relation

$$\text{prox}_{\gamma g}(x) = x - \gamma \nabla g^\gamma(x).$$

The function ∇g^γ is called the Yosida approximation of ∂g. The proximity operator $\text{prox}_{\gamma g}$ is 1-Lipschitz continuous, and so is $\gamma \nabla g^\gamma$. The Yosida approximation satisfies moreover

$$\nabla g^\gamma(x) \in \partial g \left(\text{prox}_{\gamma g}(x) \right),$$

for every $x \in \mathbb{R}^d$. Since g only takes finite values, for every $x \in \mathbb{R}^d$, $\partial g(x) \neq \emptyset$. Furthermore, the Yosida approximation satisfies for every $x \in \mathbb{R}^d$,

$$\left\| \nabla g^\gamma(x) \right\| \leq \left\| \nabla^0 g(x) \right\|.$$

Lemma 9. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then,

$$g^\gamma(x) \geq g(x) - \frac{\gamma}{2} \left\| \nabla^0 g(x) \right\|^2.$$

Proof. Let $x \in \mathbb{R}^d$. Using the convexity of g we have for every $y \in \mathbb{R}^d$,

$$g(y) + \frac{1}{2\gamma} \left\| y - x \right\|^2 \geq g(x) + \left\langle \nabla^0 g(x), y - x \right\rangle + \frac{1}{2\gamma} \left\| y - x \right\|^2. \tag{17}$$

We conclude the proof by taking the minimum over y on both sides of (A). $\square$

Lemma 10. For every $i \in \{1, \ldots, n\}$, let $g_i : \mathbb{R}^d \to \mathbb{R}$ a convex function. Consider $a, y_0, y_1, \ldots, y_n \in \mathbb{R}^d$ such that for every $k = 1, \ldots, n$, $y_k = \text{prox}_{\gamma g_k}(y_{k-1})$. Then,

$$\left\| y_n - a \right\|^2 \leq \left\| y_0 - a \right\|^2 - 2\gamma \sum_{k=1}^{n} (g_k^\gamma(y_{k-1}) - g_k(a)).$$

Proof. Iterating Equation 1, we have for every $i \in \{1, \ldots, n\}$:

$$y_i = y_0 - \gamma \sum_{j=1}^{i} \nabla g_j^\gamma(y_{j-1}).$$

Therefore,

$$\left\| y_n - a \right\|^2 = \left\| y_0 - a \right\|^2 - 2\gamma \left\langle \sum_{i=1}^{n} \nabla g_i^\gamma(y_{i-1}), y_0 - a \right\rangle + \gamma^2 \left\| \sum_{i=1}^{n} \nabla g_i^\gamma(y_{i-1}) \right\|^2. \tag{18}$$

Since $\nabla g_i^\gamma(y_{i-1}) \in \partial g_i(y_i)$,

$$\left\langle \nabla g_i^\gamma(y_{i-1}), y_0 - a \right\rangle = \left\langle \nabla g_i^\gamma(y_{i-1}), y_i - a \right\rangle + \gamma \left\| \nabla g_i^\gamma(y_{i-1}) \right\|^2 + \gamma \left\langle \nabla g_i^\gamma(y_{i-1}), \sum_{j=1}^{i-1} \nabla g_j^\gamma(y_{j-1}) \right\rangle$$

$$\geq g_i(y_i) - g_i(a) + \gamma \left\| \nabla g_i^\gamma(y_{i-1}) \right\|^2 + \gamma \left\langle \nabla g_i^\gamma(y_{i-1}), \sum_{j=1}^{i-1} \nabla g_j^\gamma(y_{j-1}) \right\rangle.$$

16

Furthermore,

$$
-2\gamma\left\langle\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1}), y_0-a\right\rangle+\gamma^2\left\|\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1})\right\|^2
$$

$$
=-2\gamma\sum_{i=1}^{n}(g_i(y_i)-g_i(a))-2\gamma^2\sum_{i=1}^{n}\|\nabla g_i^\gamma(y_{i-1})\|^2
$$

$$
+\gamma^2\left\|\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1})\right\|^2-2\gamma^2\sum_{i=1}^{n}\left\langle\nabla g_i^\gamma(y_{i-1}),\sum_{j=1}^{i-1}\nabla g_j^\gamma(y_{j-1})\right\rangle.
$$

Expanding the square norm $\gamma^2\|\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1})\|^2$, all the cross products vanish with the last term. It remains only

$$
-2\gamma\left\langle\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1}), y_0-a\right\rangle+\gamma^2\left\|\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1})\right\|^2
$$

$$
=-2\gamma\sum_{i=1}^{n}(g_i(y_i)-g_i(a))-\gamma^2\sum_{i=1}^{n}\|\nabla g_i^\gamma(y_{i-1})\|^2.
$$

Since $g_i^\gamma(y_{i-1})=\frac{1}{2\gamma}\|y_i-y_{i-1}\|^2+g_i(y_i)$,

$$
-2\gamma\left\langle\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1}), y_0-a\right\rangle+\gamma^2\left\|\sum_{i=1}^{n}\nabla g_i^\gamma(y_{i-1})\right\|^2=-2\gamma\sum_{i=1}^{n}(g_i^\gamma(y_{i-1})-g_i(a)).
$$

Plugging the last equation into (A) gives the result. $\qquad\square$

Lemma 11. There exists $C\geq 0$ which can be expressed as a linear combination of $L_{G_1}^2,\ldots,L_{G_n}^2$ with integer coefficients such that

$$
2\gamma\sum_{i=1}^{n}\left[\mathcal{E}_{G_i}(\mu_{y_0^k})-\mathcal{E}_{G_i}(\mu^\star)\right]\leq W^2(\mu_{y_0^k},\mu^\star)-W^2(\mu_{x^{k+1}},\mu^\star)+\gamma^2 C. \tag{19}
$$

Moreover, if, for every $i\in\{2,\ldots,n\}$, $g_i(\cdot,\xi)$ admits almost surely the representation $g_i(\cdot,\xi)=\tilde{g}_i(\cdot,\xi_i)$ where $\xi_2,\ldots,\xi_n$ are independent random variables, then one can set $C:=n\sum_{i=1}^{n}L_{G_i}^2$.

Proof. Using the convexity of $g_i^\gamma(\cdot,\xi^k)$ and Lemma 9,

$$
g_i^\gamma(y_{i-1}^k,\xi^k)\geq g_i^\gamma(y_0^k,\xi^k)+\left\langle\nabla g_i^\gamma(y_0^k,\xi^k), y_{i-1}^k-y_0^k\right\rangle
$$

$$
\geq g_i(y_0^k,\xi^k)-\frac{\gamma}{2}\|\nabla^0 g_i(y_0^k,\xi^k)\|^2+\left\langle\nabla g_i^\gamma(y_0^k,\xi^k), y_{i-1}^k-y_0^k\right\rangle
$$

We now look at the last term at the right hand side. If $i=1$ it is equal to zero. If $i\geq 2$, using Young's inequality,

$$
-2\gamma\left\langle\nabla g_i^\gamma(y_0^k,\xi^k), y_{i-1}^k-y_0^k\right\rangle=\sum_{j=1}^{i-1}-2\left\langle\gamma\nabla g_i^\gamma(y_0^k,\xi^k), y_j^k-y_{j-1}^k\right\rangle
$$

$$
\leq(i-1)\|\gamma\nabla g_i^\gamma(y_0^k,\xi^k)\|^2+\sum_{j=1}^{i-1}\|y_j^k-y_{j-1}^k\|^2.
$$

Combining the two last inequalities with Lemma 10 applied to $y_i\leftarrow y_i^k$ and $g_i\leftarrow g_i(\cdot,\xi^k)$, we have

$$
\|y_n^k-a\|^2\leq\|y_0^k-a\|^2 \tag{20}
$$

$$
-2\gamma\sum_{i=1}^{n}(g_i(y_0^k,\xi^k)-g_i(a,\xi^k))
$$

$$
+\gamma^2\sum_{i=1}^{n}\|\nabla^0 g_i(y_0^k,\xi^k)\|^2+\sum_{i=2}^{n}\left((i-1)\|\gamma\nabla g_i^\gamma(y_0^k,\xi^k)\|^2+\sum_{j=1}^{i-1}\|y_j^k-y_{j-1}^k\|^2\right).
$$

17

We now consider two cases. First, assume that for every $i \in \{2, \ldots, n\}$, $g_i(\cdot, \xi)$ admits almost surely the representation $g_i(\cdot, \xi) = \tilde{g}_i(\cdot, \xi_i)$ where $\xi_2, \ldots, \xi_n$ are independent random variables. In this case, ξ_j^k – the kth i.i.d copy of ξ_j – is independent of y_{j-1}^k for every $j \in \{1, \ldots, n\}$. Noting that $y_j^k - y_{j-1}^k = -\gamma \nabla \tilde{g}_j{}^\gamma(y_{j-1}^k, \xi_j^k)$ and using Assumption 5

$$\mathbb{E}\left(\left\|\nabla \tilde{g}_j{}^\gamma(y_{j-1}^k, \xi_j^k)\right\|^2 \mid y_{j-1}^k\right) = \mathbb{E}\left(\left\|\nabla \tilde{g}_j{}^\gamma(\cdot, \xi_j^k)\right\|^2\right)(y_{j-1}^k) = \mathbb{E}\left(\left\|\nabla g_j^\gamma(\cdot, \xi^k)\right\|^2\right)(y_{j-1}^k) \leq L_{G_j}^2.$$

Taking a full expectation,

$$\gamma^2 \mathbb{E}\left(\sum_{i=1}^n \left\|\nabla^0 \tilde{g}_i(y_0^k, \xi_i^k)\right\|^2 + \left((i-1)\left\|\gamma \nabla \tilde{g}_i{}^\gamma(y_0^k, \xi_i^k)\right\|^2 + \sum_{j=1}^{i-1}\left\|y_j^k - y_{j-1}^k\right\|^2\right)\right) \leq \gamma^2 n \sum_{i=1}^n L_{G_i}^2. \tag{21}$$

In this case we set $C := n \sum_{i=1}^n L_{G_i}^2$. Obviously, this cover the case $n = 2$.

Second (general case), denote $\beta_j = \mathbb{E}\left\|y_j^k - y_{j-1}^k\right\|^2$. We write

$$\nabla g_j^\gamma(y_{j-1}^k, \xi^k) = \nabla g_j^\gamma(y_0^k, \xi^k) + \left(\nabla g_j^\gamma(y_{j-1}^k, \xi^k) - \nabla g_j^\gamma(y_{j-2}^k, \xi^k)\right) + \ldots + \left(\nabla g_j^\gamma(y_1^k, \xi^k) - \nabla g_j^\gamma(y_0^k, \xi^k)\right).$$

Using Young's inequality and the fact that $\gamma \nabla g_j^\gamma(\cdot, \xi^k)$ is 1-Lipschitz continous,

$$\beta_j \leq j\left(L_{G_j}^2 + \beta_{j-1} + \ldots + \beta_1\right).$$

Noting that $\beta_1 \leq L_{G_1}^2$, it is easy to prove (by induction) that there exists a linear combination of the $L_{G_1}^2, \ldots, L_{G_n}^2$ with integer coefficients denoted $C \geq 0$ such that

$$\gamma^2 \mathbb{E}\left(\sum_{i=1}^n \left\|\nabla^0 g_i(y_0^k, \xi^k)\right\|^2 + \left((i-1)\left\|\gamma \nabla g_i^\gamma(y_0^k, \xi^k)\right\|^2 + \sum_{j=1}^{i-1}\left\|y_j^k - y_{j-1}^k\right\|^2\right)\right) \leq \gamma^2 C. \tag{22}$$

Finally, taking the expectation in (A) and plugging (A) or (A),

$$\mathbb{E}\left\|y_n^k - a\right\|^2 \leq \mathbb{E}\left\|y_0^k - a\right\|^2$$
$$- 2\gamma \sum_{i=1}^n \left(\mathbb{E}(G_i(y_0^k)) - \mathbb{E}(G_i(a))\right) + \gamma^2 C.$$

Using the definition of $\mathcal{E}_{G_i}$ and taking the inf over all couplings y_0^k, a of $\mu_{y_0^k}, \mu^\star$, we get

$$W^2(\mu_{x^{k+1}}, \mu^\star) = W^2(\mu_{y_n^k}, \mu^\star) \leq W^2(\mu_{y_0^k}, \mu^\star) + 2\gamma \sum_{i=1}^n \left[\mathcal{E}_{G_i}(\mu^\star) - \mathcal{E}_{G_i}(\mu_{y_0^k})\right] + \gamma^2 C.$$

$\square$

B Proof of Theorem 1

By summing up (8) and (7) we get

$$2\gamma \left[\mathcal{E}_F(\mu_{y_0^k}) - \mathcal{E}_F(\mu^\star) \right] \le (1 - \gamma\alpha) W^2(\mu_{x^k}, \mu^\star) - W^2(\mu_{z^k}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld).$$

Adding (6) leads to

$$2\gamma \left[\mathcal{E}_F(\mu_{y_0^k}) - \mathcal{E}_F(\mu^\star) + \mathcal{H}(\mu_{y_0^k}) - \mathcal{H}(\mu^\star) \right] \le (1 - \gamma\alpha) W^2(\mu_{x^k}, \mu^\star) - W^2(\mu_{y_0^k}, \mu^\star)$$
$$+ \gamma^2(2\sigma_F^2 + 2Ld).$$

Finally, by adding (11) we get

$$2\gamma \left[\mathcal{E}_F(\mu_{y_0^k}) - \mathcal{E}_F(\mu^\star) + \mathcal{H}(\mu_{y_0^k}) - \mathcal{H}(\mu^\star) + \sum_{i=1}^{n} \left(\mathcal{E}_{G_i}(\mu_{y_0^k}) - \mathcal{E}_{G_i}(\mu^\star) \right) \right]$$
$$\le (1 - \gamma\alpha) W^2(\mu_{x^k}, \mu^\star) - W^2(\mu_{x^{k+1}}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld + C),$$

which, along with Lemma 5, concludes the proof.

C   Proof of Corollary 2

From (3), for all $j = 0, \ldots, k$ we get

$$2\gamma \left[\mathcal{F}(\mu_{y_0^j}) - \mathcal{F}(\mu^\star) \right] \leq W^2(\mu_{x^j}, \mu^\star) - W^2(\mu_{x^{j+1}}, \mu^\star) + \gamma^2 (2\sigma_F^2 + 2Ld + C). \qquad (23)$$

Summing up (C) for $j = 0, \ldots, k$ leads to

$$2\gamma \sum_{j=0}^k \left[\mathcal{F}(\mu_{y_0^j}) - \mathcal{F}(\mu^\star) \right] \leq W^2(\mu_{x^0}, \mu^\star) - W^2(\mu_{x^{k+1}}, \mu^\star) + \gamma^2 (k+1)(2\sigma_F^2 + 2Ld + C)$$

$$\leq W^2(\mu_{x^0}, \mu^\star) + \gamma^2 (k+1)(2\sigma_F^2 + 2Ld + C).$$

Using Lemma 5 and the convexity of KL divergence [44, Theorem 11], $\mathcal{F}$ is convex on $\mathcal{P}_2(\mathbb{R}^d)$. Since $\mu_{\hat{x}^k} = \frac{1}{k+1} \sum_{j=0}^k \mu_{y_0^j}$,

$$\mathcal{F}(\mu_{\hat{x}^k}) \leq \frac{1}{k+1} \sum_{j=0}^k \mathcal{F}(\mu_{y_0^j}),$$

hence

$$\mathcal{F}(\mu_{\hat{x}^k}) - \mathcal{F}(\mu^\star) \leq \frac{1}{2\gamma(k+1)} W^2(\mu_{x^0}, \mu^\star) + \frac{\gamma}{2}(2\sigma_F^2 + 2Ld + C).$$

Hence, given any $\varepsilon > 0$, choosing stepsize $\gamma = \min \left\{ \frac{1}{L}, \frac{\varepsilon}{2\sigma_F^2 + 2Ld + C} \right\}$ leads to

$$\frac{\gamma}{2}(2\sigma_F^2 + 2Ld + C) \leq \frac{\varepsilon}{2}.$$

If the number of iterations is

$$k + 1 \geq \max \left\{ \frac{L}{\varepsilon}, \frac{2\sigma_F^2 + 2Ld + C}{\varepsilon^2} \right\} W^2(\mu_{x^0}, \mu^\star),$$

then,

$$\frac{1}{2\gamma(k+1)} W^2(\mu_{x^0}, \mu^\star) \leq \frac{\varepsilon}{2}.$$

This implies $\mathcal{F}(\mu_{\hat{x}^k}) - \mathcal{F}(\mu^\star) \leq \varepsilon$, and the proof is concluded by applying Lemma 5.

D Proof of Corollary 3

From (5), $\mathcal{F}(\mu_{y_0^j}) \geq \mathcal{F}(\mu^\star)$. From (3), for all $j = 0, \ldots, k-1$ we get

$$W^2(\mu_{x^{j+1}}, \mu^\star) \leq (1 - \gamma\alpha)W^2(\mu_{x^j}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld + C).$$

After unrolling this recurrence we get

$$W^2(\mu_{x^k}, \mu^\star) \leq (1 - \gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld + C)\sum_{j=0}^{k-1}(1 - \gamma\alpha)^j$$

$$= (1 - \gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld + C)\frac{1 - (1 - \gamma\alpha)^k}{\gamma\alpha}$$

$$\leq (1 - \gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{\alpha}.$$

The first part is proven. Setting $\gamma = \min\left\{\frac{1}{L}, \frac{\varepsilon\alpha}{2(2\sigma_F^2 + 2Ld + C)}\right\}$ gives

$$W^2(\mu_{x^k}, \mu^\star) \leq (1 - \gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) + \frac{\varepsilon}{2}.$$

If

$$k \geq \frac{1}{\gamma\alpha} \log\left(\frac{2W^2(\mu_{x^0}, \mu^\star)}{\varepsilon}\right),$$

then,

$$(1 - \gamma\alpha)^k W^2(\mu_{x^0}, \mu^\star) \leq \frac{\varepsilon}{2},$$

which concludes the proof.

E Proof of Corollary 4

From (3), for all $j = 0, \ldots, k$ we get

$$2\gamma \left[\mathcal{F}(\mu_{y_0^j}) - \mathcal{F}(\mu^\star) \right] \leq (1 - \gamma\alpha)W^2(\mu_{x^j}, \mu^\star) - W^2(\mu_{x^{j+1}}, \mu^\star) + \gamma^2(2\sigma_F^2 + 2Ld + C). \quad (24)$$

By dividing (E) by $(1 - \gamma\alpha)^j$ we get

$$\frac{2\gamma}{(1 - \gamma\alpha)^j} \left[\mathcal{F}(\mu_{y_0^j}) - \mathcal{F}(\mu^\star) \right] \leq \frac{W^2(\mu_{x^j}, \mu^\star)}{(1 - \gamma\alpha)^{j-1}} - \frac{W^2(\mu_{x^{j+1}}, \mu^\star)}{(1 - \gamma\alpha)^j} + \frac{\gamma^2(2\sigma_F^2 + 2Ld + C)}{(1 - \gamma\alpha)^j}. \quad (25)$$

Summing up (E) for $j = 0, \ldots, k$ gives

$$\sum_{j=0}^{k} \frac{2\gamma}{(1 - \gamma\alpha)^j} \left[\mathcal{F}(\mu_{y_0^j}) - \mathcal{F}(\mu^\star) \right] \leq (1 - \gamma\alpha)W^2(\mu_{x^0}, \mu^\star) - \frac{W^2(\mu_{x^{k+1}}, \mu^\star)}{(1 - \gamma\alpha)^k} + \sum_{j=0}^{k} \frac{\gamma^2(2\sigma_F^2 + 2Ld + C)}{(1 - \gamma\alpha)^j}$$

$$\leq (1 - \alpha\gamma)W^2(\mu_{x^0}, \mu^\star) + \sum_{j=0}^{k} \frac{\gamma^2(2\sigma_F^2 + 2Ld + C)}{(1 - \gamma\alpha)^j}.$$

Using Lemma 5 and the convexity of KL divergence [44, Theorem 11], $\mathcal{F}$ is convex on $\mathcal{P}_2(\mathbb{R}^d)$. Since

$$\mu_{\tilde{x}^k} = \sum_{j=0}^{k} \frac{(1 - \gamma\alpha)^{-j}}{\sum_{r=0}^{k}(1 - \gamma\alpha)^{-r}} \mu_{x^j},$$

$$\mathcal{F}(\mu_{\tilde{x}^k}) \sum_{j=0}^{k}(1 - \gamma\alpha)^{-j} = \sum_{j=0}^{k}(1 - \gamma\alpha)^{-j}\mathcal{F}(\mu_{\tilde{x}^k}) \leq \sum_{j=0}^{k}(1 - \gamma\alpha)^{-j}\mathcal{F}(\mu_{x^j}),$$

hence

$$2\gamma \sum_{j=0}^{k}(1 - \gamma\alpha)^{-j} \left[\mathcal{F}(\mu_{\tilde{x}^k}) - \mathcal{F}(\mu^\star) \right] \leq (1 - \alpha\gamma)W^2(\mu_{x^0}, \mu^\star) + \sum_{j=0}^{k} \frac{\gamma^2(2\sigma_F^2 + 2Ld + C)}{(1 - \gamma\alpha)^j}.$$

After dividing by $2\gamma \sum_{j=0}^{k}(1 - \gamma\alpha)^{-j}$ we obtain

$$\mathcal{F}(\mu_{\tilde{x}^k}) - \mathcal{F}(\mu^\star) \leq \frac{W^2(\mu_{x^0}, \mu^\star)}{2\gamma \sum_{j=0}^{k}(1 - \gamma\alpha)^{-(j+1)}} + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{2}. \quad (26)$$

Now, we perform a simplification of the sum:

$$\sum_{j=0}^{k} \gamma(1 - \gamma\alpha)^{-(j+1)} = \frac{\gamma}{(1 - \gamma\alpha)} \sum_{j=0}^{k}(1 - \gamma\alpha)^{-j} = \frac{\gamma}{(1 - \gamma\alpha)} \cdot \frac{(1 - \gamma\alpha)^{-(k+1)} - 1}{(1 - \gamma\alpha)^{-1} - 1}$$

$$= \frac{(1 - \gamma\alpha)^{-(k+1)} - 1}{\alpha}.$$

Plugging this into (E) gives

$$\mathcal{F}(\mu_{\tilde{x}^k}) - \mathcal{F}(\mu^\star) \leq \frac{\alpha W^2(\mu_{x^0}, \mu^\star)}{2((1 - \gamma\alpha)^{-(k+1)} - 1)} + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{2}$$

$$= \alpha \left[\frac{W^2(\mu_{x^0}, \mu^\star)}{2} \cdot \frac{(1 - \gamma\alpha)^{k+1}}{1 - (1 - \gamma\alpha)^{k+1}} + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{2\alpha} \right].$$

If $\gamma = \min\left\{ \frac{1}{L}, \frac{\varepsilon\alpha}{2\sigma_F^2 + 2Ld + C} \right\}$ and

$$k \geq \max\left\{ \frac{L}{\alpha}, \frac{2\sigma_F^2 + 2Ld + C}{\varepsilon\alpha^2} \right\} \log\left(2\max\left\{ 1, \frac{W^2(\mu_{x^0}, \mu^\star)}{\varepsilon} \right\} \right),$$

then $k \geq \frac{1}{\gamma\alpha} \log 2$. Moreover, $(1 - \gamma\alpha)^{k+1} \leq 1/2$,

$$\mathcal{F}(\mu_{\tilde{x}^k}) - \mathcal{F}(\mu^\star) \leq \alpha \left[(1 - \gamma\alpha)^{k+1} W^2(\mu_{x^0}, \mu^\star) + \frac{\gamma(2\sigma_F^2 + 2Ld + C)}{2\alpha} \right],$$

and

$$\mathcal{F}(\mu_{\tilde{x}^k}) - \mathcal{F}(\mu^\star) \leq \alpha \left[(1 - \gamma\alpha)^{k+1} W^2(\mu_{x^0}, \mu^\star) + \frac{\varepsilon}{2} \right].$$

The conclusion follows from Lemma 5.

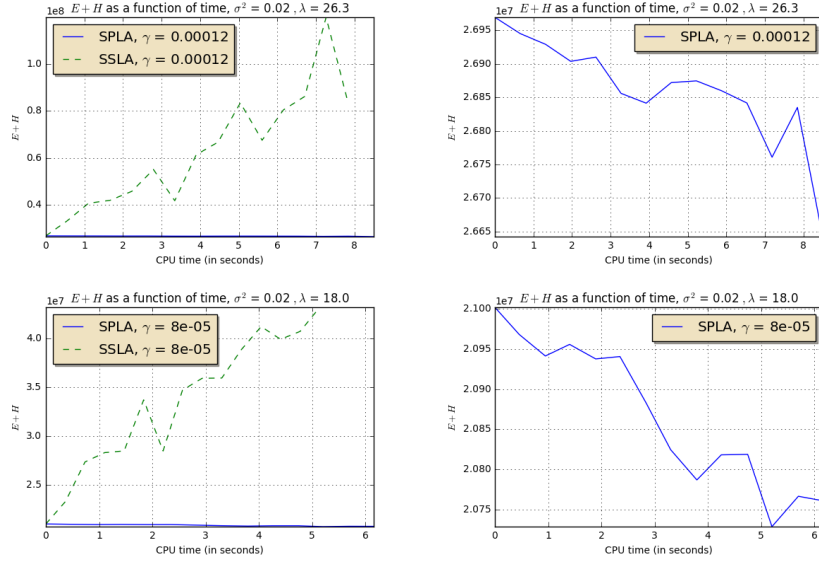## F   Additional Numerical Experiments

Figure 3: **Top row:** The functional $\mathcal{F} = \mathcal{H} + \mathcal{E}_U$ as a function of CPU time for the Amazon graph with $Y \sim N(0, I)$. Left: SSLA and SPLA. Right: Only SPLA. **Bottom row:** The functional $\mathcal{F} = \mathcal{H} + \mathcal{E}_U$ as a function of CPU time for the DBLP graph with $Y \sim N(0, I)$. Left: SSLA and SPLA. Right : Only SPLA.

These numerical experiments are conducted over the Amazon and the DBLP graphs. The left curves show the numerical stability of the proximal method (SPLA) with respect to the subgradient method (SSLA). The right curves are zoomed in view of the behavior of SPLA during the same experiments. It can be seen that SPLA still decreases the KL divergence.