Double-Loop Unadjusted Langevin Algorithm

Paul Rolland [1] **Armin Eftekhari** [2] **Ali Kavis** [1] **Volkan Cevher** [1]

Abstract

A well-known first-order method for sampling from log-concave probability distributions is the Unadjusted Langevin Algorithm (ULA). This work proposes a new annealing step-size schedule for ULA, which allows to prove new convergence guarantees for sampling from a smooth log-concave distribution, which are not covered by existing state-of-the-art convergence guarantees. To establish this result, we derive a new theoretical bound that relates the Wasserstein distance to total variation distance between any two log-concave distributions that complements the reach of Talagrand T_2 inequality. Moreover, applying this new step size schedule to an existing constrained sampling algorithm, we show state-of-the-art convergence rates for sampling from a constrained log-concave distribution, as well as improved dimension dependence.

1. Introduction

Let $\mathrm{d}\mu^*(x) \propto e^{-f(x)}\,\mathrm{d}x$ be a probability measure over $\mathbb{R}^d$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function with Lipschitz continuous gradient. In order to sample from such distributions, first-order sampling schemes based on the discretization of Langevin dynamics and, in particular the Unadjusted Langevin Algorithm (ULA), have found widespread success in various applications (Welling & Teh, 2011; Li et al., 2016b; Patterson & Teh, 2013; Li et al., 2016a).

An ever-growing body of literature has been devoted solely to the study of ULA and its variations (Ahn et al., 2012; Chen et al., 2015; Cheng & Bartlett, 2017; Cheng et al., 2017; Dalalyan & Karagulyan, 2017; Durmus et al., 2017; 2018a; Dwivedi et al., 2018; Luu et al., 2017; Welling & Teh, 2011; Ma et al., 2015). The ULA iterates are given as

$$x_{k+1} = x_k - \gamma_{k+1}\nabla f(x_k) + \sqrt{2\gamma_{k+1}}g_k, \qquad (1)$$

where ∇f is the gradient of f, $\{\gamma_k\}_{k\geq 0}$ is a non-increasing sequence of positive step-sizes, and the entries of $g_k \in \mathbb{R}^d$ are zero-mean and unit-variance Gaussian random variables, independent from each another and everything else. In its standard form (1), ULA can provably sample from any log-concave and smooth probability measure (Durmus et al., 2017; 2018a).

The recent analysis of (Durmus et al., 2018a) studies ULA through the lens of convex optimization. Their analysis shows strong resemblance with the convergence analysis of stochastic gradient descent (SGD) algorithm for minimizing a convex continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. Starting from $x_0 \in \mathbb{R}^d$, SGD iterates similarly as (1):

$$x_{k+1} = x_k - \gamma_{k+1}\nabla f(x_k) + \gamma_{k+1}\Theta(x_k),$$

where $\{\gamma_k\}_{k\geq 0}$ is a non-increasing sequence of positive step-sizes, and $\Theta : \mathbb{R}^d \to \mathbb{R}^d$ is a stochastic perturbation to ∇f. One way of proving convergence guarantees for this method is to show the following inequality:

$$2\gamma_{k+1}(\mathbb{E}[f(x_{k+1})] - f(x^*)) \leq \mathbb{E}\left[\|x_k - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{k+1} - x^*\|_2^2\right] + C\gamma_{k+1}^2 \qquad (2)$$

for some constant $C \geq 0$, $\forall k \geq 0$ and $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$. From this inequality, and using step size $\gamma_k \propto \frac{1}{\sqrt{k}}$, it is then possible to show convergence, in expectation, of the average iterate $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$ to the optimal value, i.e., $\mathbb{E}[f(\bar{x}_T)] - f(x^*) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$.

In their paper, (Durmus et al., 2018a) showed a similar descent lemma as (2) for the sequence of generated measures $\{\mu_k\}_{k\geq 0}$ denoting the distributions of the iterates $\{x_k\}_{k\geq 0}$ in (1), in which the objective gap $\mathbb{E}[f(x_k)] - f(x^*)$ is replaced with the Kullback-Leibler divergence $\mathrm{KL}(\mu_k;\mu^*)$, and the Euclidean distance $\|x_k - x^*\|_2$ is replaced with the 2-Wasserstein distance $\mathrm{W}_2(\mu_k,\mu^*)$:

$$2\gamma_{k+1}\,\mathrm{KL}(\mu_k;\mu^*) \leq \mathrm{W}_2^2(\mu_k,\mu^*) - \mathrm{W}_2^2(\mu_{k+1},\mu^*) + 2Ld\gamma_{k+1}^2, \qquad (3)$$

[1]LIONS, Ecole Polytechnique Fédérale de Lausanne, Switzerland [2]Department of Mathematics and Mathematical Statistics, Umea University, Sweden. Correspondence to: Paul Rolland <paul.rolland@epfl.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

where L is the Lipschitz constant of the gradient of f. Then again, using $\gamma_k \propto \frac{1}{\sqrt{k}}$, it is possible to show convergence of the average sample distribution $\bar{\mu}_T = \frac{1}{T}\sum_{t=0}^{T} \mu_t$ to μ^* in KL divergence, with rate $\mathcal{O}\left(\frac{d^3}{\sqrt{T}}\right)$.

In this work, we improve this convergence rate to $\mathcal{O}\left(\frac{d^3}{T^{\frac{2}{3}}}\right)$. To this end, we first establish a new bound that relates the W_2 distance and the KL divergence between any two log-concave distributions. When applied to inequality (3), this new bound can be exploited to design a new step-size sequence $\{\gamma_k\}_{k\geq 0}$ that allows to derive new convergence rates for ULA.

We introduce a new multistage decaying step size schedule, which proceeds in a double loop fashion by geometrically decreasing the step-size after a certain number of iterations, and that we call Double-loop ULA (DL-ULA). To the best of our knowledge, all existing convergence proof for ULA use either constant, or polynomially decaying step sizes, i.e. of the form $\gamma_k = k^{-\alpha}$ for some $\alpha \geq 0$, and this is the first work introducing a multistage decaying step size for a sampling algorithm. Interestingly, there is precedence to support our approach in that such step decay schedule can improve convergence of optimization algorithms (Hazan & Kale, 2014; Ge et al., 2019; Aybat et al., 2019; Yousefian et al., 2012).

Our new inequality relating KL divergence and W_2 distance serves as an alternative to the powerful T_2 inequality (Gozlan & Léonard, 2010), the latter requiring stronger assumptions on the distributions. The literature on Langevin dynamics commonly proves the convergence of an algorithm in KL divergence and then extends it to the total variation (TV) distance using the famous Pinsker's inequality (Pinsker, 1960; Cheng & Bartlett, 2017; Durmus et al., 2018a). Our new inequality enables to do the same for extending convergence results to W_2 distance in the case of general log-concave distributions, and hence, might be of independent interest. Note, however, that this inequality applied alone to extend the result of (Durmus et al., 2018a) to W_2 distance provides a suboptimal convergence rate, and modifying the step-size schedule and the analysis appears to be crucial for improving the rate.

Finally, we apply this multistage strategy to the constrained sampling algorithm MYULA (Brosse et al., 2017), which allows us to obtain improved convergence guarantees, both in terms of rate and dimension dependence. This approach provides state-of-the-art convergence guarantees for sampling from a log-concave distribution over a general convex set.

We summarize our contributions as follows:

- We introduce a variant of the Unadjusted Langevin Algorithm, using a new multistage decaying step-size schedule as well as a clipping step. Our new approach, called DL-ULA, yields new convergence guarantees, that are not covered by existing convergence result (i.e., either better convergence rate or better dimension dependence compared to state-of-the-art results).
- We apply our new step-size schedule to an existing Langevin-based constrained sampling algorithm, called MYULA (Brosse et al., 2017), and improve its convergence both in terms of iteration and dimension dependences.
- We introduce a new bound relating the 2-Wasserstein and the TV distance between any two log-concave distributions.

A summary of our convergence rates can be found in Tables 1 and 2.

Road map In section 3, we define several metrics on probability measures that we will use, recall some properties of log-concave distributions that we will exploit, as well as some results on convergence of ULA. In section 4, we present our new extension of ULA for unconstrained sampling, by introducing a new multistage step size schedule. We then prove convergence guarantees by making use of a new bound relating the KL divergence and the W_2 distance. Finally, in section 5, we apply this procedure to the existing algorithm MYULA for constrained sampling, and show that it yields improved convergence guarantees, both in terms of convergence rate and dimension dependence.

2. Related work

Unconstrained sampling Sampling algorithms based on Langevin dynamics have been widely studied (Ahn et al., 2012; Chen et al., 2015; Cheng & Bartlett, 2017; Cheng et al., 2017; Dalalyan & Karagulyan, 2017; Durmus et al., 2018a; Dwivedi et al., 2018; Durmus et al., 2017; Luu et al., 2017; Welling & Teh, 2011). Although most convergence rates have been established in the strongly log-concave setting, rates have also been shown for general log-concave distributions, and in particular exhibit larger dimension dependences (see Table 1).

Convergence guarantees for ULA applied to a general unconstrained log-concave distribution have been successively improved over the years. To the best of our knowledge, the best existing convergence results are the one obtained by (Durmus et al., 2018a) and (Durmus et al., 2017), that respectively show $\mathcal{O}(d^3\epsilon^{-4})$ and $\mathcal{O}(d^5\epsilon^{-2})$ convergence guarantees in TV distance. In this paper, we improve upon the former one, by showing a $\mathcal{O}(d^3\epsilon^{-3})$ convergence rate. This result is not absolutely better than the one of (Durmus et al., 2017), but enjoys better dimension dependence.

Until recently, convergence rate in Wasserstein distance had not been proven in the general log-concave setting. Only recently, (Zou et al., 2018) presented a method based on underdamped Langevin dynamics that provably converges in W_2-distance for a general log-concave distribution.

In (Zou et al., 2018), the authors show a $\mathcal{O}(d^{5.5}\epsilon^{-6})$ convergence rate in W_2 distance for general log-concave distributions. However, they make the assumption that $\mathbb{E}_{X\sim\mu}\left[\|X\|_2^4\right] \leq \bar{U}d^2$ for some scalar $\bar{U}$. However, let $\mathrm{d}\mu(x) \propto e^{-\|x\|_2}\,\mathrm{d}x$, which is a log-concave distribution. Then, $\mathbb{E}_{X\sim\mu}\left[\|X\|_2^4\right] = \Omega(d^4)$ and their assumption does not hold. For comparison purpose, if we replace it with our weaker Assumption 4, their rate becomes $\mathcal{O}(d^{10.5}\epsilon^{-6})$.

Constrained sampling Extensions of ULA have been designed in order to sample from constrained distributions (Bubeck et al., 2018; Brosse et al., 2017; Hsieh et al., 2018; Patterson & Teh, 2013). In (Bubeck et al., 2018), the authors propose to apply ULA, and project the sample onto the constraint at each iteration. They show a convergence rate of $\mathcal{O}(d^{12}\epsilon^{-12})$ in TV distance for log-concave distributions (i.e., $\mathcal{O}(d^{12}\epsilon^{-12})$ iterations of the algorithm are sufficient in order to obtain an error smaller than ϵ in TV distance).

In (Brosse et al., 2017), the authors propose to smooth the constraint using its Moreau-Yoshida envelope, and obtain a convergence rate of $\mathcal{O}(d^5\epsilon^{-6})$ in TV distance when the objective distribution is log-concave. To do so, they penalize the domain outside the constrain directly inside the target distribution via its Moreau-Yoshida envelop.

The analysis of MYULA in (Brosse et al., 2017) only holds when the penalty parameter is fixed and chosen in advance, leading to a natural saturation after a certain number of iterations. In this work, we extend this procedure using our Double-loop approach. This allows to obtain improved convergence both in terms of rate and dimension dependence, i.e., $\mathcal{O}(d^{3.5}\epsilon^{-5})$ in TV distance, and to ensure asymptotic convergence of the algorithm since the penalty is allowed to vary along the iterations.

The special case of sampling from simplices was solved in (Hsieh et al., 2018), introducing Mirrored Langevin Dynamics (MLD). Their work relies on finding a mirror map for the given constraint domain, and then performing ULA in the dual space. However, this method requires strong log-concavity of the distribution in the dual space. Moreover, finding a suitable mirror map for a general convex set is not an easy task.

3. Preliminaries

3.1. Various measures between distributions

Let us recall the distances/divergences between probability measures which are used frequently throughout the paper. The Kullback–Leibler (KL) divergence between two probability measures μ, ν on $\mathbb{R}^d$ is defined as

$$\mathrm{KL}(\mu;\nu) = \mathbb{E}_\mu \log(\mathrm{d}\mu/\,\mathrm{d}\nu), \tag{4}$$

assuming that μ is dominated by ν. Their Total Variation (TV) distance is defined as

$$\|\mu-\nu\|_{\mathrm{TV}} = \sup_S |\mu(S)-\nu(S)|, \tag{5}$$

where the supremum is over all measurable sets S of $\mathbb{R}^d$. Finally, the 2-Wasserstein (or W_2 for short) distance between μ and ν is defined as

$$\mathrm{W}_2^2(\mu,\nu) = \inf_{\phi\in\Phi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 d\gamma(x,y), \tag{6}$$

where $\Phi(\mu,\nu)$ denotes the set of all joint probability measures ϕ on $\mathbb{R}^{2d}$ that marginalize to μ and ν, namely, for all measurable sets $A, B \subseteq \mathbb{R}^d$, $\phi(A\times\mathbb{R}^d) = \mu(A)$ and $\phi(\mathbb{R}^d\times B) = \nu(B)$.

The main difference between W_2 and TV distances is that W_2 associates a higher cost when the difference between the distributions occurs at points that are further appart (in terms of Euclidean distance). Due to this property, errors occurring at the tail of the distributions (i.e., when $\|x\|_2 \to \infty$) can have a small impact in terms of TV distance, but a major impact in terms of W_2 distance.

3.2. Log-concave distributions and tail properties

We start by recalling the basic property that we will assume on the probability measure. We will then present some known results about this class of measures which will be exploited in the convergence analysis of our algorithm.

Definition 1. *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ has L-Lipschitz continuous gradient for $L \geq 0$ if $\forall x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2.$$

Definition 2. *We say a function $f : \mathbb{R}^d \to \mathbb{R}$ in convex if $\forall 0 \leq t \leq 1$ and $\forall x, y \in \mathbb{R}^d$,*

$$f(tx+(1-t)y) \leq tf(x) + (1-t)f(y)$$

Definition 3. *We say that probability measure $\mu \propto e^{-f(x)}\,\mathrm{d}x$ is logconcave if f is convex. Moreover, we say that μ is L-smooth if f has a L-Lipschitz continuous gradient.*

As mentioned previously, bounding the Wasserstein distance between two probability measures requires controlling the error at the tail of the distributions. In order to deal with such a distance without injecting large dependence in the dimension, we make the following assumption on the tail of the target distribution, which is quite standard when working with unconstrained non-strongly log-concave distributions (Durmus et al., 2018a; 2017):

Assumption 4. *There exists $\eta > 0, M_\eta > 0$ such that for all $x \in \mathbb{R}^d$ such that $\|x\|_2 \geq M_\eta$,*

$$f(x) - f(x^*) \geq \eta \|x - x^*\|_2$$

where $x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$. Without loss of generality, we will also assume $x^* = 0$ and $f(x^*) = 0$.

Note that in the case of a distribution constrained to a set $\Omega \subset \mathbb{R}^d$, this assumption is naturally satisfied with η arbitrary, and $M_\eta = \text{diam}(\Omega)$ where $\text{diam}(\Omega)$ is the diameter of Ω.

In order to see how this assumption transfers into a constraint on the tail of the distribution, we recall two following results shown in (Durmus et al., 2018a) and (Lovász & Vempala, 2007) respectively.

Lemma 5. *Let $X \in \mathbb{R}^d$ be a random vector from a log-concave distribution μ satisfying assumption 4. Then*

$$\mathbb{E}_{X \sim \mu} \left[\|X\|_2^2 \right] \leq \frac{2d(d+1)}{\eta^2} + M_\eta^2$$

Lemma 6. *Let $X \in \mathbb{R}^d$ be a random vector from a log-concave distribution μ such that $\mathbb{E}\left[\|X\|_2^2\right] \leq C^2$. Then, for any $R > 1$, we have*

$$Pr\left(\|X\|_2 > RC\right) < e^{-R+1}$$

It is thus possible to combine both lemmas to show that any distribution satisfying assumption 4 necessarily has a sub-exponential tail. This property will allow us to control the Wasserstein distance in terms of the total variation distance.

Lemma 7. *Let X be a random vector from a log-concave distribution μ satisfying assumption 4. Then, $\forall R > 1$,*

$$Pr\left(\|X\|_2 > R\sqrt{\frac{2d(d+1)}{\eta^2} + M_\eta}\right) < e^{-R+1}$$

3.3. Unadjusted Langevin Algorithm

Finally, we recall the standard Unadjusted Langevin Algorithm as well as a very useful inequality bounding the KL divergence between the target distribution and the k-th sample distribution.

Consider the probability space $(\mathbb{R}^d, \mathcal{B}, \mu^*)$, where $\mathcal{B}$ is the Borel sigma algebra and μ^* is the target distribution. Suppose that μ^* is log-concave and dominated by the Lebesgue measure on $\mathbb{R}^d$, namely,

$$\mathrm{d}\mu^*(x) = Ce^{-f(x)}\,\mathrm{d}x, \qquad \forall x \in S, \tag{7}$$

where C is an unknown normalizing constant and the function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and ∇f is L-Lipschitz continuous. We wish to sample from μ^* without calculating the normalizing constant C.

A well-known scheme for sampling for such a distribution is called ULA. Initialized at $x_0 \in \mathbb{R}^d$, the iterates of ULA are

$$x_{k+1} = x_k - \gamma \nabla f(x_k) + \sqrt{2\gamma} g_k \tag{8}$$

for all $k \geq 0$, where $\gamma > 0$ is the step-size and the entries of $g_k \in \mathbb{R}^d$ are zero-mean and unit-variance Gaussian random variables, independent from each another and everything else. Let μ_k be the probability measure associated to iterate x_k, $\forall k \geq 0$. It is well-known that ULA converges to the target measure in KL divergence.

More specifically, for $n \geq n_\epsilon = \mathcal{O}(d^3 L \epsilon^{-2})$ iterations, we reach $\mathrm{KL}(\overline{\mu}_n; \mu^*) \leq \epsilon$, where $\overline{\mu}_n = \frac{1}{n}\sum_{k=1}^n \mu_k$ is the average of the probability measures associated to the iterates $\{x_k\}_{k=0}^n$ (Durmus et al., 2018a). The averaging sum $\frac{1}{n}\sum_{k=1}^n \mu_k$ is to be understood in the sense of measures, i.e., sampling from the $\bar{\mu}_n$ is equivalent to choosing an index k uniformly at random among $\{1, ..., n\}$, and then sampling from μ_k.

To prove this result, the authors showed the following useful inequality that we will exploit in our analysis:

Lemma 8. *Suppose that we apply the ULA iterations* (8) *for sampling from a smooth log-concave probability measure $\mu^* \propto e^{-f(x)}\,\mathrm{d}x$ with constant step-size $\gamma > 0$, starting from $x_0 \sim \mu_0$. Then, $\forall n > 0$,*

$$\mathrm{KL}(\bar{\mu}_n; \mu^*) \leq \frac{\mathrm{W}_2^2(\mu_0, \mu^*)}{2\gamma n} + Ld\gamma. \tag{9}$$

4. DL-ULA for unconstrained sampling

In this section, we present a modified version of the standard ULA for sampling from an unconstrained distribution and provide convergence guarantees. This modified version of ULA involves a new step size schedule as well as a projection step. We will show that it allows to obtain improved convergence rate, as well as the first convergence rate in W_2-distance for overdamped Langevin dynamics.

Algorithm 1 Double-loop Unadjusted Langevin Algorithm (DL-ULA)

Input: Smooth unconstrained probability measure μ^*, step sizes $\{\gamma_k\}_{k\geq 1}$, number of (inner) iterations $\{n_k\}_{k\geq 1}$, initial probability measure μ_0 on $\mathbb{R}^d$, and thresholds $\{\tau_k\}_{k\geq 1}$.
Initialization: Draw a sample x_0 from the probability measure μ_0.
for $k = 1, \dots$ **do**
  $x_{k,0} \leftarrow x_{k-1}$
  for $n = 1, \dots, n_k$ **do**
    $x_{k,n+1} \leftarrow x_{k,n} - \gamma_k \nabla f(x_{k,n}) + \sqrt{2\gamma_k} g_{k,n}$, where $g_{k,n} \sim \mathcal{N}(0, I_d)$.
  end for
  $x_k \leftarrow x_{k,i}$, where i is drawn from the uniform distribution on $\{1, \cdots, n_k\}$.
  if $\|x_k\|_2 > \tau_k$ **then**
    $x_k \leftarrow \tau_k x_k / \|x_k\|_2$.
  end if
end for

4.1. DL-ULA algorithm

We consider the problem of sampling from a smooth and unconstrained probability measure $\mu^* \propto e^{-f(x)}\,\mathrm{d}x$, where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. To this end, we apply the standard ULA in a double-loop fashion, and decrease the step size only between each inner loop. Moreover, each inner loop is followed by a projection step onto some Euclidean ball. The procedure is summarized in Algorithm 1.

The projection step appears to be crucial in our analysis in order to control the tail of the sample distribution, which is necessary for bounding its Wasserstein distance to the target distribution.

In the following sections, we derive the convergence rate for Algorithms 1. The global idea for showing the convergence of this algorithm is to use the inequality (9) recursively between each successive outer loop. We denote as $\bar{\mu}_k$ the average distribution associated to the iterates of outer iteration k just *before* the projection step. Similarly, we denote as $\tilde{\mu}_k$ the same distribution *after* the projection step.

Each outer iteration k uses as a starting point a sample from the previous outer iteration $x_{k,0} \sim \tilde{\mu}_{k-1}$. Therefore, we can apply the inequality (9) to the outer iteration k to obtain

$$\mathrm{KL}(\bar{\mu}_k; \mu^*) \leq \frac{\mathrm{W}_2^2(\tilde{\mu}_{k-1}, \mu^*)}{2\gamma_k n_k} + Ld\gamma_k. \tag{10}$$

In order to unfold the recursion, we must have a bound on $\mathrm{W}_2^2(\tilde{\mu}_{k-1}, \mu^*)$ in terms of $\mathrm{KL}(\bar{\mu}_{k-1}, \mu^*)$. Using the light tail property of log-concave distributions, it is easy to obtain a bound between $\mathrm{W}_2^2(\tilde{\mu}_{k-1}, \mu^*)$ and $\mathrm{W}_2^2(\bar{\mu}_{k-1}, \mu^*)$. However, it is not clear how to bound $\mathrm{W}_2^2(\bar{\mu}_{k-1}, \mu^*)$ by $\mathrm{KL}(\bar{\mu}_{k-1}, \mu^*)$.

As an intermediate step in the convergence analysis, we derive in the next section a bound between the W_2-distance and the TV-distance between two general log-concave probability measures, which can then be extended to a W_2-KL bound using Pinsker's inequality.

4.2. Relation Between W_2- and TV-Distances

When μ and ν are both compactly supported on an Euclidean ball of diameter D, then it is well-known that $\mathrm{W}_2(\mu, \nu) \leq D\sqrt{\|\mu - \nu\|_{\mathrm{TV}}}$ (Gibbs & Su, 2002). Otherwise, if μ and ν are not compactly supported, their fast-decaying tail (Lemma 7) allows us to derive a similar bound, as summarized next and proved in Appendix A.

Lemma 9. **(W_2-TV distances inequality)** *Let μ, ν be log-concave probability measures on $\mathbb{R}^d$ both satisfying Assumption 4 with (η, M_η). Then, for some scalar $c \in \mathbb{R}$,*

$$\mathrm{W}_2(\mu, \nu) \leq cd \max\left(\log\left(\frac{1}{\|\mu - \nu\|_{\mathrm{TV}}}\right), 1\right) \sqrt{\|\mu - \nu\|_{\mathrm{TV}}}. \tag{11}$$

In a sense, (11) is an alternative to the powerful T_2 inequality which does not apply generally in our setting (Gozlan & Léonard, 2010). Indeed, for $C_\mu > 0$, recall that a probability measure μ satisfies Talagrand's $T_2(C_\mu)$ transportation inequality if

$$\mathrm{W}_2(\mu, \nu) \leq C_\mu \sqrt{\mathrm{KL}(\mu; \nu)}, \tag{12}$$

for any probability measure ν. Above, C_μ depends only on μ and, in particular, if μ is κ strongly log-concave,[1] then (12) holds with $C_\mu = \mathcal{O}(1/\sqrt{\kappa})$ (Gozlan & Léonard, 2010). In this work, the target measures that we consider are not necessarily strongly log-concave measures, leaving us in need for a replacement to (12). In our analysis, (11) serves as a replacement for (12). Indeed, using the Pinsker's inequality (Pinsker, 1960), an immediate consequence of (11) is that

$$\mathrm{W}_2(\mu, \nu) = \widetilde{\mathcal{O}}(\mathrm{KL}(\mu; \nu)^{\frac{1}{4}}). \tag{13}$$

In fact, (11) might also be of interest in its own right, especially when working with non-strongly log-concave measures. For example, it is easy to use (13) to extend the well-known $\mathcal{O}(\epsilon^{-2})$ convergence rate of ULA in KL divergence to a $\widetilde{\mathcal{O}}(\epsilon^{-8})$ convergence rate in W_2 distance in the non-strongly log-concave setting. To the best of our knowledge, such a result does not exist in the literature.

[1] If $\mathrm{d}\mu \propto e^{-f}\,\mathrm{d}x$, then we say that μ is κ is strongly log-concave if f is κ strongly convex.

Literature	W_2	TV	KL
(Durmus et al., 2018a)	-	$\widetilde{O}\left(Ld^3\epsilon^{-4}\right)$	$\widetilde{O}\left(Ld^3\epsilon^{-2}\right)$
(Durmus et al., 2017)	-	$\widetilde{O}\left(L^2d^5\epsilon^{-2}\right)$	-
(Zou et al., 2018)	$\widetilde{O}\left(L^2d^{10.5}\epsilon^{-6}\right)^*$	-	-
Our work	$\widetilde{O}\left(Ld^9\epsilon^{-6}\right)$	$\widetilde{O}\left(Ld^3\epsilon^{-3}\right)$	$\widetilde{O}\left(Ld^3\epsilon^{-\frac{3}{2}}\right)$

Table 1. Complexity of sampling from a smooth and log-concave probability distribution. For each metric, the entry corresponds to the total number of iterations to use in order to reach an ϵ accuracy in the specified metric. (* For comparison purpose, we extended the proof in (Zou et al., 2018) in the case where the distribution satisfies the weaker assumption 4. The dimension dependence is thus different from (Zou et al., 2018)).

4.3. Convergence Analysis of DL-ULA

Having covered the necessary technical tools above, we now turn our attention to the convergence rate of Algorithm 1. The final step to take care of is to choose the sequences $\{\gamma_k\}_{k\geq 1}$ and $\{n_k\}_{k\geq 1}$ so as to obtain the best possible convergence guarantees. We summarize our result in Theorem 10.

Theorem 10. (iteration complexity of DL-ULA) *Let μ^* be a L-smooth log-concave distribution satisfying assumption 4. Suppose that μ_0 also satisfies assumption 4. For every $k \geq 1$, let*

$$n_k = LM^2dk^2e^{3k} \tag{14}$$

$$\gamma_k = \frac{1}{Ld}e^{-2k} \tag{15}$$

$$\tau_k = Mk. \tag{16}$$

where $M = \sqrt{\frac{2d(d+1)}{\eta^2} + M_\eta^2} = \mathcal{O}(d)$.

Let $\bar{\mu}_k, \tilde{\mu}_k$ be the average distributions associated with the iterates of outer iteration k of DL-ULA using the parameters above, just before and after the projection step respectively. Then, $\forall \epsilon > 0$, we have:

- *After $N^{\mathrm{KL}} = \tilde{\mathcal{O}}(Ld^3\epsilon^{-\frac{3}{2}})$ total iterations, we obtain $\mathrm{KL}(\bar{\mu}_k; \mu^*) \leq \epsilon$.*
- *After $N^{\mathrm{TV}} = \tilde{\mathcal{O}}(Ld^3\epsilon^{-3})$ total iterations, we obtain $\|\tilde{\mu}_k - \mu^*\|_{\mathrm{TV}} \leq \epsilon$.*
- *After $N^{\mathrm{W}_2} = \tilde{\mathcal{O}}(Ld^9\epsilon^{-6})$ total iterations, we obtain $\mathrm{W}_2(\tilde{\mu}_k, \mu^*) \leq \epsilon$.*

A few remarks about Theorem 10 are in order.

Geometric sequences. Theorem 10 prescribes a geometric sequence for the choice of $\{\gamma_k\}_k$ and $\{n_k\}_k$. As outer iteration counter k increases, more and more ULA (inner) iterations are performed with the constant step-size γ_k. Asymptotically, we observe that the step size decreases at a rate $n^{-\frac{2}{3}}$ where n is the total number of ULA iterations. This decaying rate is faster than the standard decaying rate of $n^{-\frac{1}{2}}$ for ULA (Durmus et al., 2018a).

In constrast to convex optimization where a global optimum can provably be reached with constant step-size, ULA cannot converge to the target distribution μ^* when using constant step-size, since the stationary distribution of ULA iterates (8) when using a constant step size is different from the target distribution. Asymptotically, it is thus desirable to use as small a step-size as possible.

Projection step. Although the initial and target distributions are both log-concave, and thus have a sub-exponential tail, the sample distributions $\bar{\mu}_k$ are not generally log-concave, and it is not clear whether they also share the sub-exponential tail property. The projection step at the end of each outer iteration provides a way to enforce the light tail property, so that we can still apply a bound similar to (11). This procedure is made clearer in the proof of the theorem.

This procedure also provides more stability in the early outer iterations where the step-size is the largest. Moreover, since $\lim_{k\to\infty} \tau_k = \infty$, the projection step asymptotically never applies in practice.

Convergence rate comparison Table 1 summarizes various convergence rates of Langevin dynamics based methods applied to general log-concave distributions. We observe that DL-ULA achieves improved convergence guarantees either in terms of rate or dimension dependence. Compared to (Durmus et al., 2017), the convergence rate in TV distance is worse in terms of accuracy ϵ but enjoys much better dimension dependence, and is also better in terms of Lipschitz constant dependence.

5. DL-MYULA for constrained sampling

We now apply the same multistage idea to an existing constrained sampling algorithm, and show that it allows both to obtain an asymptotic convergence and improved convergence guarantees.

5.1. DL-MYULA algorithm

Consider sampling from a log-concave distribution over a convex set $\Omega \subset \mathbb{R}^d$, i.e.,

$$\mu^*(x) = \begin{cases} e^{-f(x)} / \int_\Omega e^{-f(x')} dx' & x \in \Omega \\ 0 & x \notin \Omega. \end{cases} \tag{17}$$

In (Durmus et al., 2018b; Brosse et al., 2017), the authors propose to reduce this problem to an unconstrained sampling problem by penalizing the domain outside Ω directly inside the probability measure using its Moreau-Yoshida envelop. More precisely, they propose to sample from the following unconstrained probability measure $\mathrm{d}\mu_\lambda(x) \propto e^{-f_\lambda(x)}\,\mathrm{d}x$ where $f_\lambda : \mathbb{R}^d \to \mathbb{R}$ is defined as:

$$f_\lambda(x) = f(x) + \frac{1}{2\lambda}\|x - \mathrm{proj}_\Omega(x)\|_2^2, \qquad \forall x \in \mathbb{R}^d, \tag{18}$$

where $\mathrm{proj}_\Omega : \mathbb{R}^d \to \Omega$ is the standard projection operator onto Ω defined as $\mathrm{proj}_\Omega(x) = \arg\min_{y\in\Omega} \|x - y\|_2$. Note that this penalty is easily differentiable as soon as the projection onto Ω can be computed since $\nabla f_\lambda(x) = \nabla f(x) + \frac{1}{\lambda}(x - \mathrm{proj}_\Omega(x))$.

By bounding the TV distance between μ_λ and μ^*, they showed that, by sampling from μ_λ with λ small enough, it is possible to sample from μ^* with arbitrary precision. This algorithm is called Moreau-Yoshida ULA (MYULA).

Building on this approach, we can apply our double loop algorithm, by modifying both the step size as well as the penalty parameter λ between each inner loop (Algorithm 2).

In addition to providing improved rate, as we will show later, our algorithm also has the advantage to use a decreasing penalty parameter λ so as to guarantee asymptotic convergence of the algorithm to the target distribution. On the other hand, MYULA uses constant penalty λ, and thus saturates after a certain number of iterations. Although this looks like a trivial extension, using varying penalty parameter makes the analysis more challenging since the target distribution of the algorithm is regularly changing.

Algorithm 2 DL-MYULA

Input: Smooth constrained probability measure μ^*, step sizes $\{\gamma_k\}_{k\geq 1}$, penalty parameters $\{\lambda_k\}_{k\geq 1}$, number of (inner) iterations $\{n_k\}_{k\geq 1}$, initial probability measure μ_0 on $\mathbb{R}^d$, and thresholds $\{\tau_k\}_{k\geq 1}$.
Initialization: Draw a sample x_0 from the probability measure μ_{init}.
for $k = 1, \dots$ **do**
  $x_{k,0} \leftarrow x_{k-1}$
  for $n = 1, \dots, n_k$ **do**
    $x_{k,n+1} \leftarrow x_{k,n} - \gamma_k(\nabla f(x_{k,n}) + \frac{1}{\lambda_k}(x_{k,n} - \mathrm{proj}_\Omega(x_{k,n}))) + \sqrt{2\gamma_k} g_{k,n}$, where $g_{k,n} \sim \mathcal{N}(0, I_d)$.
  end for
  $x_k \leftarrow x_{k,i}$, where i is drawn from the uniform distribution on $\{1, \cdots, n_k\}$.
  if $\|x_k\|_2 > \tau_k$ **then**
    $x_k \leftarrow \tau_k x_k / \|x_k\|_2$.
  end if
end for

5.2. Convergence analysis of DL-MYULA

We now analyze the convergence of DL-MYULA. In Algorithm 2, both the step-size γ and the penalty parameter λ are decreased after each outer iteration. Therefore, at each outer iteration k, we aim to sample from the unconstrained penalized distribution $\mathrm{d}\mu_{\lambda_k} \propto e^{-f_{\lambda_k}(x)}\,\mathrm{d}x$ where f_{λ_k} is defined in equation (18).

Similarly as for DL-ULA, we will use Lemma 9 after each outer iteration. However, since the target distribution of outer iteration is μ_{λ_k} instead of μ^*, the inequality reads as follows:

$$\mathrm{KL}(\bar{\mu}_k; \mu_{\lambda_k}) \leq \frac{\mathrm{W}_2^2(\tilde{\mu}_{k-1}, \mu_{\lambda_k})}{2\gamma_k n_k} + Ld\gamma_k.$$

where we recall that $\bar{\mu}_k$ is the average iterate distribution of outer iteration k just before the projection step, and $\tilde{\mu}_k$ is the one just after the projection step.

In order to use a similar recursion argument as previously, we must thus bound $\mathrm{W}_2(\tilde{\mu}_{k-1}, \mu_{\lambda_k})$ by $\mathrm{W}_2(\tilde{\mu}_{k-1}, \mu_{\lambda_{k-1}})$. Using the triangle inequality for W_2, we have

$$\begin{aligned}\mathrm{W}_2(\tilde{\mu}_{k-1}, \mu_{\lambda_k}) \leq \mathrm{W}_2(\tilde{\mu}_{k-1}, \mu_{\lambda_{k-1}}) + \mathrm{W}_2(\mu_{\lambda_{k-1}}, \mu^*) \\ + \mathrm{W}_2(\mu_{\lambda_k}, \mu^*).\end{aligned}$$

In (Brosse et al., 2017), the authors showed a bound for $\|\mu_\lambda - \mu^*\|_{\mathrm{TV}}$ in terms of $\lambda > 0$, and it is easy to extend their proof to obtain a bound for $\mathrm{W}_2(\mu_\lambda, \mu^*)$ (see Lemma 12 and its proof in Appendix C).

In order to prove our result, we make the same assumptions on the constraint set Ω as in (Brosse et al., 2017):

Assumption 11. *There exist $r, R, \Delta_1 > 0$ such that*

1. *$B(0, r) \subset \Omega \subset B(0, D)$ where $B(0, r_0) = \{y \in \mathbb{R}^d : \|x - y\|_2 \leq r_0\}$ $\forall r_0 > 0$,*
2. *$e^{\inf_{\Omega^c}(f) - \max_\Omega(f)} \geq \Delta_1$, where $\Omega^c = \mathbb{R}^d \backslash \Omega$.*

Lemma 12. *Let $\Omega \subset \mathbb{R}^d$ satisfy Assumption 11. Then $\forall \lambda < \frac{r^2}{8d^2}$,*

$$\mathrm{W}_2^2(\mu_\lambda, \mu^*) \leq C_\Omega^2 d\sqrt{\lambda} \tag{19}$$

for some scalar $C_\Omega > 0$ depending on D, r and Δ_1.

The proof of the previous Lemma is given in Appendix C. Using these results, the convergence proof is then very similar as for DL-ULA, and is summarized in Theorem 13, whose proof can be found in Appendix D.

Theorem 13. (iteration complexity of DL-MYULA) *Let $\Omega \subset \mathbb{R}^d$ be a convex set satisfying Assumption 11 and μ^* be a log-concave distribution given by* (17) *where f has L-Lipschitz continuous gradient. For every $k \geq 1$, let*

$$\lambda_k = \frac{1}{\frac{8d^2}{r^2} + de^{2k}} \tag{20}$$

$$n_k = Ldk^2e^{5k} \tag{21}$$

$$\gamma_k = \frac{1}{Ld}e^{-4k} \tag{22}$$

$$\tau_k = Dk \tag{23}$$

for every $k \geq 1$. Then, $\forall \epsilon > 0$, we have:

- *After $N^{\mathrm{TV}} = \mathcal{O}\left(d^{3.5}\epsilon^{-5}\right)$ total iterations, we obtain $\|\hat{\mu}_K - \mu^*\|_{\mathrm{TV}} \leq \epsilon$.*
- *After $N^{\mathrm{W}_2} = \tilde{\mathcal{O}}\left(d^{3.5}\epsilon^{-10}\right)$ total iterations, we obtain $\mathrm{W}_2(\hat{\mu}_K, \mu^*) \lesssim \epsilon$.*

We make a few comments about this convergence result.

Smoothness of μ_{λ_k} One can notice that outer iterations in DL-MYULA are longer than in DL-ULA. In order to explain this choice, first observe that the Lipschitz constant associated with the penalized distribution μ_λ grows as $\mathcal{O}\left(\frac{1}{\lambda}\right)$ as λ goes to 0. As k increases and λ_k decreases, μ_{λ_k} becomes less and less smooth. Thus, for ULA to succeed in approximating μ_{λ_k}, the step size γ_k of ULA iterations reduces accordingly, and the number of iterations increases.

The choice for λ_k ensures that $\lambda_k < \frac{r^2}{8d^2}$ as required for Lemma 12 to be applicable.

Convergence rate comparison Table 2 summarizes convergence rates in TV distance for various first-order constrained sampling algorithms. We can see that DL-MYULA outperforms existing approaches, both in terms of rate and dimension dependence.

Algorithm	TV	Literature
PLMC	$d^{12}\widetilde{O}\left(\epsilon^{-12}\right)$	(Bubeck et al., 2018)
MYULA	$d^{5}\widetilde{O}\left(\epsilon^{-6}\right)$	(Brosse et al., 2017)
DL-MYULA	$d^{3.5}\widetilde{O}\left(\epsilon^{-5}\right)$	Our work

Table 2. Upper bounds on the number of iterations required in order to guarantee an error smaller than ϵ in TV distance for various constrained sampling algorithms.

6. Conclusion

In this work, we proposed and analyzed a new step-size schedule for the well-known Unadjusted Langevin Algorithm. Our approach works by applying ULA successively with constant step-size, and by geometrically decreasing it after a certain number of iterations. Exploiting a new result on the relation between the 2-Wasserstein distance and the TV distance of two log-concave distributions, we were able to prove new convergence guarantees for this procedure. We also applied our approach to an existing first-order constrained sampling, and showed improved convergence guarantees, both in terms of rate and dimension dependence.

7. Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n○ 725594 - time-data).

This work was supported by the Swiss National Science Foundation (SNSF) under grant number 407540_167319.

This project was sponsored by the Department of the Navy, Office of Naval Research (ONR) under a grant number N62909-17-1-2111.

This work was supported by Hasler Foundation Program: Cyber Human Systems (project number 16066).

References

Ahn, S., Korattikara, A., and Welling, M. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

Aybat, N. S., Fallah, A., Gurbuzbalaban, M., and Ozdaglar, A. A universally optimal multistage accelerated stochastic gradient method. *arXiv preprint arXiv:1901.08022*, 2019.

Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. *arXiv preprint arXiv:1705.08964*, 2017.

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

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient mcmc algorithms with high-order in-

tegrators. In *Advances in Neural Information Processing Systems*, pp. 2278–2286, 2015.

Cheng, X. and Bartlett, P. Convergence of langevin mcmc in kl-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped langevin mcmc: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

Dalalyan, A. S. and Karagulyan, A. G. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

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

Durmus, A., Majewski, S., and Miasojedow, B. Analysis of langevin monte carlo via convex optimization. *arXiv preprint arXiv:1802.09188*, 2018a.

Durmus, A., Moulines, E., and Pereyra, M. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018b.

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. Log-concave sampling: Metropolis-hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018.

Ge, R., Kakade, S. M., Kidambi, R., and Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure. *arXiv preprint arXiv:1904.12838*, 2019.

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

Gozlan, N. and Léonard, C. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. Mirrored langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 2883–2892, 2018.

Li, C., Chen, C., Carlson, D., and Carin, L. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016a.

Li, W., Ahn, S., and Welling, M. Scalable mcmc for mixed membership stochastic blockmodels. In *Artificial Intelligence and Statistics*, pp. 723–731, 2016b.

Lovász, L. and Vempala, S. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Luu, T., Fadili, J., and Chesneau, C. Sampling from non-smooth distribution through langevin diffusion. 2017.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

Patterson, S. and Teh, Y. W. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in neural information processing systems*, pp. 3102–3110, 2013.

Pinsker, M. S. Information and information stability of random variables and processes. 1960.

Villani, C. Optimal transport–old and new, volume 338 of a series of comprehensive studies in mathematics, 2009.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Yousefian, F., Nedić, A., and Shanbhag, U. V. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

Zou, D., Xu, P., and Gu, Q. Stochastic variance-reduced hamilton monte carlo methods. *arXiv preprint arXiv:1802.04791*, 2018.