Journal of Machine Learning Research 22 (2021) 1-59 Submitted 7/20; Revised 6/21; Published 6/21

Replica Exchange for Non-Convex Optimization

Jing Dong JING.DONG@GSB.COLUMBIA.EDU
Graduate School of Business
Columbia University
3022 Broadway
New York, NY 10027, USA

Xin T. Tong MATTXIN@NUS.EDU.SG
Department of Mathematics
National University of Singapore
Block S17, 10 Lower Kent Ridge Road
Singapore 119077

Editor: Qiang Liu

Abstract

Gradient descent (GD) is known to converge quickly for convex objective functions, but it can be trapped at local minima. On the other hand, Langevin dynamics (LD) can explore the state space and find global minima, but in order to give accurate estimates, LD needs to run with a small discretization step size and weak stochastic force, which in general slow down its convergence. This paper shows that these two algorithms can "collaborate" through a simple exchange mechanism, in which they swap their current positions if LD yields a lower objective function. This idea can be seen as the singular limit of the replica-exchange technique from the sampling literature. We show that this new algorithm converges to the global minimum linearly with high probability, assuming the objective function is strongly convex in a neighborhood of the unique global minimum. By replacing gradients with stochastic gradients, and adding a proper threshold to the exchange mechanism, our algorithm can also be used in online settings. We also study non-swapping variants of the algorithm, which achieve similar performance. We further verify our theoretical results through some numerical experiments, and observe superior performance of the proposed algorithm over running GD or LD alone.

Keywords: Non-convex optimization, Langevin dynamics, gradient descent, stochastic gradient Langevin dynamics, stochastic gradient descent

1. Introduction

Division of labor is the secret of any efficient enterprises. By collaborating with individuals with different skillsets, we can focus on tasks within our own expertise and produce better outcomes than working independently. This paper asks whether the same principle can be applied when designing an algorithm.

Given a general smooth non-convex objective function F, we consider the unconstrained optimization problem

$$\min_{x\in\mathbb{R}^d} F(x). \tag{1}$$

It is well-known that deterministic optimization algorithms, such as gradient descent (GD), can converge to a local minimum quickly (Nesterov, 2013). However, this local minimum may not be the global minimum, and GD will be trapped there afterwards. On the other hand, sampling-based algorithms, such as Langevin dynamics (LD) can escape local minima by their stochasticity, but the additional stochastic noise contaminates the optimization results and slows down the convergence when the iterate is near the global minimum. More generally, deterministic algorithms are mostly designed to find local minima quickly, but they can be terrible at exploration. Sampling-based algorithms are better suited for exploring the state space, but they are inefficient when pinpointing the local minima. This paper investigates how they can "collaborate" to get the "best of both worlds".

The collaboration mechanism we introduce stems from the idea of replica exchange in the sampling literature. Its implementation is very simple: we run a copy of GD, denoted by X_n; and a copy of discretized LD, denoted by Y_n, in parallel. At each iteration, if $F(X_n) > F(Y_n)$, we swap their positions. At the final iteration, we output X_N. The proposed algorithm, denoted by GDxLD and formalized in Algorithm 1 below, enjoys the "expertise" of both GD and LD. In particular, we establish that if F is convex in a neighborhood of the unique global minimum, then, for any $\epsilon > 0$, there exists $N(\epsilon) = O(\epsilon^{-1})$, such that for $N \geq N(\epsilon)$, $|F(X_N) - F(x^*)| < \epsilon$ with high probability, where x^* is the global minimum. If F is strongly convex in the same neighborhood, we can further obtain linear convergence; that is, $N(\epsilon)$ can be reduced to $O(\log \frac{1}{\epsilon})$.

As we will demonstrate with more details in Section 2.2, GDxLD can be seen as the singular limit of a popular sampling algorithm, known as replica exchange or parallel tempering. The exchange mechanism is in place to make the sampling algorithm a reversible MCMC. However, for the purpose of optimization, exchanging the locations of GD and LD is not the only option. In fact, both GD and LD can obtain the location of Y_n if $F(X_n) > F(Y_n)$. This leads to a slightly different version of the algorithm, which we will denote as nGDxLD where "n" stands for "non-swapping". Since the LD part will not be swapped to the location of the GD, it is a bona-fide Langevin diffusion algorithm. In terms of performance, we establish the same complexity estimates for both GDxLD and nGDxLD. Our numerical experiments also verify that GDxLD and nGDxLD have similar performance. To simply the notation, we write (n)GDxLD when we refer to both versions of the algorithm.

Notably, the complexity bounds we establish here are the same as the complexity bounds for standard GD when F is globally convex (or strongly convex) (Nesterov, 2013), but we only need F to be convex (or strongly convex) near x^*, which is significantly weaker. It is not difficult to see intuitively why (n)GDxLD works well in such non-convex settings. The LD explores the state space and visits the neighborhood of the global minimum. Since this neighborhood is of a constant size, it can be found by the LD in constant time. Moreover, this neighborhood gives lower values of F than anywhere else, so the GD will be swapped there, if it is not there already. Finally, GD can pinpoint the global minimum as it now starts in the right neighborhood. Figure 1 provides a visualization of the mechanism.

For many modern data-driven applications, F is an empirical loss function, so its precise evaluation and its gradient evaluation can be computationally expensive. For these scenarios, we also consider an online modification of our algorithm. The natural modification of GD is stochastic gradient decent (SGD), and the modification of LD is stochastic gradient

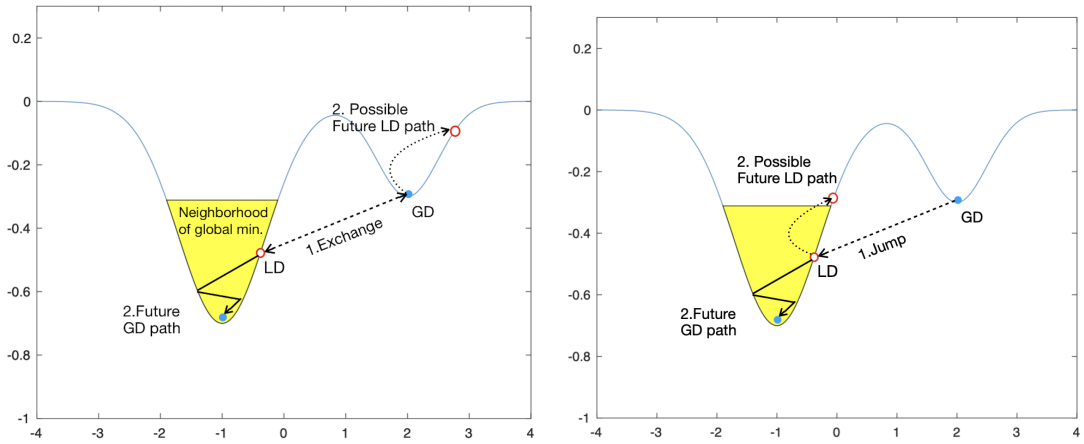


Figure 1: (Left) GDxLD in action: when LD enters a neighborhood of the global minimum (yellow region), exchange happens and helps GD to escape the local minimum that traps it. After exchange, GD can converge to the global minimum, and LD keeps exploring. (Right) nGDxLD is similar to GDxLD; the only difference is that LD will stay at its original location, instead of swapping with GD.

Langevin dynamics (SGLD). This algorithm, denoted by SGDxSGLD and formalized in Algorithm 2 below, achieves a similar complexity bound as GDxLD if F is strongly convex in the neighborhood of the unique global minimum x^*. For the theory to apply, we also need the noise of the stochastic gradient to be sub-Gaussian and is of order $O(\sqrt{\epsilon})$. This assumption can often be met by using a mini-batch of size $O(\epsilon^{-1})$, which in principle should be factored in for complexity. In this case, the overall complexity of SGDxSGLD is $O(\epsilon^{-1} \log \frac{1}{\epsilon})$. Similar to the offline scenario, the non-swapping variation, nSGDxSGLD, can be obtained by keeping the SGLD part not swapped to the location of the SGD.

1.1 Related work

Non-convex optimization problems arise in numerous advanced machine learning, statistical learning, and structural estimation settings (Geyer and Thompson, 1995; Bottou et al., 2018). How to design efficient algorithms with convergence guarantees has been an active area of research due to their practical importance. In what follows, we discuss some existing results related to our work. As this is a fast growing and expanding field, we focus mostly on algorithms related to GD, LD, or their stochastic gradient versions.

One main approach to study nonlinear optimization is to ask when an algorithm can find local minima efficiently. The motivation behind finding local minima, rather than global minima, is that in some machine learning problems, such as matrix completion and wide neural networks, local minima already have good statistical properties or prediction power (Ge et al., 2017; Park et al., 2017; Bhojanapalli et al., 2016; Du et al., 2018; Ge et al., 2018; Mei et al., 2017). Moreover, the capability to find local minima or second-order stationary points (SOSP) is nontrivial, since GD can be trapped at saddle points. When full Hessian information is available, this can be done through algorithms such as cubic-regularization or trust region (Nesterov and Polyak, 2006; Curtis et al., 2014). If only gradient is available,

one general idea is to add stochastic noise so that the algorithms can escape saddle points (Jin et al., 2017; Allen-Zhu, 2018; Jin et al., 2018; Yu et al., 2018; Daneshmand et al., 2018; Jin et al., 2019). But the "size" of the noise and the step size need to be tuned based on the accuracy requirement. This, on the other hand, reduces the learning rate and the speed of escaping saddle points. For example, to find an ϵ-accurate SOSP in the offline setting, the perturbed gradient descent method requires $O(\epsilon^{-2})$ iterations (Jin et al., 2017), and in the online setting, it requires $O(\epsilon^{-4})$ iterations (Jin et al., 2019). These convergence rates are slower than the ones of our proposed algorithms.

For problems in which local minima are not good enough, we often need to use sampling-based algorithms to find global minima. This often involves simulating an ergodic stochastic process for which the invariant measure is proportional to $\exp(-\frac{1}{\gamma}F(x))$, where γ is referred to as the "temperature", which often controls the strength of stochasticity. Because the process is ergodic, it can explore the state space. However, for the invariant measure to concentrate around the global minimum, γ needs to be small. Then for these sampling based-algorithms to find accurate approximation of global minima, they need to use weaker stochastic noise or smaller step sizes, which in general slow down the convergence. For LD in the offline setting and SGLD in the online setting, the complexity is first studied in Raginsky et al. (2017), later improved by Xu et al. (2018), and generalized by Chen et al. (2019a) to settings with decreasing step sizes. In Xu et al. (2018), it is shown that LD can find an ϵ-accurate global minimum in $O(\epsilon^{-1})$ iterations, and SGLD can do so in $O(\epsilon^{-5})$ iterations. These algorithms have higher complexity than the ones we proposed in this paper. Note that in Xu et al. (2018), it keeps the temperature at a constant order, in which case the step size needs to scale with ϵ. As we will discuss in more details in Section 2, the algorithm we propose keeps both the temperature and the step size at constant values.

We also comment that sampling-based algorithms may lead to better dependence on the dimension of the problem for some MCMC problems as discussed in Ma et al. (2019). However, if the goal is to find global minima of a Lipschitz-smooth non-convex optimization problem, the complexity in general has exponential dependence on the dimension. This has to do with the spectral gap of the sampling process and has been extensively discussed in Ma et al. (2019); Raginsky et al. (2017); Xu et al. (2018). Due to this, our developments in this paper focus on settings where the dimension of the problem is fixed, and we characterize how the complexity scales with the precision level ϵ.

Aside from optimization, LD and related Markov Chain Monte Carlo (MCMC) algorithms are also one of the main workhorses for Bayesian statistics. Our work is closely related to the growing literature on convergence rate analysis for LD. Asymptotic convergence of discretized LD with decreasing temperature (diffusion coefficient) and step sizes is extensively studied in the literature (see, e.g., Gidas (1985); Pelletier (1998)). Nonasymptotic performance bounds for discrete-time Langevin updates in the case of convex objective functions are studied in Durmus and Moulines (2017); Dalalyan (2017); Cheng et al. (2018); Dalalyan and Karagulyan (2019). In MCMC, the goal is to sample from the stationary distribution of LD, and thus the performance is often measured by the Wasserstein distance or the Kullback-Leibler divergence between the finite-time distribution and the target invariant measure (Ma et al., 2015; Bubeck et al., 2018). In this paper, we use the performance metric $\mathbb{P}(|F(X_N) - F(x^*)| > \epsilon)$, which is more suitable for the goal of optimizing a non-

convex function. This also leads to a very different framework of analysis compared to the existing literature.

1.2 Main contribution

The main message of this paper is that, by combining a sampling algorithm with an optimization algorithm, we can create a better tool for some non-convex optimization problems. From the practical point of view, the new algorithms have the following advantages:

- When the exact gradient ∇F is accessible, we propose the GDxLD algorithm and its non-swapping variant, nGDxLD. Their implementation does not require the step size h or temperature γ to change with the precision level ϵ. Such independence is important in practice, since tuning hyper-parameters is in general difficult. As we will demonstrate in Section 3.1.2, our algorithm is quite robust to a wide range of hyper-parameters.

- When the dimension is fixed, for a given precision level ϵ, (n)GDxLD beats existing algorithms in computational cost. In particular, we show (n)GDxLD can reach approximate global minimum with a high probability in $O(\log \frac{1}{\epsilon})$ iterations. Comparing to the iteration estimate of LD in Xu et al. (2018), which is $O(\epsilon^{-1} \log \frac{1}{\epsilon})$, our algorithm is much more efficient, but we require the extra assumption that F has a unique global minimum x^*, and F is strongly convex near x^*.

- When only stochastic approximation of ∇F is available, we propose the SGDxSGLD algorithm and its non-swapping variant, nSGDxSGLD. Like (n)GDxLD, their implementation does not require the temperature or the step size to change with the precision level. (n)SGDxSGLD is also more efficient when comparing with other online optimization methods using stochastic gradients. In particular, we show (n)SGDxSGLD can reach approximate global minimum with high probability with an $O(\epsilon^{-1} \log \frac{1}{\epsilon})$ complexity. This is better than the complexity estimate of VR-SGLD in Xu et al. (2018), which is $O(\epsilon^{-4}(\log \frac{1}{\epsilon})^4)$. The additional assumptions we require is that the function evaluation noise is sub-Gaussian, and F is strongly convex near the unique global minimum x^*.

In term of algorithm design, a novel aspect we introduce is the exchange mechanism. The idea comes from replica-exchange sampling, which has been designed to overcome some of the difficulties associated with rare transitions to escape local minima (Earl and Deem, 2005; Woodard et al., 2009). Dupuis et al. (2012) uses the large deviation theory to define a rate of convergence for the empirical measure of replica-exchange Langevin diffusion. It shows that the large deviation rate function is increasing with the exchange intensity, which leads to the development of infinite swapping dynamics. We comment that infinite swapping is not feasible in practice as the actual swapping rate depends on the discretization step-size. The algorithm that we propose attempts a swap at every iteration, which essentially maximizes the practical swapping intensity. Chen et al. (2019b) extend the idea to solve non-convex optimization problems, but they only discuss the exchange between two LD processes. Our work further extends the idea by combining GD with LD and provides finite-time performance bounds that are tailored to the optimization setting.

In online settings, the function and gradient evaluations are contaminated with noises. Designing the exchange mechanism in this scenario is more challenging, since we need to avoid and mitigate the possible erroneous exchanges. We demonstrate such a design is feasible by choosing a proper mini-batch size as long as the function evaluation noise is sub-Gaussian. The analysis of SGDxSGLD is hence much more challenging and nonstandard, when comparing with that of GDxLD.

While it is a natural idea to let two algorithms specializing in different tasks to collaborate, coming up with a good collaboration mechanism is not straightforward. For example, one may propose to let the LD explore the state-space first, after it finds the region of the global minimum, one would turn-off the temperature, i.e., setting $\gamma = 0$, and run the GD. However, in practice, it is hard to come up with a systematic way to check whether the process is in the neighborhood of the global minimum or not. One may also propose to turn down the temperature of the LD gradually. This is the idea of simulated annealing (Kirkpatrick et al., 1983; Mangoubi and Vishnoi, 2018). The challenge there is to come up with a good "cooling schedule". To ensure convergence to global minima theoretically, the temperature needs to be turned down very slowly, which could jeopardize the efficiency of the algorithm (Gelfand and S.K., 1991). For example, Granville et al. (1994) shows that for the simulated annealing to converge, the temperature at the n-th iteration needs to be of order $1/\log n$.

Readers who are familiar with optimization algorithms might naturally think of doing replica exchange between other deterministic algorithms and other sampling-based algorithms. Standard deterministic algorithms include GD with line search, Newton's method, and heavy-ball methods such as Nestrov acceleration. Sampling-based algorithms include random search, perturbed gradient descent, and particle swarm optimization. We investigate GDxLD instead of other exchange combinations, not because GDxLD is superior in convergence speed, but because of the following two reasons:

1. GDxLD can be seen as a natural singular limit of replica-exchange LD, which is a mathematical subject with both elegant theoretical properties and useful sampling implications. We will explain the connection between GDxLD and replica-exchange LD in Section 2.2.

2. GDxLD is very simple to implement. It can be easily adapted to the online setting. We will explain how to do so in Section 2.5.

On the other hand, it would be interesting to see whether exchange between other algorithms can provide faster rate of convergence in theory or in practice. We leave this as a future research direction, and think the analysis framework and proving techniques we developed here can be extended to more general settings. In particular, due to the division of labor, LD is only used to explore the state-space. Therefore, instead of establishing its convergence as in Raginsky et al. (2017); Xu et al. (2018), we only need to prove that it is suitably ergodic, such that there is a positive probability of visiting the neighborhood of the global minimum. GD is only used for exploitation and the complexity only depends on its behavior in the neighborhood of the global minimum.

Lastly, one of the key element in the complexity of (n)GDxLD is how long it takes LD to find the neighborhood of the global minimum. High dimensional and multi-modal

structures can be difficult to sample using just LD. In the literature, sampling techniques such as dimension reduction, Laplace approximation, Gibbs-type-partition can be applied to improve the sampling efficiency for high dimensional problems (Hairer et al., 2014; Cui et al., 2014; Morzfeld et al., 2019; Tong et al., 2020). As for multi-modality, it is well known that LD may have slow convergence when sampling from mixture of log-concave densities if each component density is close to singular (Menz and Schlichting, 2014). The replica exchange method, discussed in Section 2.2 below, considers running multiple LDs with different temperatures and exchanging among them (Swendsen and Wang, 1986). It has been shown to perform well for multi-modal densities (Woodard et al., 2009; Dong and Tong, 2020). Other than replica exchange, simulated tempering, which considers dynamically changing the temperature of the LD can also handle such multimodality in general (Marinari and Parisi, 1992; Woodard et al., 2009; Lee et al., 2018). Replacing LD with these more advanced sampler may facilitate more efficient exploration when facing more challenging energy landscapes.

2. Main results

In this section, we present the main algorithms: GDxLD and nGDxLD (Algorithm 1), SGDxSGLD and nSGDxSGLD (Algorithm 2). We also provide the corresponding complexity analysis. We start by developing (n)GDxLD in the offline setting (Section 2.1) and discuss the connection between GDxLD and replica-exchange LD (Section 2.2). The rigorous complexity estimate is given by Theorem 2 in Section 2.3. We also study how the complexity depends on the exploration speed of LD in Section 2.4 (see Theorem 3). We then discuss how to adapt the algorithm to the online setting and develop (n)SGDxSGLD in Section 2.5. Section 2.6 provides the corresponding complexity analysis – Theorems 4 and 5. To highlight the main idea and make the discussion concise, we defer the proof of Theorems 2 and 3 to Appendix A, and the proof of Theorems 4 and 5 to Appendix B.

2.1 Replica exchange in offline setting

We begin with a simple offline optimization scenario, where the objective function F and its gradient ∇F are directly accessible. GD is one of the simplest deterministic iterative optimization algorithms. It involves iterating the formula

$$X_{n+1} = X_n - \nabla F(X_n)h. \tag{2}$$

The hyper-parameter h is known as the step size, which can often be set as a small constant. If F is convex and x^* is a global minimum, the GD iterates can converge to a global minimum "linearly" fast; that is, $F(X_n) - F(x^*) \leq \epsilon$ if $n = O(\frac{1}{\epsilon})$. If F is strongly convex with convexity parameter m, then the convergence will be "exponentially" fast; that is, $F(X_n) - F(x^*) \leq \epsilon$ if $n = O((\log \frac{1}{\epsilon})/m)$.

In the general non-convex optimization setting, F can still be strongly convex near a global minimum x^*. In this case, if GD starts near x^*, the iterates can converge to x^* very fast. However, this is hard to implement in practice, since it requires a lot of prior knowledge of F. If we start GD at an arbitrary point, the iterates are likely to be trapped at a local minimum. To resolve this issue, one method is to add stochastic noise to GD and

generate iterates according to

$$Y_{n+1} = Y_n - \nabla F(Y_n)h + \sqrt{2\gamma h}Z_n, \tag{3}$$

where Z_n's are i.i.d. samples from $\mathcal{N}(0, I_d)$. For a small enough step size or learning rate h, the iterates (3) can be viewed as a temporal discretization of the LD

$$dY_t = -\nabla F(Y_t)dt + \sqrt{2\gamma}dB_t, \tag{4}$$

where B_t is a d-dimensional standard Brownian motion. The diffusion process (4) is often called the overdamped Langevin dynamic or unadjusted Langevin dynamic (Durmus and Moulines, 2017). Here for notational simplicity, we refer to algorithm (3) as LD as well. γ is referred to as the temperature of the LD.

It is known that under certain regularity conditions, (4) has an invariant measure $\pi_\gamma(x) \propto \exp(-\frac{1}{\gamma}F(x))$. Note that the stationary measure is concentrated around the global minimum for small γ. Therefore, it is reasonable to hypothesize that by iterating according to (3) enough times, the iterates will be close to x^*. Adding the stochastic noise $\sqrt{2\gamma h}Z_n$ in (3) partially resolves the non-convexity issue, since the iterates can now escape local minima. On the other hand, in order for (3) to be a good approximation of (4), and consequently to be a good sampler of π_γ, it is necessary to use a small step size h and hence a lot of iterations. In particular, Xu et al. (2018) shows in Corollary 3.3 that in order for $\mathbb{E}[F(Y_n)] - F(x^*) = O(\epsilon)$, h needs to scale linearly as ϵ, and the computational complexity is $O(\epsilon^{-1}\log\frac{1}{\epsilon})$. This is slower than GD if F is strongly convex.

In summary, when optimizing a non-convex F, one has a dilemma in choosing the types of algorithms. Using stochastic algorithms like LD will eventually find a global minimum, but they are inefficient for accurate estimation. Deterministic algorithms like GD is more efficient if initialized properly, but there is the danger that they can get trapped in local minima when initialized arbitrarily.

To resolve this dilemma, idealistically, we can use a stochastic algorithm first to explore the state space. Once we detect that the iterate is close to x^*, we switch to a deterministic algorithm. However, it is in general difficult to write down a detection criterion a-priori. Our idea is to devise a simple on-the-fly criterion. It involves running a copy of GD, X_n in (2), and a copy of LD, Y_n in (3), simultaneously. Since a smaller F-value implies the iterate is closer to x^* in general, we apply GD in the next iteration for the one with a smaller F-value, and LD to the one with a larger F-value. In other words, we want to exchange the locations of X_n and Y_n if $F(X_n) > F(Y_n) + t_0$, where $t_0 \geq 0$ is a properly chosen threshold. Finally, to implement the non-swapping variation, that is nGDxLD, it suffices to let LD stay at its original location when the exchange takes place. A more detailed description of the algorithm is given in Algorithm 1.

2.2 GDxLD as a singular limit of replica-exchange Langevin diffusion

In this section, we review the idea of replica-exchange LD, and show its connection with GDxLD. Consider an LD

$$dX_t = -\nabla F(X_t)dt + \sqrt{2\nu}dW_t,$$

Algorithm 1: GDxLD and nGDxLD: offline optimization

Input: Temperature γ, step size h, number of steps N, exchange threshold $t_0 \in [0, r_0/8)$, and initial X_0, Y_0.

for $n = 0$ *to* $N-1$ **do**

$X'_{n+1} = X_n - \nabla F(X_n)h$;

$Y'_{n+1} = Y_n - \nabla F(Y_n)h + \sqrt{2\gamma h}Z_n$, where $Z_n \sim N(0, I_d)$.;

if $F(Y'_{n+1}) < F(X'_{n+1}) - t_0$ **then**

$$(X_{n+1}, Y_{n+1}) = \begin{cases} (Y'_{n+1}, X'_{n+1}), & \text{if GDxLD is applied} \\ (Y'_{n+1}, Y'_{n+1}), & \text{if nGDxLD is applied} \end{cases};$$

else

$(X_{n+1}, Y_{n+1}) = (X'_{n+1}, Y'_{n+1})$.

end

end

Output: X_N as an optimizer for F.

where W_t is a Brownian motion independent of B_t. Under certain regularity conditions on F, $\{X_t\}_{t\geq 0}$ has a unique stationary measure π_ν satisfying

$$\pi_\nu(x) \propto \exp\left(-\tfrac{1}{\nu}F(x)\right).$$

Note that the stationary measure is concentrated around the unique global minimum x^*, and the smaller the value of ν, the higher the concentration is. In particular, if $\nu \to 0$, then π_ν is a Dirac measure at x^*. However, from the algorithmic point of view, the smaller the value of ν, the slower the stochastic process converges to its stationary distribution. In practice, we can only sample the LD for a finite amount of time, which gives rise to the tradeoff between the concentration around the global minimum and the convergence rate to stationarity. One idea to overcome this tradeoff is to run two Langevin diffusions with two different temperatures, high and low, in parallel, and "combining" the two in an appropriate way so that we can enjoy the benefit of both high and low temperatures. This idea is known as replica-exchange LD.

Consider

$$dX_t^a = -\nabla F(X_t^a)dt + \sqrt{2\nu}dW_t$$
$$dY_t^a = -\nabla F(Y_t^a)dt + \sqrt{2\gamma}dB_t$$

The way we would connect them is to allow exchange between X_t and Y_t at random times. In particular, we swap the positions of X_t and Y_t according to a state-dependent rate

$$s(x, y; a) := a \exp\left(0 \wedge \left\{(\tfrac{1}{\nu} - \tfrac{1}{\gamma})(F(x) - F(y)\right\}\right).$$

We refer to a as the exchange intensity. The joint process (X_t^a, Y_t^a) has a unique stationary measure

$$\pi_{\nu,\gamma}(x, y) \propto \exp\left(-\tfrac{1}{\nu}F(x) - \tfrac{1}{\gamma}F(y)\right).$$

Based on $\pi_{\nu,\gamma}$, one would want to set ν small so that X_t can exploit local minima, and set γ large so that Y_t can explore the state space to find better minima. We exchange the

positions of the two with high probability if Y_t finds a better local minimum neighborhood to exploit than the one where X_t is currently at.

For optimization purposes, we can send ν to zero. In this case

$$dX_t^a = -\nabla F(X_t^a)dt, \quad s(x,y;a) = a1_{\{F(x)<F(y)\}},$$

and

$$\pi_{0,\gamma}(x,y) \propto \delta_{x^*}(x)\exp\left(-\tfrac{1}{\gamma}F(y)\right).$$

We would also like to send a to infinity to allow exchange as soon as Y_t^a find a better region to explore. However, the processes (X^a, Y^a) do not converge in natural senses when $a \to \infty$, because the number of swap-caused discontinuities, which are of size $O(1)$, will grow without bound in any time interval of positive length. Dupuis et al. (2012) uses a temperature swapping idea to accommodate this. Consider a temperature exchange process

$$d\tilde{X}_t^a = -\nabla F(\tilde{X}_t^a)dt + \sqrt{2\gamma 1_{\{\tilde{Z}_t^a=1\}}}dW_t$$
$$d\tilde{Y}_t^a = -\nabla F(\tilde{Y}_t^a)dt + \sqrt{2\gamma 1_{\{\tilde{Z}_t^a=0\}}}dB_t,$$

where $\{\tilde{Z}_t^a\}_{t\geq 0}$ is a Markov jump process that switches from 0 to 1 at rate $a1_{\{F(x)>F(y)\}}$ and 1 to 0 at rate $a1_{\{F(x)<F(y)\}}$. Now, sending a to infinity, we have

$$d\tilde{X}_t^\infty = -\nabla F(\tilde{X}_t^\infty)dt + \sqrt{2\gamma 1_{\{F(\tilde{X}_t^\infty)>F(\tilde{Y}_t^\infty)\}}}dW_t$$
$$d\tilde{Y}_t^\infty = -\nabla F(\tilde{Y}_t^\infty)dt + \sqrt{2\gamma 1_{\{F(\tilde{X}_t^\infty)<F(\tilde{Y}_t^\infty)\}}}dB_t.$$

In actual implementations, we will not be able to exactly sample the continuous-time processes. We thus implement a discretized version of it as described in Algorithm 1.

2.3 Complexity bound for GDxLD and nGDxLD

We next present the performance bound for GDxLD and nGDxLD. We start by introducing some assumptions on F.

First we need the function to be smooth in the following sense:

Assumption 1 *F is Lipschitz continuous with Lipschitz constant $L \in (0,\infty)$, i.e.,*

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

Second, we need some conditions under which the iterates or their function values will not diverge to infinity. The following assumption ensures the gradient will push the iterates back once they get too large:

Assumption 2 *The utility function is coercive. In particular, there exist constants $\lambda_c, M_c \in (0,\infty)$, such that*

$$\|\nabla F(x)\|^2 \geq \lambda_c F(x) - M_c,$$

and $F(x) \to \infty$ when $\|x\| \to \infty$.

Note that another more commonly used definition of coerciveness (or dissipation) (Raginsky et al., 2017; Xu et al., 2018) is

$$-\langle \nabla F(x), x\rangle \leq -\lambda_0 \|x\|^2 + M_0. \tag{5}$$

The condition (5) is stronger than Assumption 2. In general, (5) can be enforced by adding proper regularizations. These are explained by Lemma 1. Its proof can be found in Appendix A.

Lemma 1 *Under Assumption 1,*

1. *For any $\lambda > 0$,* (5) *holds for $F_\lambda(x) = F(x) + \lambda\|x\|^2$.*
2. *Suppose* (5) *holds, then Assumption 2 holds.*

The last condition we need is that F is twice differentiable and strongly convex near the global minimum x^*. This allows GD to find x^* efficiently when it is close to x^*.

Assumption 3 *x^* is the unique global minimum for F. There exists $r_0 > 0$, such that the sub-level set $B_0 = \{x : F(x) \leq F(x^*) + r_0\}$ is star-convex with x^* being the center, i.e., a line segment connecting x^* and any $x \in B_0$ is in B_0. There exists $m, M, r > 0$ such that the Hessian $\nabla^2 F$ exists for $B_0 \subset \{x : \|x - x^*\| < r\}$, and for any vector v*

$$m\|v\|^2 \leq v^T \nabla^2 F(x) v \leq M\|v\|^2, \quad \forall x : \|x - x^*\| < r.$$

Note that when Assumption 1 holds, $m \leq M \leq L$.

Assumption 3 is a mild one in practice, since checking the Hessian matrix is positive definite is often the most direct way to determine whether a point is a local minimum. Figuratively speaking, Assumption 3 essentially requires that F has only one global minimum x^* at the bottom of a "valley" of F. It is important to emphasize that we do not assume knowing where this "valley" is, otherwise it suffices to run GD within this "valley". Moreover, when the LD process enters a "valley", there is no mechanism required to detect whether this "valley" contains the global minimum. In fact, designing such detection mechanism can be quite complicated, since it is similar to finding the optimal solution in a multi-arm bandit problem. Instead, for GDxLD, we only need to implement the simple exchange mechanism.

Requiring B_0 to be star-convex rules out the scenario where there are multiple local minima taking values only ϵ-apart from the global optimal value where ϵ can be arbitrarily small. In this scenario, GDxLD will have a hard time to identify the true global local minima. Indeed, GD iterates may stuck at one of the local minima, and the exchange will take place only when LD iterate is $o(\epsilon)$ away from x^*, which will happen rarely due to the noisy movement of LD. On the other hand, finding solutions whose functional values are only ϵ away from the global optimal is often considered good enough in practice.

When $\nabla^2 F$ is not available, or F is only convex near x^*, we can also use the following more general version of Assumption 3. Admittedly the complexity estimate under it will be be worse than the one under Assumption 3.

Assumption 4 *x^* is the unique global minimum for F. There exists $r_0 > 0$, such that the sub-level set $B_0 = \{x : F(x) \leq F(x^*) + r_0\}$ is star-convex with x^* being the center, i.e., a line segment connecting x^* and any $x \in B_0$ is in B_0. F is convex in B_0. In addition, there exist $0 < r_l < r_u < \infty$, such that $\{\|x - x^*\| \leq r_l\} \subset B_0 \subset \{\|x - x^*\| \leq r_u\}$.*

With all assumptions stated, the complexity of GDxLD and nGDxLD is given in the following theorem.

Theorem 2 *Consider the iterates following Algorithm 1 ((n)GDxLD). Under Assumptions 1, 2, 3, and fixed* $\gamma = O(1)$, $0 < h \leq \frac{1}{2L}$, $0 \leq t_0 \leq r_0/8$, *for any* $\epsilon > 0$ *and* $\delta > 0$, *there exists* $N(\epsilon, \delta) = O(-\log \epsilon) + O(-\log \delta)$ *such that for any* $n \geq N(\epsilon, \delta)$,

$$\mathbb{P}(F(X_n) - F(x^*) \leq \epsilon) \geq 1 - \delta. \tag{6}$$

Alternatively, under Assumptions 1, 2, 4, and fixed $\gamma = O(1)$, $0 < h \leq \min\{\frac{1}{2L}, \frac{r_u^2}{r_0}\}$, (6) *holds with* $N(\epsilon, \delta) = O(\epsilon^{-1}) + O(\log(1/\delta))$.

In particular, if we hold δ fixed, to achieve an ϵ accuracy, the complexity is $O(\log(1/\epsilon))$ when F is strongly convex in B_0 and $O(\epsilon^{-1})$ when F is convex in B_0. These complexity estimates are of the same order as GD in the convex setting. However, F does not need to be convex globally for our results to hold.

The fast convergence rate comes partly from the fact that GDxLD does not require the hyper-parameters h and γ to change with the error tolerance ϵ. This is quite unique when compared with other "single-copy" stochastic optimization algorithms. This feature is of great advantage in both practical implementations and theoretical analysis.

Finally, as pointed out in the introduction, our analysis and subsequent results focus on a fixed dimension setting. The big O terms in the definition of $N(\epsilon, \delta)$ "hide" constants that are independent of ϵ and δ. Such constants can scale exponentially with d and $\frac{1}{\gamma h}$. In particular, the $O(\log(1/\delta))$ term is associated with how long it takes the LD to visit B_0, which is further determined by how fast LD converges to stationarity. With a fixed value of γ, the convergence speed in general scales inversely with the exponential of d (see the proof of Theorem 2 in Appendix A for more details). Curse of dimensionality is a common issue for sampling-based optimization algorithms. In (Raginsky et al., 2017; Xu et al., 2018), this is mentioned as a problem of the spectral gap. A more detailed discussion can also be found in Ma et al. (2019).

2.4 Exploration speed of LD

Intuitively, the efficiency of (n)GDxLD depends largely on how often LD visits the neighborhood of x^*, B_0. A standard way to address this question is to study the convergence rate of LD to stationarity. Let

$$\pi_\gamma \propto \exp(-F(x)/\gamma),$$

which denotes the "target" invariant distribution, i.e., it is the invariant distribution of the Langevin diffusion defined in (4). We also write U_γ as the normalizing constant of π_γ.

In the literature, the convergence of a stochastic processes is usually obtained using either a small set argument or a functional inequality argument. Theorem 2 uses the small set argument. In particular, in Lemma 9, we show that each "typical" LD iterate can reach B_0 with a probability lower bounded by $\alpha > 0$. While such α is usually not difficult to obtain, it is often pessimistically small and does not reveal how the convergence depends on F. In contrast, the functional inequality approach usually provides more details about how the convergence depends on F. In particular, using functional inequalities like log Sobolev

inequality (LSI) or Poincaré inequality, one can show that the convergence speed of LD to π_γ depends on the pre-constant of the functional inequality. This pre-constant usually only depends on the target density π_γ, i.e., F, and can be seen as an intrinsic measure of how difficult it is to sample π_γ. In addition, the convergence speed derived from this approach is in general better. For example, Proposition 2.8 in Hairer et al. (2014) indicates that the convergence speed estimate from the small set argument is in general smaller than the one derived from Poincaré inequality. Therefore, it is of theoretical importance to show how the computational complexity estimates of (n)GDxLD depend on the pre-constant of the functional inequality. In this paper, we use LSI as it controls the convergence speed of LD in Kullback-Leibler (KL) divergence. Note that in general, KL divergence between two distributions has a linear dependence on the dimension, whereas the χ^2 divergence, which is associated with Poincaré inequality, often scales exponentially in the dimension.

Let μ_t denote the distribution of Y_t, i.e., the Langevin diffusion (4) at time t. Let $\hat{\mu}_n$ denote the distribution of Y_n, i.e., the discrete LD update (3) at step n. For two measures, μ and π, with $\mu \ll \pi$, the KL divergence of μ with respect to π is defined as

$$\mathrm{KL}(\mu\|\pi) = \int \log\left(\frac{\mu(x)}{\pi(x)}\right)\mu(x)dx.$$

We impose the following assumption on π_γ.

Assumption 5 *π_γ satisfies a log Sobolev inequality with pre-constant $\beta > 0$:*

$$\int \pi_\gamma(x) f(x)^2 \log f(x)^2 \leq \frac{2}{\beta}\mathbb{E}_{\pi_\gamma}\|\nabla f\|^2.$$

It is well known that with LSI, the Langevin diffusion (4) satisfies

$$\mathrm{KL}(\mu_t|\pi_\gamma) \leq \exp(-2\gamma\beta t)\mathrm{KL}(\mu_0|\pi).$$

In addition, the LD (3) satisfies a similar convergence rate (Vempala and Wibisono, 2019):

$$\mathrm{KL}(\hat{\mu}_n|\pi) \leq \exp(-\gamma\beta nh)\mathrm{KL}(\mu_0|\pi) + \frac{8hdL^2}{\beta\gamma}.$$

Adapting these results, we can establish the following convergence results for (n)GDxLD.

Theorem 3 *Consider the iterates following Algorithm 1 ((n)GDxLD). Suppose Assumptions 1, 2, 3, and 5 hold. In addition, for GDxLD, $t_0 > 0$. Fix*

$$h < \left(\frac{\pi_\gamma(B_0)}{2R_V + 6\sqrt{d}L/\sqrt{\beta\gamma}}\right)^2,$$

where $R_V = 8(M_c/4 + 4\gamma Ld)/\lambda_c$. For any $\epsilon > 0$ and $\delta > 0$, set

$$K = \frac{\log(\delta)}{\log(1 - \pi_\gamma(B_0)/2)} \text{ and } n_0 = \frac{1}{\beta\gamma}\frac{1}{h}\log\left(\frac{1}{h}\right).$$

There exists

$$N(\beta, \epsilon, \delta) = O(Kn_0) + O(\log(1/\epsilon)) = O(\log(1/\delta)/\beta) + O(\log(1/\epsilon))$$

such that for $n \geq N(\beta, \epsilon, \delta)$

$$\mathbb{P}(|X_n - x^*| \leq \epsilon) > 1 - \delta.$$

The big O terms in the definition of $N(\beta, \epsilon, \delta)$ in Theorem 3 "hide" constants that have polynomial dependence on d, $1/\pi_\gamma(B_0)$, and $\log U_\gamma$ (see the proof of Theorem 3 in Appendix A.4 for more details). The complexity bound in Theorem 3 can be viewed as a refinement of the bound established in Theorem 2, where the hidden constant scales exponentially with d and $\frac{1}{\gamma h}$. This is due to the fact that the LSI constant β characterizes how difficult it is to sample π_γ. To interpret this refined complexity bound, we note that n_0 represents the time it takes the LD to sample the approximate stationary distribution. If we check whether the LD is in B_0 every n_0 steps, K is the number of trials we need to run to ensure a larger than $1-\delta$ probability of visiting B_0. Then, the time it takes the LD to visit B_0 with high probability is upper bounded by Kn_0. Note that K scales with $1/\pi_\gamma(B_0)$. In order to have a smaller K, we want to have a larger $\pi_\gamma(B_0)$. This can be achieved by using a smaller γ, which gives rise to a π_γ that is more concentrated on B_0. On the other hand, to achieve a smaller n_0, we want to have a larger γ so the LD can escape local minima and converge to stationarity faster. Thus, there is a non-trivial tradeoff to be made here to optimize the constant term "hidden" in $O(\log(1/\delta)/\beta)$ in the quantification of $N(\beta, \epsilon, \delta)$. However, as will be demonstrated in our numerical experiments in Section 3.1.2, (n)GDxLD achieves good and robust performance for a reasonable range of γ (not too small or too big). This is because in (n)GDxLD, γ does not need to depend on ϵ.

We also note that the analysis of nGDxLD is conceptually simpler than GDxLD, since the LD part is a bona-fide unadjusted Langevin algorithm, so the result of (Vempala and Wibisono, 2019) can be implemented rather straightforwardly. In contrast, the LD part in GDxLD can be swapped with the GD part. Thus, the analysis of GDxLD is more challenging. In particular, we need to impose the assumption that $t_0 > 0$ in Algorithm 1. This technical assumption limits the number of swaps between GD and LD.

2.5 Online optimization with stochastic gradient

In a lot of data science applications, we define a loss function for a given parameter x and data sample s as $f(x, s)$, and the loss function we wish to minimize is the average of $f(x, s)$ over a distribution S. Let

$$F(x) = \mathbb{E}_S[f(x, S)].$$

Since the distribution of S can be complicated or unknown, the precise evaluation of F and the gradient ∇F may be computationally too expensive or practically infeasible. However, we often have access to a large number of samples of S in applications. So given an iterate X_n, we can draw two independent batches of independent and identically distributed (iid) samples, $s_{n,1}, \ldots, s_{n,\Theta}$ and $s'_{n,1}, \ldots, s'_{n,\Theta}$, and use

$$\hat{F}_n(X_n) = \frac{1}{\Theta}\sum_{i=1}^{\Theta} f(X_n, s_{n,i}), \quad \nabla\hat{F}_n(X_n) = \frac{1}{\Theta}\sum_{i=1}^{\Theta} \nabla_x f(X_n, s'_{n,i})$$

to approximate F and ∇F. Here we require $\{s_{n,i}, 1 \le i \le \Theta\}$ and $\{s'_{n,i}, 1 \le i \le \Theta\}$ to be two independent batches, so that the corresponding approximation errors are uncorrelated.

When we replace ∇F with $\nabla\hat{F}_n$ in GD and LD, the resulting algorithms are called SGD and SGLD. They are useful when the data samples are accessible only online: to run the algorithms, we only need to get access to and operate on a batch of data. This is very important when computation or storage capacities are smaller than the size of the data.

To implement the replica exchange idea in the online setting, it is natural to replace GD and LD with their online versions. In addition, we need to pay special attention to the exchange criterion. Since we only have access to $\hat{F}_n$, not F, $\hat{F}_n(X'_{n+1}) > \hat{F}_n(Y_{n+1})$ may not guarantee that $F(X'_{n+1}) > F(Y_{n+1})$. Incorrect exchanges may lead to bad performance, and thus we need to be cautious to avoid that. One way to avoid/reduce incorrect exchanges is to introduce a threshold $t_0 > 0$ when comparing $\hat{F}_n$'s. In particular, if t_0 is chosen to be larger than the "typical" size of approximation errors of $\hat{F}_n$, then, $\hat{F}_n(X'_{n+1}) > \hat{F}_n(Y_{n+1}) + t_0$ indicates that $\hat{F}(X'_{n+1})$ is very "likely" to be larger than $\hat{F}(Y_{n+1})$. Lastly, since the approximation error of $\hat{F}_n(x)$ in theory can be very large when x is large, we avoid exchanges if the iterates are very large, i.e., when $\min\{\|X_n\|, \|Y_n\|\} > \hat{M}_v$ for some large $\hat{M}_v \in (0, \infty)$.

Putting these ideas together, the SGDxSGLD algorithm is given in Algorithm 2:

Algorithm 2: SGDxSGLD and nSGDxSGLD: online optimization

Input: Temperature γ, step size h, number of steps N, initial X_0, Y_0, estimation error parameter Θ (when using batch means, Θ is the batch size, it controls the accuracy of $\hat{F}_n$ and $\nabla\hat{F}_n$), threshold $t_0 \in (0, r_0/8]$, and exchange boundary $\hat{M}_v$.

for $n = 0$ *to* $N-1$ **do**

$X'_{n+1} = X_n - h\nabla\hat{F}_n(X_n)$;

$Y'_{n+1} = Y_n - h\nabla\hat{F}_n(Y_n) + \sqrt{2\gamma h}Z_n$, where $Z_n \sim N(0, I_d)$;

if $\hat{F}_n(Y'_{n+1}) < \hat{F}_n(X'_{n+1}) - t_0$, $\|X'_{n+1}\| \le \hat{M}_V$, *and* $\|Y'_{n+1}\| \le \hat{M}_V$ **then**

$$(X_{n+1}, Y_{n+1}) = \begin{cases} (Y'_{n+1}, X'_{n+1}), & \text{if SGDxSGLD is applied} \\ (Y'_{n+1}, Y'_{n+1}), & \text{if nSGDxSGLD is applied} \end{cases};$$

else

$(X_{n+1}, Y_{n+1}) = (X'_{n+1}, Y'_{n+1})$.

end

end

Output: X_N as an optimizer for F.

2.6 Complexity bound for SGDxSGLD

To implement SGDxSGLD, we require three new hyper-parameters, Θ, t_0 and $\hat{M}_v$. We discuss how they can be chosen next.

First of all, the batch-size Θ controls the accuracy of the stochastic approximations of F and ∇F. In particular, we define

$$\zeta_n(x) := \hat{F}_n(x) - F(x) \text{ and } \xi_n(x) = \nabla\hat{F}_n(x) - \nabla F(x),$$

where $\zeta_n(x)$'s and $\xi_n(x)$'s are independent random noise with $\mathbb{E}[\zeta_n(x)] = \mathbb{E}[\xi_n(x)] = 0$. By controlling the number of samples we generate at each iteration, we can control the accuracy of the estimation, as the variances of the estimation errors are of order $1/\Theta$. We will see in Theorem 4 and the discussions following it that $1/\Theta$ should be of the same order as the error tolerance ϵ. For the simplicity of exposition, we introduce a new parameter

$\theta = O(\epsilon)$ to describe the "scale" of ζ_n and ξ_n. In addition, we assume that the errors have sub-Gaussian tails:

Assumption 6 *There exists a constant $\theta > 0$, such that for any $0 < b < \frac{1}{2\theta}$,*

$$\mathbb{E}\left[\exp\left(a^T\xi_n(x) + b\|\xi_n(x)\|^2\right)\right] \leq \frac{1}{(1-2b\theta)^{d/2}} \exp\left(\frac{\|a\|^2\theta}{2(1-2b\theta)}\right),$$

and for $\|x\|, \|y\| \leq \hat{M}_V$, we have for any $z > 0$,

$$\mathbb{P}(\zeta_n(x) - \zeta_n(y) \geq z) \leq \exp\left(-\frac{z^2}{\theta}\right).$$

Note that Assumption 6 implies

$$\mathbb{E}_n[\|\xi_n(x)\|^2] \leq d\theta.$$

We also remark that Assumption 6 holds if $\xi_n(x) \sim \mathcal{N}(0, \theta I_d)$ and $\zeta_n(x) - \zeta_n(y) \sim \mathcal{N}(0, \frac{1}{2}\theta)$. In practice, Assumption 6 can be verified using Hoeffding inequality or other concentration inequalities if the stochastic gradients are bounded.

Second, the threshold t_0 is related to the shape of the "valley" around x^*. To keep the exposition concise, we set $t_0 \leq r_0/8$ where r_0 is defined in Assumption 4. Heuristically, it can be chosen as a generic small constant such as 10^{-2}.

Lastly, $\hat{M}_V$ is introduced to facilitate theoretical verification of Assumption 6. In other words, if Assumption 6 holds for $\hat{M}_V = \infty$, then we can set $\hat{M}_V = \infty$. More generally, under Assumptions 1 and 2, we set

$$\hat{C}_V = \frac{M_c}{4} + (8\gamma L d + 4L\theta d),\ \hat{R}_V = 8\lambda_c^{-1}\hat{C}_V \text{ and } \hat{M}_V = \sup\{x : F(x) \leq \hat{R}_V\}.$$

In practice, one can set $\hat{M}_V$ as a generic large number.

We are now ready to present the complexity of (n)SGDxSGLD:

Theorem 4 *Consider the iterates following Algorithm 2 ((n)SGDxSGLD). Under Assumptions 1, 2, 3, 6, and fixed $\gamma = O(1)$, $0 < h \leq \min\{\frac{1}{2L}, \frac{r_u^2}{r_0}\}$, and $0 < t_0 \leq r_0/8$, for any $\epsilon > 0$ and $\delta > 0$, there exists $N(\epsilon, \delta) = O(-\log(\delta) - \log(\epsilon))$, such that for any fixed $N > N(\epsilon, \delta)$, there exists $\theta(N, \epsilon, \delta) = O\left(\min\{N^{-1}, \epsilon\delta\}\right)$, and for $\theta \leq \theta(N, \epsilon, \delta)$, we have*

$$\mathbb{P}(F(X_N) - F(x^*) \leq \epsilon) \geq 1 - \delta.$$

In particular, if we hold δ fixed, then to achieve an ϵ accuracy, we need to set the number of iterations $N = O(-\log\epsilon)$ and the batch size $\Theta = O(\theta^{-1}) = O(\epsilon^{-1})$. In this case, the total complexity (including data sample) is $O(N\Theta) = O(\epsilon^{-1}\log\frac{1}{\epsilon})$.

To see where the $O(\epsilon^{-1})$ batch size comes from, we can look at a simple example where $F(x) = \frac{1}{2}x^2$. As this function is strongly convex, we can focus on the SGD part. The iterates in this case takes the form

$$X_{n+1} = X_n - hX_n + h\xi_n(X_n).$$

For simplicity, we assume $\xi_n(X_n)$'s are iid $\mathcal{N}(0, \theta I_d)$. Then, X_n is a linear auto-regress sequence with the invariant measure $\mathcal{N}(0, \frac{h\theta}{2-h} I_d)$. Now, for $\mathbb{E}[F(X_N)] = \frac{h\theta d}{2(2-h)} = O(\epsilon)$ when h is a constant, $\theta = O(\epsilon)$.

Similar to the offline case, we can also characterize how the computational cost depends on the LSI pre-constant. Let

$$\hat{B}_0 = \left\{ x : F(x) \leq F(x^*) + \frac{r_0}{4} \right\}.$$

Theorem 5 *Consider the iterates following Algorithm 2 ((n)SGDxSGLD). Suppose Assumptions 1, 2, 3, 5, and 6 hold. Fix.*

$$h < \left(\frac{\pi_\gamma(\hat{B}_0)}{4\hat{R}_V + 8\sqrt{d}L/\sqrt{\beta\gamma}} \right)^2.$$

For any $\epsilon > 0$ and $\delta > 0$, let

$$K = \frac{\log(\delta)}{\log(1 - \pi_\gamma(\hat{B}_0)/2)} \text{ and } n_0 = \frac{4}{\beta\gamma} h^{-1} \log(1/h) + 1.$$

There exists

$$N(\beta, \epsilon, \delta) = O(Kn_0) + O(\log(1/\epsilon)) = O(\log(1/\delta)/\beta + \log(1/\epsilon)),$$

such that for $N > N(\beta, \epsilon, \delta)$, there exists $\theta(N, \epsilon, \delta) = O(\min\{N^{-1}, \epsilon\delta\})$, and for $\theta < \theta(N, \epsilon, \delta)$,

$$\mathbb{P}(F(X_N) - F(x^*) < \epsilon) > 1 - \delta.$$

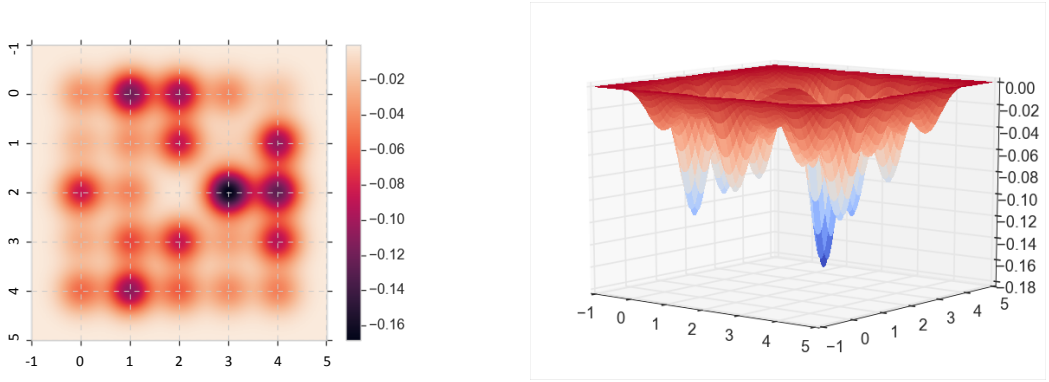
Similar to the offline versions, the big O term in the definition of $N(\beta, \epsilon)$ in Theorem 4 "hides" constants that scale exponentially with d and $\frac{1}{\gamma h}$, while the big O terms in the definition of $N(\beta, \epsilon, \delta)$ in Theorem 5 "hide" constants that have polynomial dependence on d (see the proofs of Theorems 4 and 5 in Appendices B.3 and B.4 for more details). Lastly, it is worth noting that the step size h and temperature γ in (n)SGDxSGLD is independent of the error tolerance ϵ. This is one of the reason why it can beat existing sampling-based algorithms on the convergence speed.

3. Numerical experiments

In this section, we provide some numerical experiments to illustrate the performance of (n)GDxLD and (n)SGDxSGLD. Our main focus is to demonstrate that by doing exchange between the two algorithms, (S)GD and (SG)LD, the performance of the combined algorithm can be substantially better than running isolated copies of the individual algorithms. We also demonstrate the robustness of our algorithm to different choices of the temperature γ and step size h.

3.1 Two-dimensional Problems

We start by looking at two-dimensional examples, which are easier to visualize.



(a) Heat map of F

(b) 3-d plot of F

Figure 2: The objective function F in (7)

3.1.1 Offline setting

First, we consider how to find the mode of a two-dimensional Gaussian-mixture density. The loss function is given by

$$F(x) = -\sum_{i=1}^{M} \frac{w_i}{\sqrt{\det(2\pi\Sigma_i)}} \exp\left(-\tfrac{1}{2}(x-m_i)^T\Sigma_i^{-1}(x-m_i)\right) + L(x). \tag{7}$$

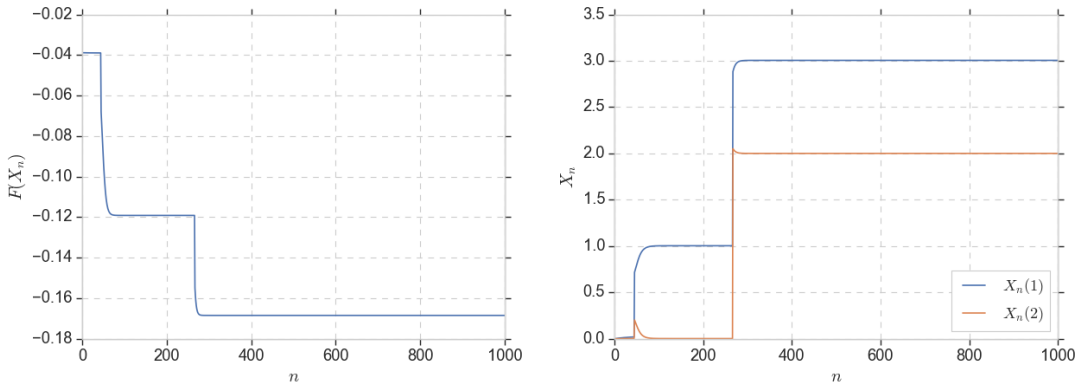
For simplicity, we choose $M = 5 \times 5 = 25$, each m_i is a point in the meshgrid $\{0,1,2,3,4\}^2$, and Σ_i =diag(0.1) so the "valleys" are distinctive. The weights are generated randomly. As the Gaussian-mixture density and its gradient are almost zero for x far away from m_i's, we add a quadratic regularization term

$$L(x) = \sum_{i=1}^{2} \left\{(x(i)+1)^2 1_{x(i)\le -1} + (x(i)-5)^2 1_{x(i)\ge 5}\right\}, \tag{8}$$

where $x(i)$ is the i-th element of x.

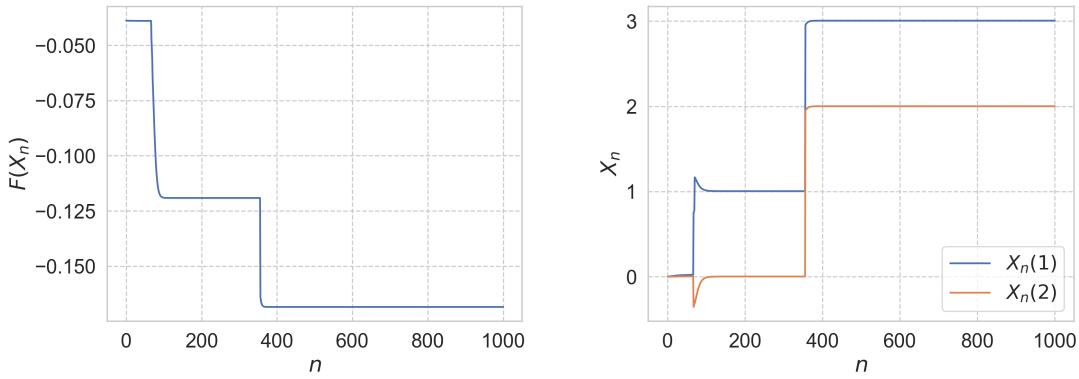
Figure 2 shows the heat map and the 3-d plot of one possible realization of F. We can see that it is highly non-convex with 25 local minima. In this particular realization of F, the global minimum is at $(3,2)$ and $F(3,2) = -0.168$.

We implement GDxLD for the objective function plotted in Figure 2 with $h = 0.1$, $\gamma = 1$, $X_0 = (0,0)$, and $Y_0 = (1,1)$. We plot $F(X_n)$ and X_n at different iterations n in Figure 3. We do not plot Y_n, the sample path of the LD, since it is used for exploration, not optimization. We observe that the convergence happens really fast despite the underlying non-convexity. In particular, we find the global minimum with less than 300 iterations. We run 100 independent copies of GDxLD and are able to find the global minimum within 1000 iterations in all cases. In Figure 4, we plot $F(X_n)$ and X_n at different iterations for a typical nGDxLD implementation. We again observe that the convergence happens really fast, i.e., with less than 400 iterations.



(a) $F(X_n)$ (b) X_n

Figure 3: Convergence of the iterates under GDxLD



(a) $F(X_n)$ (b) X_n

Figure 4: Convergence of the iterates under nGDxLD

For comparison, we also implement GD and LD with the same F. For GD, the iteration takes the form

$$X_{n+1} = X_n - h\nabla F(X_n)$$

with $h = 0.1$. In this case, X_n gets stuck at different local minima depending on where we start, i.e., the value of X_0. For example, Figure 5 plots the trajectories of X_n under GD with $X_0 = (0, 0)$, which is the same as the X_0 we used in GDxLD. As for LD, Figure 6 plots X_n following

$$X_{n+1} = X_n - h\nabla F(X_n) + \sqrt{2\gamma h} Z_n.$$

We set $h = 0.1$ and test two different values of γ: $\gamma = 1$, which is the γ used in GDxLD, and $\gamma = 0.01$. When $\gamma = 1$ (Figure 6 (a)), we do not see convergence to the global minimum at all. The process is doing random exploration in the state-space. When $\gamma = 0.01$ (Figure 6 (b)), we do observe convergence of X_n to the neighborhood of the global minimum. However, compared with GDxLD, the convergence is much slower under LD, since the exploration is slowed down by the small γ. In particular, we find the approximate global minimum with around 1.2×10^5 iterations.

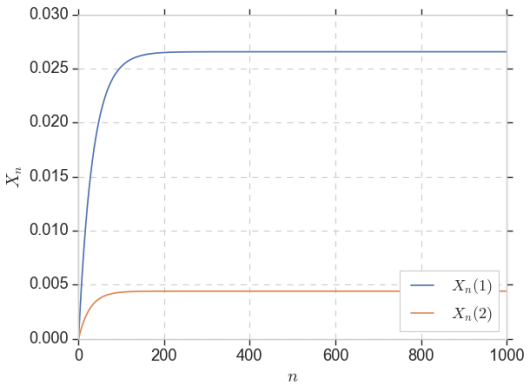
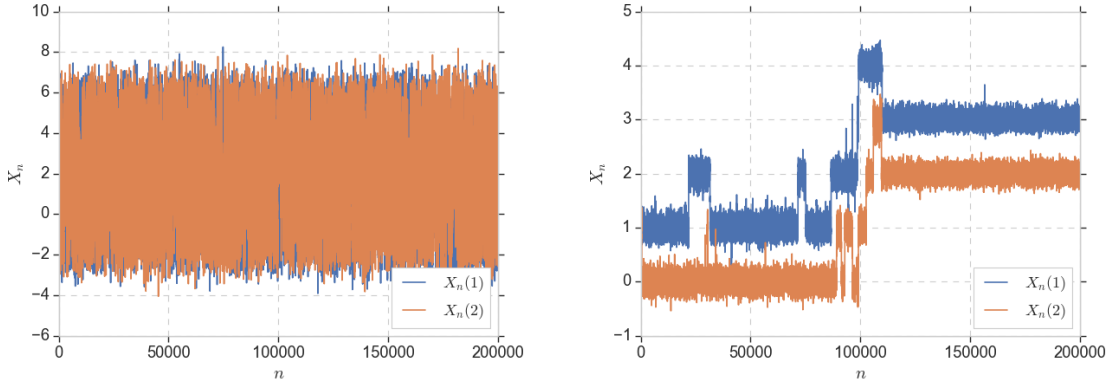


Figure 5: Convergence of the iterates under GD

3.1.2 Sensitivity analysis of the hyper-parameters

One attractive feature of GDxLD is that we do not require the temperature γ and the step size h to change with the precision level ϵ. In the theoretical analysis, we fix them as constants. From a practical point of view, we want γ to be large enough so that it is easy for LD to escape the local minima. On the other hand, we do not want γ to be too large as we want it to focus on exploring the "relevant" region so that there is a good chance of visiting the neighborhood of the global minimum. As for h, we want it to be small enough so that the GD converges once it is in the right neighborhood of the global minimum. On the other hand, we do not want h to be too small, as the convergence rate of GD, when it is in the right neighborhood, increases with h.

In this section, we conduct some sensitivity analysis for different values of γ and h. We use the same objective function as the one plotted in Figure 2.



(a) $\gamma = 1$ (b) $\gamma = 0.01$

Figure 6: Convergence of the iterates under LD

In Figure 7, we keep $h = 0.1$ fixed and vary the value of γ from 0.1 to 100. The left plot shows $\mathbb{E}[F(X_n)]$ estimated based on 100 independent replications of GDxLD. The right plot shows $\mathbb{P}(\|X_n - x^*\| \leq 10^{-3})$, which is again estimated based on 100 independent replications. We observe that as long as γ is not too small or too large, i.e., $0.5 \leq \gamma \leq 10$, GDxLD achieves very fast convergence. For $0.5 \leq \gamma \leq 10$, the convergence speed is slightly different for different values of γ, with $\gamma = 2.5, 5$ among the fastest.

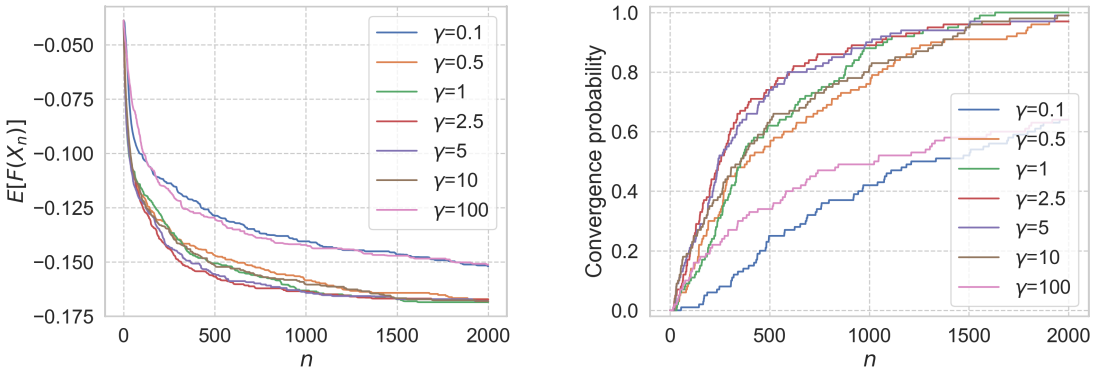


Figure 7: Convergence of the iterates under GDxLD with different values of γ. $h = 0.1$. Averages are estimated based on 100 independent replications.

In Figure 8, we keep $\gamma = 1$ fixed and vary the value of h from 0.1 to 1.5. The left plot shows $\mathbb{E}[F(X_n)]$ and the right plot shows $\mathbb{P}(\|X_n - x^*\| \leq 10^{-3})$. We observe that as long as h is not too big or too small, i.e., $0.05 \leq \gamma \leq 1$, GDxLD achieves very fast convergence. Taking a closer look at the sample path of GDxLD when $h = 1.5$, we note that X_n is swapped to the region around the global minimum fairly quickly but it keeps oscillating around the global minimum due to the large step size. For $0.05 \leq h \leq 1$, the convergence

speed is slightly different for different values of h with $h = 1$ being the fastest and $h = 0.05$ being the slowest.

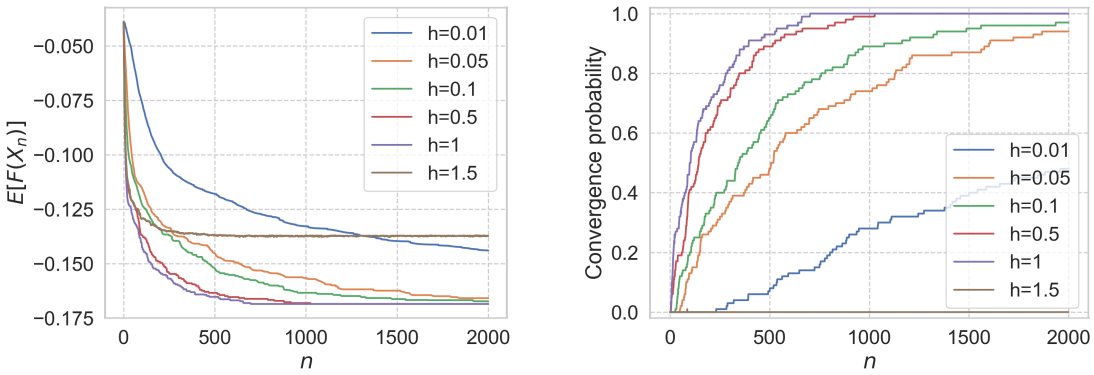


Figure 8: Convergence of the iterates under GDxLD with different values of h. $\gamma = 1$. Averages are estimated based on 100 independent replications.

Above all, our numerical experiments suggest that while different hyper-parameters may lead to different performances of GDxLD, the differences are fairly small as long as γ and h are within a reasonable range. This suggests the robustness of GDxLD to the hyper-parameters.

In Figure 9, we conduct the same analysis for nGDxLD. In particular, we plot $\mathbb{E}[F(X_n)]$ at different iterations for different values of the temperature γ and step size h in nGDxLD. The performances are very similar as those in GDxLD. In all our subsequent numerical experiments, we implement both GDxLD and nGDxLD. Since their performances are very similar, we only show the results for GDxLD in the figures.

3.1.3 Online setting

In this section, we consider an online version of the test problem from Section 3.1.1. In particular, we consider the setting of kernel density estimation (KDE)

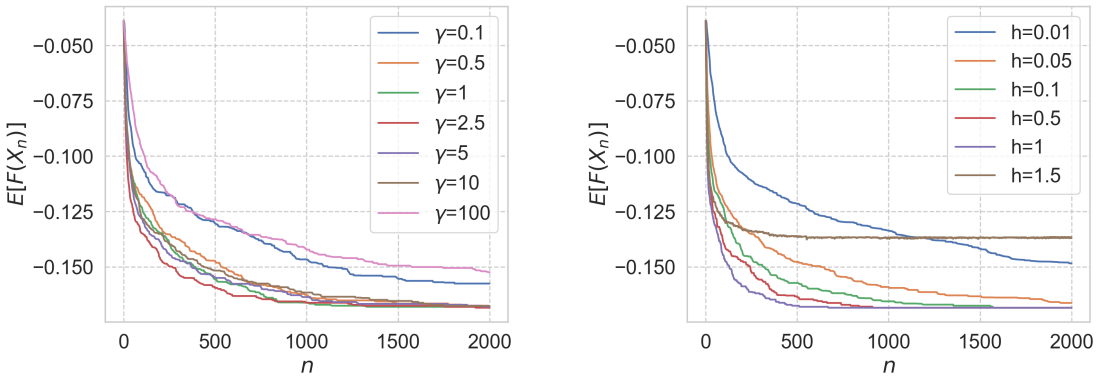
$$\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(x, s_i).$$

κ_σ is known as a kernel function with tuning parameter σ. It measures the similarity between x and the sample data s_i's. There are many choices of kernel functions, and here we use the Gaussian kernel

$$\kappa_\sigma(x, s_i) = \frac{1}{(2\pi\sigma)^{\frac{d}{2}}} \exp(-\tfrac{1}{2\sigma}(x - s_i)^2).$$

Then, $\hat{p}_N(x)$ can be seen as a sample average version of

$$p(x) = \mathbb{E}_s \kappa_\sigma(x, s).$$



(a) $h = 0.1$ (b) $\gamma = 1$

Figure 9: Convergence of the iterates under nGDxLD with different values of γ and h. Averages are estimated based on 100 independent replications.

Notably, p is the density function of $X = S + \sqrt{\sigma}Z$, where S follows the distribution of the sample data and $Z \sim \mathcal{N}(0,1)$. In the following example, we assume S follows a mixture of Gaussian distribution with density

$$f(x) = \sum_{i=1}^{M} \frac{w_i}{\sqrt{\det(2\pi\Sigma_i)}} \exp\left(-\tfrac{1}{2}(x-m_i)^T\Sigma_i^{-1}(x-m_i)\right). \tag{9}$$

As in Section 3.1.1, we set $M = 5\times 5 = 25$, each m_i is a point in the meshgrid $\{0,1,2,3,4\}^2$, Σ_i =diag(0.1), and the weights are randomly generated. Our goal is to find the mode of p. In this case, we write

$$F(x) = -p(x) + L(x) = -\mathbb{E}_s\kappa_\sigma(x,s) + L(x),$$

where L is the quadratic regularization function defined in (8). Then, we can run SGDxSGLD with the mini-batch average approximations of F and ∇F:

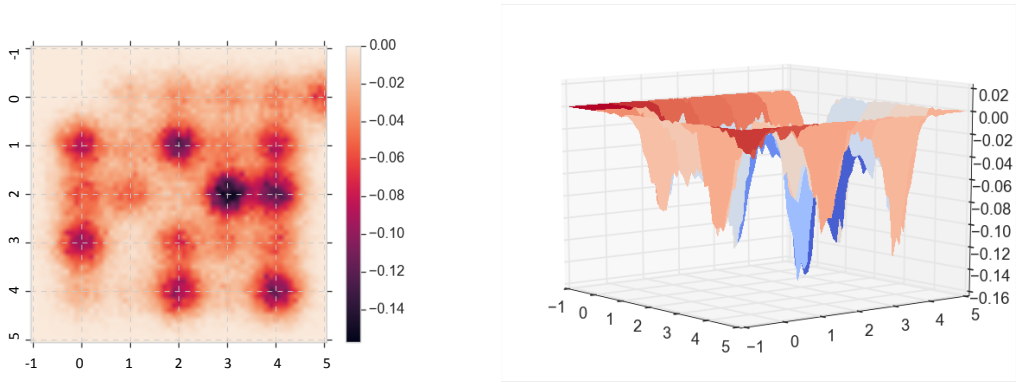
$$\hat{F}_n(X_n) = \frac{1}{\Theta}\sum_{i=1}^{\Theta}\kappa_\sigma(X_n, s_{n,i}) + L(x), \text{ and } \nabla\hat{F}_n(X_n) = \frac{1}{\Theta}\sum_{i=1}^{\Theta}\nabla_x\kappa_\sigma(X_n, s'_{n,i}) + \nabla L(x),$$

where the data-specific gradient takes the form

$$\nabla_x\kappa_\sigma(x,s_i) = \frac{1}{(2\pi\sigma)^{\frac{d}{2}}\sigma}\exp(-\tfrac{1}{2\sigma}|x-s_i|^2)(x-s_i).$$

In Figure 10, we plot the heat map and 3-plot of one possible realization of $\hat{F}_n$ with $\sigma = 0.1^2$ and $n = 10^4$. Note that in this particular realization, the global minimum is achieved at $(3,2)$.

In Figure 11, we plot X_n for different values of n under SGDxSGLD with the objective function plotted in Figure 10. We set $h = 0.1$, $\gamma = 1$, $\Theta = 10^3$, $t_0 = 0.05$, $\hat{M}_v = 5$,



(a) Heat map of $\hat{F}_n$ (b) 3-d plot of $\hat{F}_n$

Figure 10: The estimated objective function

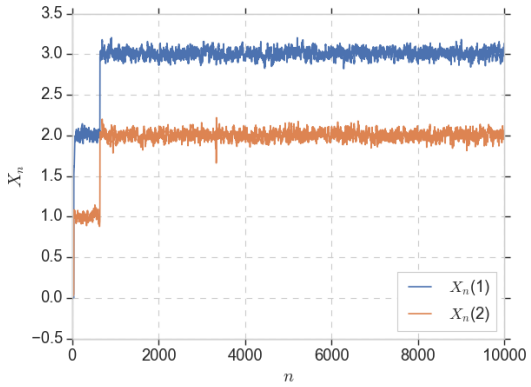
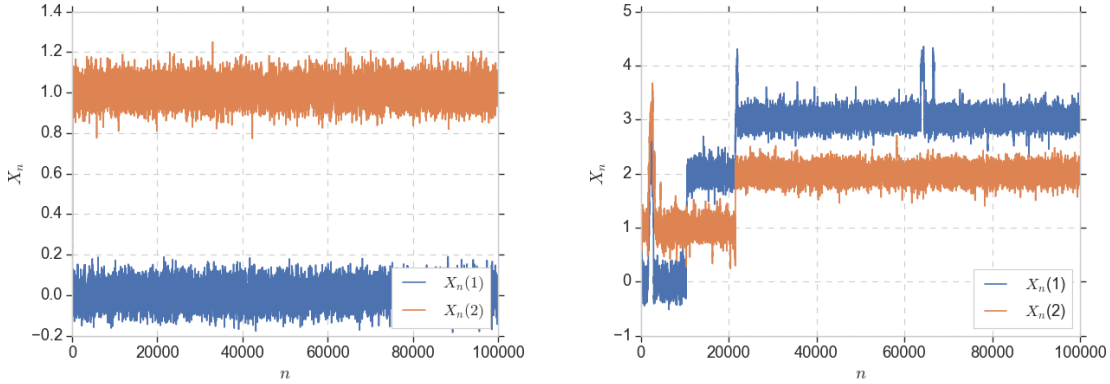


Figure 11: Convergence of the iterates under SGDxSGLD



(a) SGD (b) SGLD

Figure 12: Convergence of the iterates under SGD and SGLD

$X_0 = (0, 0)$, and $Y_0 = (1, 1)$. We observe that SGDxSGLD converges to the approximate global minimum very fast, within 10^3 iterations.

For comparison, in Figure 12, we plot the sample path of X_n under SGD and SGLD with the objective function plotted in Figure 10. For SGD, the iteration takes the form

$$X_{n+1} = X_n - h\nabla\hat{F}(X_n).$$

For SGLD, the iteration takes the form

$$X_{n+1} = X_n - h\nabla\hat{F}(X_n) + \sqrt{2\gamma h}Z_n.$$

We set $h = 0.1$, $\gamma = 0.01$, and $\Theta = 10^3$. Note that $\gamma = 0.01$ is tuned to ensure convergence. We observe that SGD still gets stuck in local minima. For example, in Figure 12 (a), when $X_0 = (0, 0)$, X_n gets stuck at $(0, 1)$. SGLD is able to attain the global minimum, but at a much slower rate than SGDxSGLD. In particular, SGLD takes more than 2×10^4 iterations to converge to the approximate global minimum in Figure 12 (b).

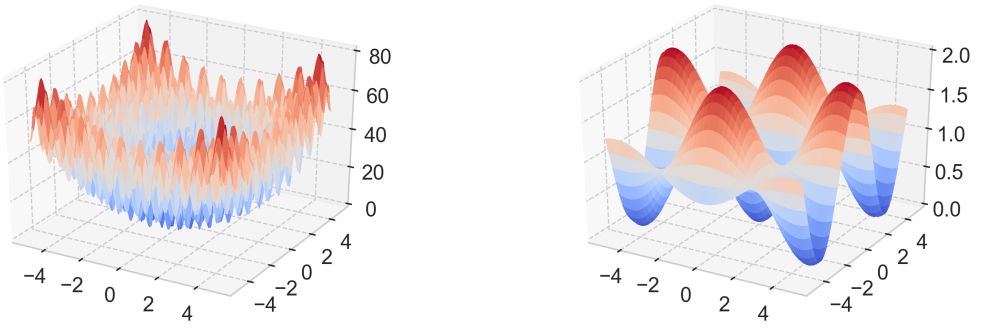
3.2 Non-Convex Optimization Problems with $d \geq 2$

In this section, we demonstrate the performance of (n)GDxLD on some classical non-convex optimization test problems. In particular, we consider the Rastrigin function

$$F(x) = 10n + \sum_{i=1}^{d} \left(x_i^2 - 10\cos(2\pi x_i)\right)$$

restricted to $x \in [-5, 5]^d$. We also consider the Griewank function

$$F(x) = \sum_{i=1}^{d} \frac{x_i^2}{4000} - \prod_{i=1}^{d} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$$



(a) Rastrigin Function (b) Griewank Function

Figure 13: 3-d plot of F when $d = 2$

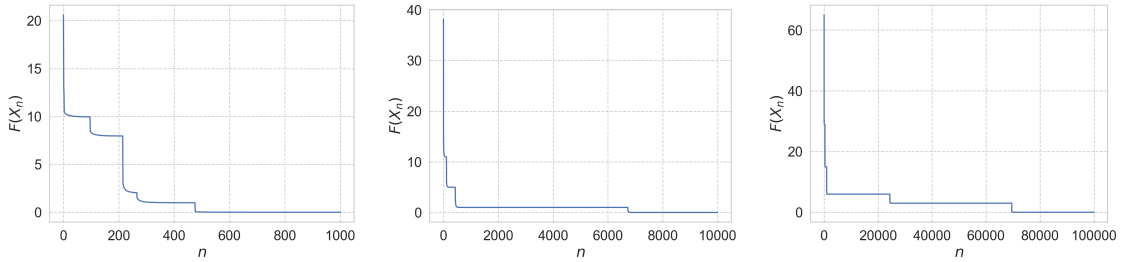
restricted to $x \in [-5, 5]^d$. For both functions, the global minimum is at the origin, i.e., $x^* = (0, \ldots, 0)$, with $F(x^*) = 0$. We select these two functions because they are classic simple test problems for non-convex, multimodal optimization algorithms. Existing tools for optimizing these functions often involve metaheurstics, which lack rigorous complexity analysis (Gämperle et al., 2002; Esquivel and Coello, 2003; Cheng and Jin, 2014). Figure 13 provides an illustration of the two test functions when $d = 2$. We note that the Rastrigin function is especially challenging to optimize as it has many local minima, while the Griewank function restricted to $[-5, 5]^d$ has a relatively smooth landscape with only a few local minima.

For Rastrigin functions of different dimensions, we plot $F(X_n)$ under at different iterations under GDxLD in Figure 15. We observe that as d increases, it takes longer to find the global minimum. For example, when $d = 2$, it takes around 5×10^2 iterations to find the global minim, while when $d = 5$, it takes around 7×10^4 iterations to find the the global minimum. When $d \geq 10$, we are not able to find the global minimum within 10^5 iterations, but the function value reduces substantially. nGDxLD achieves very similar performances as GDxLD. To avoid repetition, we do not include the corresponding plots here.

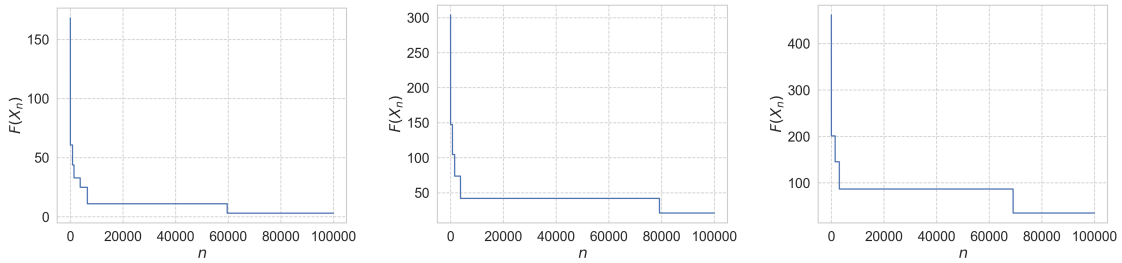
For Griewank functions of different dimensions, we plot $F(X_n)$ at different iterations under GDxLD in Figure 15. We observe that for d as large as 50, the algorithm is able to find the global minimum within 10^4 iterations.

4. Conclusion

GD is known to converge quickly for convex objective functions, but it is not designed for exploration and it can be trapped at local minima. LD is better at exploring the state space. But in order for the stationary distribution of the LD to concentrate around the global minimum, it needs to run with a weak stochastic force, which in general slows down its convergence. This paper considers a novel exchange mechanism to exploit the expertise of both GD and LD. The proposed algorithm, (n)GDxLD, can converge to the

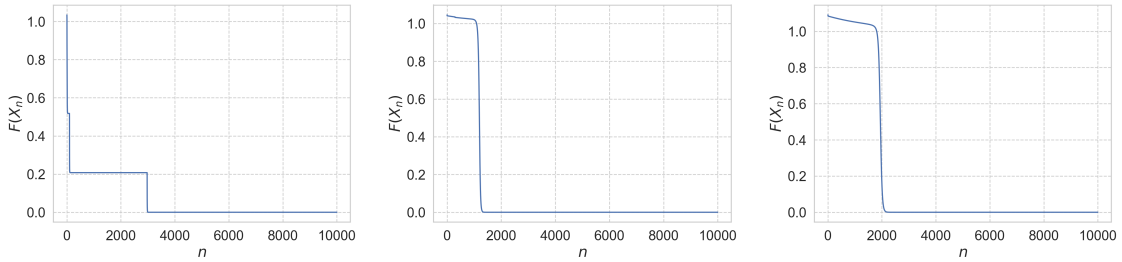


(a) $d = 2$ (b) $d = 3$ (c) $d = 5$



(d) $d = 10$ (e) $d = 20$ (f) $d = 30$

Figure 14: GDxLD for Rastrigin functions on $[-5, 5]^d$. For (a), (b), and (c), $h = 0.005$ and $\gamma = 5$. For (d), (e), and (f), $h = 0.001$ and $\gamma = 5$. The initial values are generated uniformly on $[-5, 5]^d$



(a) $d = 5$ (b) $d = 25$ (c) $d = 50$

Figure 15: GDxLD for Griewank functions on $[-5, 5]^d$. $h = 0.1$ and $\gamma = 5$. For (d), (e), and (f), $h = 0.001$ and $\gamma = 5$. The initial values are generated uniformly on $[-5, 5]^d$

global minimum linearly with high probability for non-convex objective functions, under the assumption that the objective function is strongly convex in a neighborhood of the unique global minimum. Our algorithms can be generalized to online settings. To do so, we replace the exact gradient and function evaluation with their corresponding batch-average versions, and introduce an appropriate threshold for exchange. Lastly, we demonstrate the strength of our algorithms through numerical experiments.

Acknowledgments

We thank anonymous reviewers for their valuable comments and suggestions. The research of Jing Dong is supported by National Science Foundation Grant DMS-1720433. The research of Xin T. Tong is supported by the Singapore Ministry of Education Academic Research Funds Tier 1 grant R-146-000-292-114.

Appendix A. Detailed complexity analysis of (n)GDxLD

In this section, we provide the detailed proof of Theorem 2. The proof uses a constructive stochastic control argument, under which we can "drive" the iterates into the desired neighborhood. We start by providing an overview of the construction, which can be of interests to the analysis of other sampling-based numerical methods. We first note once $X_n \in B_0$, by convexity, X_n will converge to x^* with properly chosen step size h (see details in Lemma 11 and 12). It thus remains to show that X_n will be in B_0 with high probability for n large enough. This task involves two key steps.

Step 1. We construct a proper exponential-type Lyapunov function V with corresponding parameters $C > 0$ and $0 < \rho < 1$ (Lemma 7). In particular, if $\|Y_n - x^*\| > C$ and $\|X_n - x^*\| > C$,

$$\mathbb{E}[V(Y'_{n+1})|Y_n] \le \rho V(Y_n) \text{ and } \mathbb{E}[V(X'_{n+1})|X_n] \le \rho V(X_n)$$

Utilizing this Lyapunov function, we can show for Y_n, the k-th, $k \ge 1$, visit time to the set $\{x : \|x - x^*\| \le C\}$ has a finite moment generating function in a neighborhood of the origin (Lemma 8). This implies that Y_n visits the set $\{x : \|x - x^*\| \le C\}$ quite "often" (i.e., the inter-visit time has a sub-exponential distribution).

Step 2. We then show that during each visit to $\{x : \|x - x^*\| \le C\}$, there is positive probability that Y_n will also visit B_0 (Lemma 9). This essentially creates a sequence of geometric trials whenever $Y_n \in \{x : \|x - x^*\| \le C\}$. Note that once $Y_n \in B_0$, $X_k \in B_0$ for any $k \ge n$ due to the exchange mechanism.

Remark 6 *The positive probability of visiting B_0 in Step 2 can decay exponentially with the dimension d. Therefore, the complexity estimates in Theorem 2 and likewise Theorem 4 can grow exponentially in dimension as well. This is not due to the techniques we are using, as the estimates in Raginsky et al. (2017) and Xu et al. (2018) depend on a quantity called "spectral gap", which can scale exponentially with the dimension as well.*

To facilitate subsequent discussion, we introduce a few more notations. We will use the filtration

$$\mathcal{F}_n = \sigma\{X_k, Y_k, Z_{k-1}, k \le n\},$$

to denote the information up to iteration n. We use $\mathbb{P}_n$ to denote the conditional probability, conditioned on $\mathcal{F}_n$, and $\mathbb{E}_n$ to denote the corresponding conditional expectation. Note that these notations generalize to stopping times.

To keep our derivation concise, we assume $x^* = 0$ and $F(x^*) = 0$. This does not sacrifice any generality, since we can always shift the function and consider

$$F_c(x) = F(x - x^*) - F(x^*).$$

It is easy to check if F satisfy the assumptions introduced in Section 2, F_c also satisfy the assumptions with slightly different constants that depend on x^*.

A.1 Recurrence of the small set

In this section, we provide details about Step 1 and 2 in the proof outline. We start from checking Lemma 1.

Proof [Proof of Lemma 1] For Claim 1), note that

$$-\langle \nabla F_\lambda(x), x\rangle = -2\lambda\|x\|^2 - \langle \nabla F(x), x\rangle.$$

By Assumption 1,

$$\begin{aligned}
|\langle \nabla F(x), x\rangle| &\le \|\nabla F(x)\|\|x\| \text{ by Cauchy-Schwarz inequality} \\
&\le (\|\nabla F(x) - \nabla F(0)\| + \|\nabla F(0)\|)\|x\| \\
&\le (L\|x\| + \|\nabla F(0)\|)\|x\| \text{ by Assumption 1} \\
&= L\|x\|^2 + \|\nabla F(0)\|\|x\|.
\end{aligned}$$

Then, applying Young's inequality, we have

$$\begin{aligned}
-\langle \nabla F_\lambda(x), x\rangle &\le -2\lambda\|x\|^2 + L\|x\|^2 + \|\nabla F(0)\|\|x\| \\
&\le -2\lambda\|x\|^2 + 2L\|x\|^2 + \frac{\|\nabla F(0)\|^2}{L} \\
&\le -\lambda\|x\|^2 + \frac{L^2}{\lambda} + \frac{\|\nabla F(0)\|^2}{L},
\end{aligned}$$

which is of form (5).

For Claim 2), note that

$$\begin{aligned}
&-\langle \nabla F(x), x - x^*\rangle \\
\le& -\lambda_0\|x\|^2 + M_0 + \|\nabla F(x)\|\|x^*\| \text{ by (5)} \\
\le& -\lambda_0\|x\|^2 + M_0 + L\|x - x^*\|\|x^*\| \text{ by Assumption 1} \\
\le& -\lambda_0\|x - x^*\|^2 + \lambda_0\|x^*\|^2 + (2\lambda_0 + L)\|x - x^*\|\|x^*\| + M_0 \\
\le& -\frac{1}{2}\lambda_0\|x - x^*\|^2 + \lambda_0\|x^*\|^2 + \frac{2\lambda_0 + L}{2\lambda_0}\|x^*\|^2 + M_0 \text{ by Young's inequality.}
\end{aligned}$$

Setting $M_1 = \lambda_0\|x^*\|^2 + \frac{2\lambda_0+L}{2\lambda_0}\|x^*\|^2 + M_0$, then,

$$-\langle \nabla F(x), x - x^* \rangle \leq -\frac{1}{2}\lambda_0\|x - x^*\|^2 + M_1.$$

Using Young's inequality again, we have,

$$\frac{1}{\lambda_0}\|\nabla F(x)\|^2 + \frac{1}{4}\lambda_0\|x^* - x\|^2 \geq \langle \nabla F(x), x - x^* \rangle \geq \frac{1}{2}\lambda_0\|x - x^*\|^2 - M_1.$$

Thus,

$$\|\nabla F(x)\|^2 \geq \frac{1}{4}\lambda_0^2\|x - x^*\|^2 - \lambda_0 M_1.$$

■

Our first result provides a proper construction of the Lyapunov function V. It also establishes that $F(X_n)$ is monotonically decreasing.

Lemma 7 *For (n)GDxLD, under Assumption 1, if $Lh \leq 1/2$,*

1. *The value of $F(X_n)$ keeps decreasing:*

$$F(X_{n+1}) \leq F(X'_{n+1}) \leq F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h \leq F(X_n).$$

2. *Assume also Assumption 2, for $\eta \leq (8\gamma)^{-1}$, $V(x) := \exp(\eta F(x))$ satisfies the following:*

$$\begin{aligned}\mathbb{E}_n[F(Y'_{n+1})] &\leq F(Y_n) - \frac{1}{2}\|\nabla F(Y_n)\|^2 h + 2\gamma L h d \\ &\leq (1 - \frac{1}{2}\lambda_c h)F(Y_n) + 2C_V h.\end{aligned}$$

$$\mathbb{E}_n[V(Y'_{n+1})] \leq \exp\left(-\frac{1}{4}\eta h \lambda_c F(Y_n) + \eta h C_V\right) V(Y_n),$$

$$\mathbb{E}_n[V(X_{n+1})] \leq \mathbb{E}_n[V(X'_{n+1})] \leq \exp\left(-\frac{1}{4}\eta h \lambda_c F(X_n) + \eta h C_V\right) V(X_n),$$

where $C_V = M_c/4 + 4\gamma L d$.

Proof For claim 1), note that by Rolle's theorem, there exits x_n on the line segment between X_n and X'_{n+1}, such that

$$\begin{aligned}F(X'_{n+1}) =& F(X_n) - \nabla F(x_n)^T \nabla F(X_n) h \\ =& F(X_n) - \nabla F(X_n)^T \nabla F(X_n) h - (\nabla F(x_n) - \nabla F(X_n))^T \nabla F(X_n) h \\ \leq& F(X_n) - \|\nabla F(X_n)\|^2 h + h\|\nabla F(x_n) - \nabla F(X_n)\|\|\nabla F(X_n)\| \\ & \text{by Cauchy-Schwarz inequality} \\ \leq& F(X_n) - \|\nabla F(X_n)\|^2 h + Lh^2\|\nabla F(X_n)\|^2 \text{ by Assumption 1} \\ \leq& F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h \text{ as } Lh < \tfrac{1}{2}.\end{aligned}$$

Claim 1) then follows, as $F(X'_{n+1}) \le F(X_{n+1})$.

Next, we turn to claim 2). We start by establishing a bound for $F(Y'_{n+1})$. Let $\Delta Y_n = Y'_{n+1} - Y_n = -\nabla F(Y_n)h + \sqrt{2\gamma h}Z_n$. Note that again by Rolle's theorem, there exits y_n on the line segment between Y_n and Y'_{n+1}, such that

$$\begin{aligned}
F(Y'_{n+1}) =& F(Y_n) + \nabla F(y_n)^T \Delta Y_n \\
=& F(Y_n) + \nabla F(Y_n)^T \Delta Y_n - (\nabla F(y_n) + \nabla F(Y_n))^T \Delta Y_n \\
\le& F(Y_n) + \nabla F(Y_n)^T \Delta Y_n + L\|\Delta Y_n\|^2 \\
& \text{by Cauchy-Schwarz inequality and Assumption 1} \\
\le& F(Y_n) - \nabla F(Y_n)^T \nabla F(Y_n)h + \sqrt{2\gamma h}\nabla F(Y_n)^T Z_n \\
& + L\nabla F(Y_n)^T \nabla F(Y_n)h^2 - 2\sqrt{2\gamma h}hL\nabla F(Y_n)^T Z_n + 2L\gamma h Z_n^T Z_n \\
\le& F(Y_n) - \frac{1}{2}\|\nabla F(Y_n)\|^2 h + \beta\sqrt{2\gamma h}\nabla F(Y_n)^T Z_n + 2\gamma Lh\|Z_n\|^2 \text{ as } Lh < \tfrac{1}{2},
\end{aligned}$$

with $\beta = 1 - 2Lh \in (0,1)$. Taking conditional expectation and using Assumption 2, we have our first estimate.

Recall that $V(y) = \exp(\eta F(y))$. Then,

$$\begin{aligned}
\mathbb{E}_n[V(Y'_{n+1})] \le& V(Y_n)\exp\left(-\frac{\eta h}{2}\|\nabla F(Y_n)\|^2\right)\mathbb{E}_n\left[\exp\left(\beta\eta\sqrt{2\gamma h}\nabla F(Y_n)^T Z_n + 2\eta\gamma Lh\|Z_n\|^2\right)\right] \\
=& V(Y_n)\exp\left(-\frac{\eta h}{2}\|\nabla F(Y_n)\|^2\right)(1-4\eta\gamma Lh)^{-d/2}\exp\left(\frac{\beta^2\eta^2\gamma h}{1-4\eta\gamma Lh}\|\nabla F(Y_n)\|^2\right) \\
& \text{as } Z_n \sim \mathcal{N}(0, I_d) \text{ and } 4\eta\gamma Lh < 1/4 \\
\le& V(Y_n)\exp\left(-\frac{\eta h}{2}\|\nabla F(Y_n)\|^2\right)\exp\left(\frac{\eta h}{4}\|\nabla F(Y_n)\|^2 + 8\eta\gamma Lh\frac{d}{2}\right) \\
& \text{as } \eta\gamma < 1/8 \text{ and } \beta < 1 \\
\le& V(Y_n)\exp\left(-\frac{\eta h}{4}\|\nabla F(Y_n)\|^2 + 4\eta\gamma Lhd\right) \\
\le& V(Y_n)\exp\left(-\frac{\eta h}{4}\lambda_c F(Y_n) + \frac{\eta h}{4}M_c + 4\eta\gamma Lhd\right) \text{ by Assumption 2.}
\end{aligned}$$

Similarly, from the derivation of claim 1), we have

$$F(X'_{n+1}) \le F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h.$$

Then,

$$\begin{aligned}
\mathbb{E}_n[V(X'_{n+1})] &\le V(X_n)\exp\left(-\frac{\eta h}{2}\|\nabla F(X_n)\|^2\right) \\
&\le V(X_n)\exp\left(-\frac{\eta h}{4}\lambda_c F(X_n) + \frac{\eta h}{4}M_c\right).
\end{aligned}$$

■

In the following, we set $R_V = 8\lambda_c^{-1}C_V$ and define a sequence of stopping times:

$$\tau_0 = \min\{n \geq 0 : F(X_n) \leq R_V\},$$

and for $k = 1, 2, \ldots,$

$$\tau_k = \min\{n > \tau_{k-1}, F(Y_n) \leq R_V\}.$$

Utilizing the Lyapunov function V, our second result establishes bounds for the moment generating function of τ_k's, $k \geq 0$.

Lemma 8 *For (n)GDxLD, under Assumptions 1 and 2, if $Lh \leq 1/2$ and $\eta \leq (8\gamma)^{-1}$, then for any $K \geq 0$, the stopping time τ_K satisfies*

$$\mathbb{E}[\exp(h\eta C_V \tau_K)] \leq \exp(2Kh\eta C_V + K\eta R_V)(V(X_0) + V(Y_0)).$$

Proof Note that for $n < \tau_0$, $F(Y_n) \geq F(X_n) > R_V = 8\lambda_c^{-1}C_V$. Then, by Lemma 7,

$$\mathbb{E}_n[V(X_{n+1}) + V(Y_{n+1})] \leq \exp(-h\eta C_V)(V(X_n) + V(Y_n)).$$

This implies

$$(V(X_{\tau_0 \wedge n}) + V(Y_{\tau_0 \wedge n}))\exp(h\eta C_V(\tau_0 \wedge n))$$

is a supermartingale. As $V(x) \geq 1$, we have, by sending $n \to \infty$,

$$\mathbb{E}[\exp(h\eta C_V \tau_0)] \leq V(X_0) + V(Y_0).$$

By Lemma 7, $F(X_{n+1}) \leq F(X'_{n+1}) < R_V$ for $n \geq \tau_0$. Therefore, for $k \geq 0$, if $\tau_{k+1} > \tau_k + 1$, $F(Y'_n) > R_V$ and there is no jump for X_n at step n for $\tau_k < n < \tau_{k+1}$.

Given $\mathcal{F}_{\tau_k}$ (starting from τ_k), for $n \leq \tau_{k+1} - 1$, we have $F(Y_n) > R_V$, and by Lemma 7,

$$\mathbb{E}_n[V(Y_{n+1})] \leq \exp(-h\eta C_V)V(Y_n).$$

This implies $V(Y_{\tau_{k+1} \wedge n})\exp(h\eta C_V(\tau_{k+1} \wedge n))$ is a supermartingale. Then, because $V(x) \geq 1$, by sending $n \to \infty$, we have

$$\mathbb{E}_{\tau_k+1}[\exp(h\eta C_V(\tau_{k+1} - \tau_k - 1))] \leq V(Y_{\tau_k+1}).$$

Next,

$$\begin{aligned}
\mathbb{E}_{\tau_k}[\exp(h\eta C_V(\tau_{k+1} - \tau_k - 1))] \leq & \mathbb{E}_{\tau_k}[V(Y_{\tau_k+1})] \\
\leq & \exp\left(-\frac{1}{4}h\eta F(Y_{\tau_k}) + h\eta C_V\right) V(Y_{\tau_k}) \text{ by Lemma 7} \\
\leq & \exp(h\eta C_V + \eta R_V).
\end{aligned}$$

Now because

$$\mathbb{E}[\exp(h\eta C_V \tau_0)] \leq V(X_0) + V(Y_0),$$

we have

$$\mathbb{E}[\exp(h\eta C_V(\tau_K - K))] = \mathbb{E}\left[\prod_{k=0}^{K-1} \exp(h\eta C_V(\tau_{k+1} - \tau_k - 1))\right]$$
$$\leq \left(\prod_{k=0}^{K-1} \exp(h\eta C_V + \eta R_V)\right)(V(X_0) + V(Y_0)).$$

Then,

$$\mathbb{E}[\exp(h\eta C_V \tau_K)] = \exp(2h\eta K C_V + K\eta R_V)(V(X_0) + V(Y_0)).$$

■

Let

$$D = \max\{\|x - h\nabla F(x)\| : F(x) \leq R_V\}.$$

The last result of this subsection shows that if $F(Y_n) \leq R_V$, there is a positive probability that $\|Y'_{n+1}\| \leq r$ for any $r > 0$. In particular, this includes $r = r_l$.

Lemma 9 *For (n)GDxLD, if $F(Y_n) \leq R_V$, for any $r > 0$, there exist an $\alpha(r, D) > 0$, such that*

$$\mathbb{P}_n(\|Y'_{n+1}\| \leq r) > \alpha(r, D).$$

In particular, a lower bound for $\alpha(r, D)$ is given by

$$\alpha(r, D) \geq \frac{S_d r^d}{(4\gamma h\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\gamma h}(D^2 + r^2)\right) > 0,$$

where S_d is the volume of a d-dimensional unit-ball.

Proof

$$\mathbb{P}_n(\|Y'_{n+1}\| \leq r) = \mathbb{P}_n(\|Y_n - h\nabla F(Y_n) + \sqrt{2\gamma h} Z_n\| \leq r)$$
$$= \mathbb{P}_n\left(\|Z_n - Q_n\| \leq \frac{r}{\sqrt{2\gamma h}}\right),$$

where $Q_n := -(Y_n - h\nabla F(Y_n))/\sqrt{2\gamma h}$. Note that as $F(Y_n) \leq R_V$, $\|Q_n\| \leq D/\sqrt{2\gamma h}$. Thus,

$$\mathbb{P}_n\left(\|Z_n - Q_n\| \leq \frac{r}{\sqrt{2\gamma h}}\right) = \int_{\|z\| \leq r/\sqrt{2\gamma h}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\|z + Q_n\|^2\right) dz$$
$$\geq \int_{\|z\| \leq r/\sqrt{2\gamma h}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\gamma h}(D^2 + r^2)\right) dz$$
$$\geq \frac{S_d r^d}{(4\gamma h\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\gamma h}(D^2 + r^2)\right).$$

■

A.2 Convergence to global minimum

In this subsection, we analyze the "speed" of convergence for $\{X_{n+k} : k \geq 0\}$ to $x^* = 0$ when $X_n \in B_0$. Most of these results are classical. In particular, if we assume $B_0 = R^d$, then these rate-of-convergence results can be found in Nesterov (2005). For self-completeness, we provide the detailed arguments here adapted to our settings. First of all, we show Assumption 3 leads to Assumption 4.

Lemma 10 *Under Assumption 3, Assumption 4 holds with*

$$r_u = r, \quad r_0 = \frac{1}{2}a, \quad r_l = \sqrt{\frac{a}{M}}.$$

Moreover for any $x, y \in B_0$,

$$F(y) - F(x) \geq \langle \nabla F(x), y - x \rangle + \frac{m}{2}\|y - x\|^2. \tag{10}$$

Proof Without loss of generality, we assume $x^* = 0$ and $F(0) = 0$. Let $a = \min_{x:\|x\| \geq r} F(x)$. Then $a > 0$ by our assumption. We choose $r_0 = \frac{1}{2}a, r_u = r$, then $B_0 = \{x : F(x) \leq F(x^*) + r_0\} \subset \{\|x\| \leq r\}$.

Next by Taylor expansion, we know F is convex within B_0 and $\{|x| \leq r\}$. This also leads to B_0 being convex since it is a sublevel set. Moreover for any x so that $\|x\| \leq r$, for some x' on the line between x and 0,

$$F(x) = F(0) + x^T \nabla F(0) + \frac{1}{2}x^T \nabla^2 F(x')x = \frac{1}{2}x^T \nabla^2 F(x')x.$$

So $F(x) \leq \frac{1}{2}M\|x\|^2$. So if we let $r_l = \sqrt{\frac{a}{M}}$, $\{\|x - x^*\| \leq r_l\} \subset B_0$. Finally, using Taylor's expansion leads to (10). ■

Lemma 11 *For (n)GDxLD, under Assumptions 1 and 4, and assuming $h \leq \min\{1/(2L), r_u^2/r_0\}$, if $X_n \in B_0$,*

$$F(X_{n+k}) \leq \frac{1}{1/r_0 + kh/r_u^2}$$

for all $k \geq 0$.

Proof From Lemma 7, we have, if $F(X_n) \leq r_0$, i.e, $X_n \in B_0$, $F(X_{n+1}) \leq r_0$.

We first note for any $k \geq 0$,

$$F(X_{n+k}) \leq \langle \nabla F(X_{n+k}), X_{n+k} \rangle \leq r_u \|\nabla F(X_{n+k})\|, \tag{11}$$

where the first inequality follows by convexity (Assumption 4) and the second inequality follows by Hölder's inequality.

Next, by Lemma 7,

$$F(X_{n+k}) \leq F(X_{n+k-1}) - \frac{1}{2}\|\nabla F(X_{n+k-1})\|^2 h \leq F(X_{n+k-1}) - \frac{h}{2r_u^2}F(X_{n+k-1})^2,$$

where the last inequality follows from (11). This implies

$$\frac{1}{F(X_{n+k})} \geq \frac{1}{F(X_{n+k-1}) - \frac{h}{2r_u^2}F(X_{n+k-1})^2} = \frac{1}{F(X_{n+k-1})} + \frac{1}{2r_u^2/h - F(X_{n+k-1})}.$$

Because $F(X_{n+k}) \leq r_0 < r_u^2/h$ by assumption,

$$\frac{1}{F(X_{n+k})} \geq \frac{1}{F(X_{n+k-1})} + \frac{h}{r_u^2}.$$

Then, by induction, we have

$$\frac{1}{F(X_{n+k})} \geq \frac{1}{F(X_n)} + \frac{kh}{r_u^2} \geq \frac{1}{r_0} + \frac{kh}{r_u^2}.$$

■

Lemma 12 *For (n)GDxLD, under Assumptions 1 and 3, if $F(X_n) \leq r_0$, $Lh \leq 1/2$,*

$$F(X_{n+k}) \leq (1 - mh)^k F(X_n)$$

for all $k \geq 0$.

Proof We first note if $F(x)$ is strongly convex in B_0, for $x \in B_0$,

$$F(x^*) - F(x) \geq \langle \nabla F(x), x^* - x \rangle + \frac{m}{2}||x - x^*||^2.$$

By rearranging the inequality, we have

$$F(x) - F(x^*) \leq \langle \nabla F(x), x - x^* \rangle - \frac{m}{2}||x - x^*||^2 \leq \frac{1}{2m}\|\nabla F(x)\|^2, \tag{12}$$

where the last inequality follows from Young's inequality.

Next, from Lemma 7, we have

$$F(X_{n+1}) \leq F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h \leq (1 - mh)F(X_n),$$

where the second inequality follows from (12). Note that by Lemma 7, $F(X_{n+k}) \leq r_0$ for $k \geq 0$. Thus, by induction, we have

$$F(X_{n+k}) \leq (1 - mh)^k F(X_n).$$

■

Remark 13 *The proof of Lemma 12 deals with $F(X_n)$ and $\nabla F(X_n)$ directly. It is thus easily generalizable to the online setting as the noise is additive (see Lemma 23). In contrast, the proof for Lemma 11 requires investigating $(F(X_n))^{-1}$. Its generalization to the online setting can be much more complicated, as the stochastic noise can make the inverse singular.*

A.3 Proof of Theorem 2

We are now ready to prove the main theorem.

Note from Lemma 11 that if $X_n \in B_0$, for any

$$k \geq \left(\frac{1}{\epsilon} - \frac{1}{r_0}\right)\frac{r_u^2}{h},$$

$F(X_{n+k}) \leq \epsilon$. We set

$$k(\epsilon) = (1/\epsilon - 1/r_0)r_u^2/h = O(\epsilon^{-1}).$$

Next, we study "how long" it takes for X_n to reach the set B_0. From Lemma 9, every time $Y_n \in \{x : F(x) \leq R_V\}$,

$$\mathbb{P}_n(\|X_{n+1}\| \leq r_l) \geq \mathbb{P}_n(\|Y'_{n+1}\| \leq r_l) \geq \alpha(r_l, D) > 0.$$

Then,

$$\mathbb{P}(F(X_{\tau_K+1}) > r_0) = \mathbb{E}\left[\prod_{k=1}^{l} \mathbb{P}_{\tau_k}(F(X_{\tau_k+1}) > r_0)\right] \leq (1 - \alpha(r_l, D))^K.$$

Thus, if

$$K > \frac{\log(\delta/2)}{\log(1 - \alpha(r_l, D))},$$

$\mathbb{P}(F(X_{\tau_K+1}) > r_0) < \delta/2$. We set

$$K(\delta) = \log(\delta/2)/\log(1 - \alpha(r_l, D)) = O(-\log\delta).$$

Lastly, we establish a bound for τ_K. From Lemma 8, we have, by Markov inequality,

$$\mathbb{P}(\tau_K > T) \leq \frac{\mathbb{E}[\exp(h\eta C_V \tau_K)]}{\exp(h\eta C_V T)} \leq \frac{\exp(2Kh\eta C_V + K\eta R_V)(V(X_0) + V(Y_0))}{\exp(h\eta C_V T)}.$$

Thus, if

$$T > -\frac{\log(\delta/2)}{h\eta C_V} + 2K + \frac{KR_V}{hC_V} + \frac{\log(V(X_0) + V(Y_0))}{h\eta C_V}$$

$\mathbb{P}(\tau_K > T) < \delta/2$. We set

$$T(\delta) = -\frac{\log(\delta/2)}{h\eta C_V} + 2K(\delta) + \frac{K(\delta)R_V}{hC_V} + \frac{\log(V(X_0) + V(Y_0))}{h\eta C_V} = O(-\log\delta).$$

Above all, we have, for any $N \geq T(\delta) + k(\epsilon)$,

$$\begin{aligned}\mathbb{P}(F(X_N) > \epsilon) &= \mathbb{P}(F(X_N) > \epsilon, \tau_K > T(\delta)) + \mathbb{P}(F(X_N) > \epsilon, \tau_K < T(\delta)) \\ &\leq \mathbb{P}(\tau_K > T(\delta)) + \mathbb{P}(F(X_{\tau_K+1}) > r_0) \leq \delta.\end{aligned}$$

When Assumption 3 holds, F is strongly convex in B_0, from Lemma 12, if $X_n \in B_0$, for any

$$k \geq \frac{\log(\epsilon) - \log(r_0)}{\log(1 - mh)},$$

$F(X_{n+k}) \leq \epsilon$. In this case, we can set

$$k(\epsilon) = (\log(\epsilon) - \log(r_0))\,/\log(1 - mh) = O(-\log\epsilon). \tag{13}$$

Note that T scales with K and $1/\alpha(r_l, D)$. Lemma 9 shows that $1/\alpha(r_l, D)$ depends exponentially on d and $\frac{1}{\gamma h}$.

A.4 Proof of Theorem 3

The proof of Theorem 3 follows similar lines of arguments as the proof of Theorem 2. In particular, it relies on a geometric trial argument. The key difference is that the success probability of the geometric trial is bounded using the pre-constant of LSI rather than the small set argument in Lemma 9.

Let

$$\begin{aligned}\tilde{Y}_{n+1} &= \tilde{Y}_n - \nabla F(\tilde{Y}_n)h + \sqrt{2\gamma h}Z_n \\ &= \tilde{Y}_n - \frac{1}{\gamma}\nabla F(\tilde{Y}_n)\gamma h + \sqrt{2\gamma h}Z_n\end{aligned}$$

This is to be differentiated with Y_n in Algorithm 1, which can swap position with X_n. Note that $\tilde{Y}_n$ is also known as the unadjusted Langevin algorithm (ULA). We denote $\tilde{\mu}_n$ as the distribution of $\tilde{Y}_n$. Recall that $\pi_\gamma(x) = \frac{1}{U_\gamma}\exp(-\frac{1}{\gamma}F(x))$. We first prove an auxiliary bound for $V(x) = \exp(\eta F(x))$ that will be useful in our subsequent analysis.

Lemma 14 *Under Assumptions 1 and 2, for $\eta \le (8\gamma)^{-1}$, $h \le 1/(2L)$,*

$$\mathbb{E}_n[V(\tilde{Y}_{n+m})] \le V(\tilde{Y}_n) + 4\exp(\eta R_V),$$

where $R_V = 8C_V/\lambda_c$.

Proof From the second claim in Lemma 7, we have

$$\mathbb{E}_n[V(\tilde{Y}_{n+1})] \le \exp\left(-\frac{1}{4}\eta h\lambda_c F(\tilde{Y}_n) + \eta h C_V\right)\exp(\eta F(\tilde{Y}_n)).$$

When $F(\tilde{Y}_n) < R_V$,

$$\begin{aligned}\mathbb{E}_n[V(\tilde{Y}_{n+1})] &\le \exp\left(\eta h C_V\right)\exp(\eta F(\tilde{Y}_n)) \\ &\le \left(1 - \frac{1}{2}\eta h C_V\right)\exp(\eta F(\tilde{Y}_n)) + \left(\exp\left(\eta h C_V\right) - 1 + \frac{1}{2}\eta h C_V\right)\exp(\eta R_V) \\ &\le \left(1 - \frac{1}{2}\eta h C_V\right)\exp(\eta F(\tilde{Y}_n)) + 2\eta h C_V \exp(\eta R_V)\end{aligned}$$

When $F(\tilde{Y}_n) \ge R_V$, we have

$$\mathbb{E}_n[V(\tilde{Y}_{n+1})] \le \exp\left(-\eta h C_V\right)\exp(\eta F(\tilde{Y}_n)) \le \left(1 - \frac{1}{2}\eta h C_V\right)\exp(\eta F(\tilde{Y}_n))$$

In both cases,

$$\mathbb{E}_n[V(\tilde{Y}_{n+1})] \le \left(1 - \frac{1}{2}\eta h C_V\right)V(\tilde{Y}_n) + 2\eta h C_V \exp(\eta R_V),$$

which implies that

$$\mathbb{E}_n[V(\tilde{Y}_{n+m})] \le V(\tilde{Y}_n) + 4\exp(\eta R_V).$$

■

Lemma 15 *Under Assumptions 1 and 5,*

$$KL(\tilde{\mu}_n\|\pi_\gamma) \le \exp(-\beta(n-1)h\gamma)\left(\frac{1}{\gamma}F(\tilde{Y}_0)+\tilde{C}_d\right)+\frac{8hdL^2}{\beta\gamma}$$

where $\tilde{M}_d := 2Lhd - \frac{d}{2}(\log(4\pi h)+1)+\log(U_\gamma)$.

Proof Given $\tilde{Y}_0$,

$$\tilde{Y}_1 \sim \mathcal{N}(m(\tilde{Y}_0), 2\gamma h I_d), \text{ where } m(\tilde{Y}_0) = \tilde{Y}_0 - \nabla F(\tilde{Y}_0)h.$$

Then, given $\tilde{Y}_0$,

$$\begin{aligned}
\mathrm{KL}(\tilde{\mu}_1\|\pi_\gamma) &= \int \log\left(\frac{\tilde{\mu}_1(y)}{\pi_\gamma(y)}\right)\tilde{\mu}_1(y)dy \\
&= \int\left(-\frac{d}{2}\log(4\pi\gamma h) - \frac{\|y-m(\tilde{Y}_0)\|^2}{4\gamma h} + \log U_\gamma + \frac{F(y)}{\gamma}\right)\tilde{\mu}_1(y)dy \\
&= -\frac{d}{2}\log(4\pi\gamma h) - \frac{d}{2} + \log U_\gamma + \frac{1}{\gamma}\mathbb{E}[F(\tilde{Y}_1)|\tilde{Y}_0] \\
&\le -\frac{d}{2}\log(4\pi\gamma h) - \frac{d}{2} + \log U_\gamma + \frac{1}{\gamma}(F(\tilde{Y}_0)+2\gamma Lhd) \text{ by Lemma 7 claim 2.}
\end{aligned}$$

For

$$\tilde{M}_d = -\frac{d}{2}(\log(4\pi\gamma h)+1)+\log U_\gamma + 2Lhd,$$

we have

$$\mathrm{KL}(\tilde{\mu}_1\|\pi_\gamma) \le \frac{1}{\gamma}F(\tilde{Y}_0)+\tilde{M}_d$$

Next, from Theorem 2 in (Vempala and Wibisono, 2019) with $L = L/\gamma$ and $\epsilon = \gamma h$, we have

$$\begin{aligned}
\mathrm{KL}(\tilde{\mu}_n\|\pi_\gamma) &\le \exp(-\beta(n-1)h\gamma)\mathrm{KL}(\tilde{\mu}_1\|\pi_\gamma) + \frac{8hdL^2}{\beta\gamma} \\
&\le \exp(-\beta(n-1)h\gamma)\left(\frac{1}{\gamma}F(\tilde{Y}_0)+\tilde{M}_d\right)+\frac{8hdL^2}{\beta\gamma}.
\end{aligned}$$

■

Based on Lemma 15, let

$$n_0 = \frac{2}{\beta\gamma}h^{-1}\log(1/h)+1.$$

Note that for $n \ge n_0$ and h small enough,

$$\begin{aligned}
\mathrm{KL}(\tilde{\mu}_n\|\pi_\gamma) &\le h^2\left(\frac{1}{\gamma}F(\tilde{Y}_0)+\tilde{M}_d\right)+\frac{8hdL^2}{\beta\gamma} \\
&\le hF(\tilde{Y}_0)+\frac{9hdL^2}{\beta\gamma}.
\end{aligned} \tag{14}$$

We next draw connection between the bound (14) and the hitting time of Y_n to B_0. For nGDxLD, let

$$\phi = \min_n \{Y_n \in B_0\}.$$

With a slight abuse of notation, for GDxLD, let

$$\phi = \min_n \{X_n = Y_n' \text{ or } Y_n \in B_0\}.$$

Lemma 16 *For nGDxLD or GDxLD, under Assumptions 1, 2, and 5,*

$$\mathbb{P}(\phi \le n_0) \ge \pi_\gamma(B_0) - \sqrt{h} F(\tilde{Y}_0) - 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}}.$$

Proof For nGDxLD, $Y_n = \tilde{Y}_n$. For GDxLD, before replica exchange, we also have $Y_n = \tilde{Y}_n$. Therefore, $\mathbb{P}(\phi > n) \le \mathbb{P}(Z_n \notin B_0)$, which further implies that

$$\mathbb{P}(\phi \le n) \ge \mathbb{P}(Z_n \in B_0).$$

By Pinsker's inequality, the KL divergence provides an upper bound for the total variation distance. Let $d_{tv}(\mu, \pi)$ denote the total variation distance between μ and π. Then, for $n \ge n_0$,

$$d_{tv}(\tilde{\mu}_n, \pi_\gamma) \le \sqrt{\frac{1}{2}\mathrm{KL}(\tilde{\mu}_n \| \pi_\gamma)} \le \sqrt{\frac{1}{2} h F(\tilde{Y}_0) + \frac{1}{2}\frac{9hdL^2}{\beta\gamma}} \le \sqrt{h} F(\tilde{Y}_0) + 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}},$$

where the second inequality follows from (14). Thus,

$$\mathbb{P}(\tilde{Y}_{n_0} \in B_0) \ge \pi_\gamma(B_0) - d_{tv}(\tilde{\mu}_n, \pi_\gamma) \ge \pi_\gamma(B_0) - \sqrt{h} F(\tilde{Y}_0) - 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}}.$$

■

Lemma 17 *For (n)GDxLD, fix any K and $T \ge (n_0+1)K$,*

$$\begin{aligned}\mathbb{P}_0(\phi > T) \le & \exp(2(n_0+1)Kh\eta C_V + (n_0+1)K\eta R_V - h\eta C_V T)(V(X_0) + V(Y_0)) \\ & + \left(1 - \pi_\gamma(B_0) + \sqrt{h} R_V + 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}}\right)^K.\end{aligned}$$

Proof Recall the sequence of stopping times $\tau_j = \inf\{n > \tau_{j-1} : F(Y_n) \le R_V\}$. From Lemma 8, we have

$$\mathbb{E}[\exp(h\eta C_V \tau_k)] \le \exp(2kh\eta C_V + k\eta R_V)(V(X_0) + V(Y_0)).$$

We next define a new sequence of stopping times:

$$\psi_0 = \inf\{n : F(Y_n) \le R_V\}, \quad \psi_0' = \psi_0 + n_0,$$

and for $k = 1, \ldots,$

$$\psi_k = \inf\left\{n \geq \psi'_{k-1} + 1, F(Y_n) \leq R_V\right\}, \quad \psi'_k = \psi_k + n_0 + 1.$$

Note that ψ_i always coincide with one of τ_j's, and as $\tau_j - \tau_{j-1} \geq 1$,

$$\psi_k \leq \tau_{(n_0+1)k}.$$

Thus,

$$\begin{aligned}\mathbb{P}_0(\phi \geq T) \leq & \mathbb{P}_0\left(\tau_{(n_0+1)K} \geq T\right) + \mathbb{P}_0\left(Y_{\psi'_k} \notin B_0, \forall k \leq K\right) \\ \leq & \exp(2(n_0+1)Kh\eta C_V + (n_0+1)K\eta R_V - h\eta C_V T)(V(X_0) + V(Y_0)) \\ & + \left(1 - \pi_\gamma(B_0) + \sqrt{h}R_V + 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}}\right)^K \text{ by Lemma 16.}\end{aligned}$$

■

We are now ready to prove Theorem 3. Set

$$h < \left(\frac{\pi_\gamma(B_0)}{2R_V + 6\sqrt{d}L/\sqrt{\beta\gamma}}\right)^2.$$

Then,

$$1 - \pi_\gamma(B_0) + \sqrt{h}R_V + 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}} < 1 - \frac{\pi_\gamma(B_0)}{2}.$$

For nGDxLD, based on Lemma 17, let

$$K(\delta) = \frac{\log(\delta/2)}{\log(1 - \pi_\gamma(B_0)/2)}.$$

For $K \geq K(\delta)$,

$$\left(1 - \pi_\gamma(B_0) + \sqrt{h}R_V + 3\frac{\sqrt{hd}L}{\sqrt{\gamma\beta}}\right)^K \leq \frac{\delta}{2}.$$

In addition, let

$$T(\beta, \delta) = -\frac{\log(\delta/2)}{h\eta C_V} + 2(n_0+1)K(\delta) + \frac{(n_0+1)K(\delta)R_V}{hC_V} + \frac{\log(V(X_0) + V(Y_0))}{h\eta C_V}.$$

Note that $T(\beta, \delta) = O(K(\delta)n_0) = O(\log(1/\delta)/\beta)$. For $T \geq T(\delta)$, $\mathbb{P}(\tau_K > T) \leq \delta/2$. Above all, $\mathbb{P}(\phi > T(\delta)) \leq \delta$.

Lastly, note that once Y_n is in B_0, X_n will be moved there if it is not in B_0 already. After X_n is in B_0, it takes at most $k(\epsilon) = (\log(\epsilon) - \log(r_0)) / \log(1 - mh)$ iterates to achieve the desired accuracy. Therefore, we can set

$$N(\beta, \epsilon, \delta) = T(\beta, \delta) + k(\epsilon).$$

For GDxLD, let $J = \lceil R_v/t_0 \rceil$. We also define

$$\phi_0 = \min\{n \geq 0 : F(X_n) \leq R_V\}$$

and for $k = 1, 2, \ldots,$

$$\phi_k = \min\{n \geq 0 : X_{\phi_{k-1}+n} = Y'_{\phi_{k-1}+n} \text{ or } Y_{\phi_{k-1}+n} \in B_0\}.$$

We next make a few observations. First, for $n \geq \phi_0$, $F(X_n) \leq R_V$. Second, $F(Y_{\phi_k}) \leq R_V$ for $k \geq 1$. Lastly, the value of $F(X_n)$ will decrease by t_0 every time swapping takes place. Thus, after ϕ_0, there can be at most J swapping events. The last observation implies that before time ϕ_J, Y_n must visit B_0 at least once. Let

$$\Xi = \min\{n \geq 0 : Y_n \in B_0 \text{ or } F(X_n) \leq \epsilon\}.$$

Then,

$$\begin{aligned}
&\mathbb{P}_0(\Xi > (J+1)T) \\
\leq &\mathbb{P}_0\left(\sum_{k=0}^{J} \phi_k > T\right) \\
\leq &\mathbb{E}_{\phi_0}\left[\sum_{k=0}^{J} \mathbb{P}_{\phi_{k-1}}(\phi_k > T)\right] \text{ where } \phi_{-1} \equiv 0 \\
\leq &(J+1)\exp(2(n_0+1)Kh\eta C_V + (n_0+1)K\eta R_V - h\eta C_V T)(V(X_0) + V(Y_0) + 2R_V) \\
&+ (J+1)\left(1 - \pi_\gamma(B_0) + \sqrt{h}R_V + 3\frac{\sqrt{hd}L}{\sqrt{\beta\gamma}}\right)^K \\
&\text{by Lemma 17 and the fact that } V(X_{\phi_k}) \leq R_V,\ V(Y_{\phi_k}) \leq R_V.
\end{aligned}$$

Let

$$\tilde{K}(\delta) = \frac{\log(\delta) - \log(4(J+1))}{\log(1 - \pi_\gamma(B_0)/2)}$$

and

$$\begin{aligned}
T(\beta, \delta) = &- \frac{\log(\delta) - \log(4(J+1))}{h\eta C_V} + 2(n_0+1)K(\delta) + \frac{(n_0+1)K(\delta)R_V}{hC_V} \\
&+ \frac{\log(V(X_0) + V(Y_0) + 2R_V)}{h\eta C_V}.
\end{aligned}$$

For $T > T(\beta, \delta)$, $\mathbb{P}(\Xi > (J+1)T) \leq \delta/2$. In this case, we can set

$$N(\beta, \epsilon, \delta) = (J+1)T(\beta, \delta) + k(\epsilon).$$

Appendix B. Detailed complexity analysis of (n)SGDxSGLD

In this section, we provide the proof of Theorem 4. The proof follows a similar construction as the proof of Theorem 2. However, the stochasticity of $\hat{F}(x)$ and $\nabla\hat{F}(x)$ substantially complicates the analysis.

To facilitate subsequent discussions, we start by introducing some additional notations. We denote

$$\zeta_n^X = \hat{F}_n(X_n) - F(X_n) \text{ and } \zeta_n^Y = \hat{F}_n(Y_n) - F(Y_n).$$

Similarly, we denote

$$\xi_n^X = \nabla\hat{F}_n(X_n) - \nabla F(X_n) \text{ and } \xi_n^Y = \nabla\hat{F}_n(Y_n) - \nabla F(Y_n).$$

We also define

$$\mathcal{F}_n = \sigma\{X_k, Y_k, Z_{k-1}, k \le n\},$$

and

$$\mathcal{G}_n = \mathcal{F}_n \vee \sigma\{X'_{n+1}, Y'_{n+1}\}.$$

We use $\tilde{\mathbb{P}}_n$ to denote the conditional probability, conditioned on $\mathcal{G}_n$, and $\tilde{\mathbb{E}}_n$ to denote the corresponding conditional expectation.

Following the proof of Theorem 2, the proof is divided into two steps. We first establish the positive recurrence of some small sets centered around the global minimum. We then establish convergence to the global minimum conditional on being in the properly defined small set. Without loss of generality, we again assume $x^* = 0$ and $F(x^*) = 0$.

B.1 Recurrence of the small set

Our first result establishes some bounds for the decay rate of $F(X_n)$.

Lemma 18 *For (n)SGDxSGLD, under Assumptions 1 and 6, if $Lh \le 1/2$,*

1. *The value of $F(X_n)$ keeps decreasing on average:*

$$\mathbb{E}_n[F(X'_{n+1})] \le F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h + dLh^2\theta.$$

If $\theta \le -\frac{t_0^2}{\log(2Lh^2 t_0)}$,

$$\mathbb{E}_n[F(X_{n+1})] \le F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h + 2dLh^2\theta.$$

2. *Assume also Assumption 2, for $\hat{\eta} \le \min\{(16\gamma)^{-1}, (8h\theta)^{-1}\}$, $\hat{V}(x) := \exp(\hat{\eta}F(x))$ satisfies the following:*

$$\mathbb{E}_n[F(Y_{n+1})] \le (1 - \frac{1}{2}\lambda_c h)F(Y_n) - \frac{1}{2}\|\nabla F(Y_n)\|^2 h + 2C_V h.$$

$$\mathbb{E}_n[\hat{V}(Y'_{n+1})] \le \exp\left(-\frac{1}{4}\hat{\eta}h\lambda_c F(Y_n) + \hat{\eta}h\hat{C}_V\right)\hat{V}(Y_n),$$

$$\mathbb{E}_n[\hat{V}(X'_{n+1})] \le \exp\left(-\frac{1}{4}\hat{\eta}h\lambda_c F(X_n) + \hat{\eta}h\hat{C}_V\right)\hat{V}(X_n),$$

where $\hat{C}_V = M_c/4 + (8\gamma Ld + 4Lh\theta d)$.

3. If $\hat{\eta} < \min\left\{(8h\theta)^{-1}, 2t_0/\theta\right\}$, *for* $\theta \leq -\frac{t_0^2}{\log(\exp(\hat{\eta}h\theta d)-1)}$,

$$\mathbb{E}_n[\hat{V}(X_{n+1})] \leq \hat{V}(X_n)\exp\left(\tfrac{1}{4}d\right).$$

Proof For Claim 1), by Rolle's theorem, there exits x_n on the line segment between X_n and X'_{n+1}, such that

$$\begin{aligned}
\mathbb{E}_n[F(X'_{n+1})] =& F(X_n) - \mathbb{E}_n\left[\nabla F(x_n)^T(\nabla F(X_n) + \xi_n^X)\right]h \\
=& F(X_n) - \mathbb{E}_n\left[\nabla F(X_n)^T(\nabla F(X_n) + \xi_n^X)\right]h \\
&+ \mathbb{E}_n\left[(\nabla F(x_n) - \nabla F(X_n))^T(\nabla F(X_n) + \xi_n^X)\right]h \\
\leq& F(X_n) - \|\nabla F(X_n)\|^2 h \\
&+ h\left(\mathbb{E}_n\left[\|\nabla F(x_n) - \nabla F(X_n)\|^2\right]\right)^{1/2}\left(\mathbb{E}_n\left[\|\nabla F(X_n) + \xi_n^X\|^2\right]\right)^{1/2} \\
&\text{by Hölder inequality} \\
\leq& F(X_n) - \|\nabla F(X_n)\|^2 h + Lh^2\mathbb{E}_n\left[\|\nabla F(X_n) + \xi_n^X\|^2\right] \text{ by Assumption 1} \\
\leq& F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h + dLh^2\theta \\
&\text{as } Lh < \tfrac{1}{2} \text{ and } \mathbb{E}_n[\|\xi_n^X\|^2] \leq d\theta \text{ by Assumption 6.}
\end{aligned}$$

For X_{n+1}, we first note that when $\|X'_{n+1}\| > \hat{M}_V$ or $\|Y'_{n+1}\| > \hat{M}_V$, $F(X_{n+1}) = F(X'_{n+1})$. When $\|X'_{n+1}\| \leq \hat{M}_V$ and $\|Y'_{n+1}\| \leq \hat{M}_V$, we note that if $F(Y'_{n+1}) \leq F(X'_{n+1})$, then $F(X_{n+1}) \leq F(X'_{n+1})$. If $F(Y'_{n+1}) > F(X'_{n+1})$, we may "accidentally" move X_{n+1} to Y'_{n+1} due to the estimation errors. In particular,

$$\begin{aligned}
\tilde{\mathbb{E}}_n[F(X_{n+1})] =& F(X'_{n+1})\tilde{\mathbb{P}}_n(F_n(X'_{n+1}) \leq F_n(Y'_{n+1}) + t_0) \\
&+ F(Y'_{n+1})\tilde{\mathbb{P}}_n(F_n(X'_{n+1}) > F_n(Y'_{n+1}) + t_0).
\end{aligned}$$

This implies

$$\begin{aligned}
&\tilde{\mathbb{E}}_n[F(X_{n+1})] - F(X'_{n+1}) \\
=& (F(Y'_{n+1}) - F(X'_{n+1}))\tilde{\mathbb{P}}_n(\hat{F}_n(X'_{n+1}) > \hat{F}_n(Y'_{n+1}) + t_0) \\
=& (F(Y'_{n+1}) - F(X'_{n+1}))\tilde{\mathbb{P}}_n(\zeta_{n+1}^X - \zeta_{n+1}^Y > F(Y'_{n+1}) - F(X'_{n+1}) + t_0) \\
\leq& (F(Y'_{n+1}) - F(X'_{n+1}))\exp(-(F(Y'_{n+1}) - F(X'_{n+1}) + t_0)^2/\theta) \text{ by Assumption 6} \\
\leq& \frac{\theta}{2t_0}\exp\left(-\frac{t_0^2}{\theta}\right) \text{ as } F(Y'_{n+1}) > F(X'_{n+1}) \text{ and } e^{-x} \leq 1/x \text{ for } x > 0 \\
\leq& Lh^2\theta \text{ as } \theta \leq -\tfrac{t_0^2}{\log(2Lh^2 t_0)}.
\end{aligned}$$

Thus, $\mathbb{E}_n[F(X_{n+1})] \leq F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h + 2dLh^2\theta$.

For Claim 2), we start by establishing a bound for $F(Y'_{n+1})$. Let

$$W_n = -\sqrt{h}\xi_n^Y + \sqrt{2\gamma}Z_n \text{ and } \Delta Y_n = Y'_{n+1} - Y_n = -\nabla F(Y_n)h + \sqrt{h}W_n.$$

By Rolle's theorem, there exits y_n on the line segment between Y_n and Y'_{n+1}, such that

$$\begin{aligned}
F(Y'_{n+1}) =& F(Y_n) + \nabla F(y_n)^T \Delta Y_n \\
=& F(Y_n) + \nabla F(Y_n)^T \Delta Y_n + (\nabla F(y_n) - \nabla F(Y_n))^T \Delta Y_n \\
\leq& F(Y_n) + \nabla F(Y_n)^T \Delta Y_n + L\|\Delta Y_n\|^2 \\
& \text{by Cauchy-Schwarz inequality and Assumption 1} \\
\leq& F(Y_n) - \nabla F(Y_n)^T \nabla F(Y_n) h + \sqrt{h} \nabla F(Y_n)^T W_n \\
& + L\nabla F(Y_n)^T \nabla F(Y_n) h^2 - 2\sqrt{h} h L \nabla F(Y_n)^T W_n + LhW_n^T W_n \\
\leq& F(Y_n) - \frac{1}{2}\|\nabla F(Y_n)\|^2 h + \beta\sqrt{h}\nabla F(Y_n)^T W_n + Lh\|W_n\|^2 \text{ as } Lh < \tfrac{1}{2}, \qquad (15)
\end{aligned}$$

where $\beta = 1 - 2hL \in (0,1)$. Taking conditional expectation yields the first estimate. Next, note that for any $0 < b < \min\{1/(8\gamma), 1/(4h\theta)\}$, we have

$$\begin{aligned}
& \mathbb{E}\left[\exp(a^T W_n + b\|W_n\|^2)\right] \\
\leq& \mathbb{E}\left[\exp\left(-\sqrt{h}a^T \xi_n^Y + \sqrt{2\gamma}a^T Z_n + 4\gamma b\|Z_n\|^2 + 2bh\|\xi_n^Y\|^2\right)\right] \\
& \text{by Young's inequality } \|W_n\|^2 \leq 4\gamma\|Z_n\|^2 + 2h\|\xi_n^Y\|^2 \\
\leq& (1-8\gamma b)^{-d/2} \exp\left(\frac{\gamma\|a\|^2}{1-8\gamma b}\right)(1-4bh\theta)^{-d/2}\exp\left(\frac{h\|a\|^2\theta}{2(1-4bh\theta)}\right) \qquad (16) \\
& \text{by Assumption 6 and the fact that } Z_n \sim \mathcal{N}(0, I).
\end{aligned}$$

Then,

$$\begin{aligned}
& \mathbb{E}_n[\hat{V}(Y'_{n+1})] \\
\leq& \hat{V}(Y_n)\exp\left(-\frac{\hat{\eta}h}{2}\|\nabla F(Y_n)\|^2\right)\mathbb{E}_n\left[\exp\left(\hat{\eta}\beta\sqrt{h}\nabla F(Y_n)^T W_n + \hat{\eta}Lh\|W_n\|^2\right)\right] \text{ by (15)} \\
=& \hat{V}(Y_n)\exp\left(-\frac{\hat{\eta}h}{2}\|\nabla F(Y_n)\|^2\right)(1-8\gamma\hat{\eta}Lh)^{-d/2}\exp\left(\frac{\gamma\hat{\eta}^2\beta^2 h\|\nabla F(Y_n)\|^2}{1-8\gamma\hat{\eta}Lh}\right) \\
& (1-4\hat{\eta}Lh^2\theta)^{-d/2}\exp\left(\frac{h\theta\hat{\eta}^2\beta^2 h\|\nabla F(Y_n)\|^2}{2(1-4\hat{\eta}Lh^2\theta)}\right) \\
& \text{by (16) as } \gamma\hat{\eta} < 1/16,\ \hat{\eta}h\theta < 1/8 \text{ and } Lh < 1/2 \\
\leq& \hat{V}(Y_n)\exp\left(-\frac{\hat{\eta}h}{2}\|\nabla F(Y_n)\|^2\right)\exp\left(\frac{\hat{\eta}h}{4}\|\nabla F(Y_n)\|^2 + 16\hat{\eta}\gamma Lh\frac{d}{2} + 8\hat{\eta}Lh^2\theta\frac{d}{2}\right) \\
& \text{as } 8\gamma\hat{\eta}Lh < 1/4,\ 4\hat{\eta}Lh^2\theta < 1/4 \text{ and } \beta < 1 \\
\leq& \hat{V}(Y_n)\exp\left(-\frac{\hat{\eta}h}{4}\|\nabla F(Y_n)\|^2 + \hat{\eta}h(8\gamma Ld + 4Lh\theta d)\right) \\
\leq& \hat{V}(Y_n)\exp\left(-\frac{\hat{\eta}h}{4}\lambda_c F(Y_n) + \frac{\hat{\eta}h}{4}M_c + \hat{\eta}h(8\gamma Ld + 4Lh\theta d)\right) \text{ by Assumption 2.}
\end{aligned}$$

The upper bound for $\mathbb{E}_n[\hat{V}(X'_{n+1})]$ can be obtained in a similar way.

Lastly, for Claim 3), we first note following the same argument as (15), we have

$$F(X'_{n+1}) \leq F(X_n) - \frac{1}{2}\|\nabla F(X_n)\|^2 h - \beta h \nabla F(X_n)^T \xi_n^X + Lh^2\|\xi_n^X\|^2.$$

Then,

$$\begin{aligned}\mathbb{E}_n[\hat{V}(X'_{n+1})] \leq & \hat{V}(X_n)\exp\left(-\frac{1}{2}\|\nabla F(X_n)\|^2\hat{\eta}h\right)\\ & (1-2\hat{\eta}Lh^2\theta)^{-d/2}\exp\left(\frac{\hat{\eta}^2\beta^2h^2\theta\|\nabla F(X_n)\|^2}{2(1-2\hat{\eta}Lh^2\theta)}\right)\\ \leq & \hat{V}(X_n)\exp\left(4\hat{\eta}Lh^2\theta\tfrac{d}{2}\right) \text{ as } \hat{\eta}h\theta < 1/8 \text{ and } hL < 1/2\\ \leq & \hat{V}(X_n)\exp(\hat{\eta}h\theta d) \text{ as } hL < 1/2.\end{aligned}$$

Next, note that when $\|X'_{n+1}\| > \hat{M}_V$ or $\|Y'_{n+1}\| > \hat{M}_V$, $\hat{V}(X_{n+1}) \leq \hat{V}(X'_{n+1})$. When $\max\{\|X'_{n+1}\|, \|Y'_{n+1}\|\} \leq \hat{M}_V$, if $F(Y'_{n+1}) \leq F(X'_{n+1})$, $\hat{V}(X_{n+1}) \leq \hat{V}(X'_{n+1})$. If $F(Y'_{n+1}) > F(X'_{n+1})$, we have

$$\begin{aligned}&\tilde{\mathbb{E}}_n[\hat{V}(X_{n+1})]\\ =&\hat{V}(X'_{n+1})\mathbb{P}\left(\hat{F}_n(X'_{n+1}) \leq \hat{F}_n(Y'_{n+1}) + t_0\right) + \hat{V}(Y'_{n+1})\mathbb{P}\left(\hat{F}_n(X'_{n+1}) > \hat{F}_n(Y'_{n+1}) + t_0\right)\\ \leq&\hat{V}(X'_{n+1})\left(1+\exp(\hat{\eta}(F(Y'_{n+1}) - F(X'_{n+1})))\exp\left(-\frac{1}{\theta}\left(F(Y'_{n+1}) - F(X'_{n+1}) + t_0\right)^2\right)\right)\\ &\text{by Assumption 6}\\ \leq&\hat{V}(X'_{n+1})\left(1+\exp\left(-\tfrac{t_0^2}{\theta}\right)\right) \text{ as } \hat{\eta} < 2t_0/\theta\\ \leq&\hat{V}(X'_{n+1})\exp(\hat{\eta}h\theta d) \text{ as } 1+\exp(-t_0^2/\theta) < \exp(\hat{\eta}h\theta d).\end{aligned}$$

Thus,

$$\mathbb{E}_n[\hat{V}(X_{n+1})] \leq \mathbb{E}_n[\hat{V}(X'_{n+1})\exp(\hat{\eta}h\theta d)] \leq \hat{V}(X_n)\exp(2\hat{\eta}h\theta d).$$

■

Recall that $\hat{R}_V = 8\lambda_c^{-1}\hat{C}_V$. We define a sequence of stopping times:

$$\hat{\tau}_0 = \min\left\{n \geq 0 : F(Y_n) \leq \hat{R}_V\right\},$$

and for $k = 1, 2, \ldots,$

$$\hat{\tau}_k = \min\left\{n > \hat{\tau}_{k-1} : F(Y_n) \leq \hat{R}_V\right\}.$$

Utilizing the Lyapunov function $\hat{V}$, our second result establishes bounds for the moment generating function of the stopping times.

Lemma 19 *For (n)SGDxSGLD, under Assumptions 1, 2 and 6, if $Lh \leq 1/2$ and $\hat{\eta} \leq \min\{(16\gamma)^{-1}, (8h\theta)^{-1}\}$, for any $K \geq 0$, the stopping time $\hat{\tau}_K$ satisfies*

$$\mathbb{E}[\exp(h\hat{\eta}\hat{C}_V\hat{\tau}_K)] \leq \exp(2Kh\hat{\eta}\hat{C}_V + K\hat{\eta}\hat{R}_V)\hat{V}(Y_0).$$

The proof of Lemma 19 follows exactly the same lines of arguments as the proof of Lemma 8. We thus omit it here.

Let

$$\hat{D} = \max\{\|x - h\nabla F(x)\| : F(x) \leq \hat{R}_V\}.$$

Following the similar lines of arguments as Lemma 9, we have the following result.

Lemma 20 *For (n)SGDxSGLD, under Assumption 6, if $F(Y_n) \leq \hat{R}_V$, then, for any $r > 0$, there exists $\hat{\alpha}(r, \hat{D}) > 0$, such that*

$$\mathbb{P}_n(\|Y'_{n+1}\| \leq r) > \hat{\alpha}(r, \hat{D}).$$

In particular,

$$\hat{\alpha}(r, \hat{D}) \geq \frac{S_d r^d}{(8\tau h\pi)^{\frac{d}{2}}} \left(1 - \frac{4h^2\theta}{r^2}\right) \exp\left(-\frac{1}{2\tau h}(\hat{D}^2 + r^2)\right) > 0.$$

Proof

$$\begin{aligned}\mathbb{P}_n(\|Y'_{n+1}\| \leq r) &= \mathbb{P}_n(\|Y_n - h\nabla F(Y_n) - h\xi_n^Y + \sqrt{2\gamma h}Z_n\| \leq r) \\ &\geq \mathbb{P}_n\left(\|Z_n - Q_n\| \leq \frac{r}{2\sqrt{2\tau h}}\right) \mathbb{P}_n\left(h\|\xi_n^Y\| \leq \frac{r}{2}\right),\end{aligned}$$

where $Q_n = -(Y_n - h\nabla F(Y_n))/\sqrt{2\tau h}$. Note that as $F(Y_n) \leq \hat{R}_V$, $\|Q_n\| \leq \hat{D}/\sqrt{2\tau h}$. Thus,

$$\begin{aligned}\mathbb{P}_n\left(\|Z_n - Q_n\| \leq \frac{r}{\sqrt{2\tau h}}\right) &= \int_{\|z\| \leq \frac{r}{2\sqrt{2\tau h}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\|z + Q_n\|^2\right) dz \\ &\geq \int_{\|z\| \leq \frac{r}{2\sqrt{2\tau h}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\tau h}(\hat{D}^2 + r^2)\right) dz \\ &\geq \frac{S_d r^d}{(8\tau h\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\tau h}(\hat{D}^2 + r^2)\right).\end{aligned}$$

Lastly, by Markov inequality,

$$\mathbb{P}_n\left(h\|\xi_n^Y\| \leq \frac{r}{2}\right) \geq 1 - \frac{\mathbb{E}[4h^2\|\xi_n^Y\|^2]}{r^2} \geq 1 - \frac{4h^2\theta d}{r^2}.$$

■

Our next result shows that if $F(Y'_{n+1}) \leq \frac{1}{4}r_0$, there is a positive probability that $F(X_{n+1}) \leq \frac{1}{2}r_0$.

Lemma 21 *For (n)SGDxSGLD, under Assumption 6, if $t_0 \leq \frac{1}{8}r_0$, $\theta \leq \frac{r_0^2}{64\log 2}$, $F(Y'_{n+1}) \leq \frac{1}{4}r_0$, and $\|X'_{n+1}\| \leq \hat{M}_V$, then*

$$\tilde{\mathbb{P}}_n\left(F(X_{n+1}) \leq \frac{1}{2}r_0\right) \geq \frac{1}{2}.$$

Proof Note that if $F(X'_{n+1}) \leq \frac{1}{2}r_0$, $F(X_{n+1})$ is guaranteed to be less than $\frac{1}{2}r_0$. If $F(X'_{n+1}) > \frac{1}{2}r_0$, the probability of exchange is

$$\begin{aligned}
&\tilde{\mathbb{P}}_n\left(\hat{F}_n(X'_{n+1}) \geq \hat{F}_n(Y'_{n+1}) + t_0\right)\\
=&\tilde{\mathbb{P}}_n\left(F(X'_{n+1}) - F(Y'_{n+1}) + \xi_n^X - \xi_n^Y \geq t_0\right)\\
\geq&\tilde{\mathbb{P}}_n\left(\xi_n^X - \xi_n^Y \geq \tfrac{1}{8}r_0\right) \text{ as } F(X'_{n+1}) - F(Y'_{n+1}) > \tfrac{1}{4}r_0 \text{ and } t_0 \leq \tfrac{1}{8}r_0\\
\geq&1 - \exp\left(-\frac{r_0^2}{64\theta}\right) \text{ under Assumption 6}\\
\geq&\frac{1}{2} \text{ as } \theta \leq \tfrac{r_0^2}{64\log 2}.
\end{aligned}$$

If exchange takes place, $X_{n+1} = Y'_{n+1}$ and $F(X_{n+1}) \leq \frac{1}{4}r_0$. ■

B.2 Convergence to global minimum

In this subsection, we analyze the "speed" of convergence $\{X_{n+k} : k \geq 0\}$ to x^* when $F(X_n) \leq \frac{1}{2}r_0$. Let

$$\kappa_n = \inf\{k > 0 : F(X_{n+k}) > r_0\}.$$

Lemma 22 *For (n)SGDxSGLD, under Assumption 1 and 6, and assuming $Lh \leq 1/2$ and $\hat{\eta} < (8h\theta)^{-1}$, if $F(X_n) \leq \frac{1}{2}r_0$, then for any fixed $k > 0$,*

$$\mathbb{P}_n(\kappa_n > k) \geq 1 - \exp\left(-\frac{r_0}{16h\theta} + \frac{1}{4}dk\right).$$

Proof From Lemma 18, the following is a supermartingale

$$\hat{V}(X_{n+k})\exp\left(-\frac{1}{4}dk\right)1_{\kappa_n \geq k}.$$

In particular,

$$\mathbb{E}_{n+k}\left[\hat{V}(X_{n+k+1})]\exp\left(-\frac{1}{4}d(k+1)\right)1_{\kappa_n \geq k+1}\right] \leq \hat{V}(X_{n+k})\exp\left(-\frac{1}{4}dk\right)1_{\kappa_n \geq k}.$$

Therefore,

$$\mathbb{E}_n[\hat{V}(X_{n+(\kappa_n \wedge k)})\exp(-\tfrac{1}{4}d(\kappa_n \wedge k))] \leq \hat{V}(X_n) \leq \exp\left(\frac{1}{2}\hat{\eta}r_0\right).$$

We also note

$$\mathbb{E}_n\left[\hat{V}(X_{n+(\kappa_n \wedge k)})]\exp\left(-\frac{1}{4}d(\kappa_n \wedge k)\right)\right] \geq \mathbb{E}_n\left[\exp(\hat{\eta}r_0)\exp\left(-\frac{1}{4}dk\right)1_{\kappa_n \leq k}\right].$$

Then,

$$\mathbb{P}_n(\kappa_n \le k) \le \exp\left(-\frac{1}{2}\hat{\eta} r_0 + \frac{1}{4}dk\right) \le \exp\left(-\frac{r_0}{16h\theta} + \frac{1}{4}dk\right),$$

since $8\hat{\eta}h\theta < 1$. ■

Lemma 23 *For (n)SGDxSGLD, under Assumptions 1 and 6, and assuming F is strongly convex in B_0 and $h \le \min\{1/(2L), 1/m\}$, if $F(X_n) \le \frac{1}{2}r_0$,*

$$\mathbb{E}_n[F(X_{n+k})1_{\kappa_n > k}] \le (1-mh)^k F(X_n) + \frac{d\theta}{m},$$

for all $k \ge 0$.

Proof We first note from Lemma 18, if $F(X_n) \le r_0$,

$$\begin{aligned}\mathbb{E}_{n+j}[F(X_{n+j+1})] &\le F(X_{n+j}) - \frac{1}{2}\|\nabla F(X_{n+j})\|^2 h + 2dLh^2\theta \\ &\le (1-mh)F(X_{n+j}) + 2dLh^2\theta,\end{aligned}$$

where the second inequality follows from (12) as $F(x)$ is strongly convex in B_0.

Next, we note

$$\begin{aligned}\mathbb{E}_n[F(X_{n+(\kappa_n \wedge k)})] &\le \mathbb{E}_n\left[(1-mh)^{(\kappa_n \wedge k)-1}F(X_n) + \sum_{j=1}^{\kappa_n \wedge k}(1-mh)^{(\kappa_n \wedge k)-j}dLh^2\theta\right] \\ &\le F(X_n) + \frac{2dLh\theta}{m} \le F(X_n) + \frac{d\theta}{m} \text{ as } Lh < 1/2.\end{aligned}$$

Because $\mathbb{E}_n[F(X_{n+(\kappa_n \wedge k)})] > \mathbb{E}_n[F(X_{n+k})1_{\kappa_n > k}]$,

$$\mathbb{E}_n[F(X_{n+k})1_{\kappa_n > k}] \le (1-mh)^k F(X_n) + \frac{d\theta}{m}.$$

■

B.3 Proof of Theorem 4

For any fixed accuracy $\epsilon > 0$ and δ, we set

$$K(\delta) = \frac{\log(\delta/3)}{\log(1-\hat{\alpha}(r_0/4, \hat{D}))} = O(-\log(\delta)),$$

$$k(\epsilon, \delta) = \frac{\log(2\delta\epsilon/(9r_0))}{\log(1-mh)} = O(-\log(\delta) - \log(\epsilon)),$$

and

$$N(\epsilon, \delta) = k(\epsilon, \delta) + \frac{2K(\delta)h\hat{\eta}vC_V + K(\delta)\hat{\eta}\hat{R}_V + \log \hat{V}(Y_0) - \log(\delta/3)}{h\hat{\eta}\hat{C}_V} = O(-\log(\delta) - \log(\epsilon)).$$

For any fixed $N > N(\epsilon, \delta)$, we set

$$\begin{aligned}
&\theta(N, \epsilon, \delta) \\
\leq &\min\left\{\frac{\delta\epsilon m}{9d}, \frac{r_0}{16h(dN/4 - \log(\delta/9))}, -\frac{t_0^2}{\log(2Lh^2 t_0)}, -\frac{t_0^2}{\log(\exp(d/8) - 1)}, \frac{r_0^2}{64\log 2}\right\} \\
= &O(\min\{N^{-1}, \epsilon\delta\}).
\end{aligned}$$

Now for fixed $N > N(\epsilon, \delta)$ and $\theta \leq \theta(N, \epsilon, \delta)$, we first note if $F(X_n) \leq \frac{1}{2}r_0$ for $n \leq N - k(\epsilon, \delta)$,

$$\begin{aligned}
\mathbb{P}_n(F(X_N) > \epsilon) =&\mathbb{P}_n(F(X_N) > \epsilon, \kappa_n > N - n) + \mathbb{P}_n(F(X_{n+k}) > \epsilon, \kappa_n \leq N - n) \\
\leq&\mathbb{P}_n(F(X_N)1_{\kappa_n > N-n} > \epsilon) + \mathbb{P}_n(\kappa_n \leq N - n) \\
\leq&\frac{1}{\epsilon}\left((1 - mh)^{N-n}F(X_n) + \frac{d\theta}{m}\right) + \exp\left(-\frac{r_0}{16h\theta} + \frac{1}{4}dN\right) \\
&\text{by Markov inequality, Lemma 23, and Lemma 22} \\
\leq&\frac{1}{\epsilon}\left((1 - mh)^{k(\epsilon,\delta)}\frac{r_0}{2} + \frac{d\theta}{m}\right) + \exp\left(-\frac{r_0}{16h\theta} + \frac{1}{4}dN\right) \\
\leq&\frac{1}{3}\delta \text{ by our choice of } \theta(N, \epsilon, \delta) \text{ and } k(\epsilon, \delta).
\end{aligned}$$

Next, we study how long it takes X_n to visit the set $\{x : F(x) \leq r_0/2\}$. In particular, we denote $T = \inf\{n : F(X_n) \leq r_0/2\}$. From Lemma 20 and 21, every time $Y_n \in \{x : F(x) \leq \hat{R}_V\}$,

$$\mathbb{P}_n\left(F(X_{n+1}) \leq \frac{1}{2}r_0\right) \geq \frac{1}{2}\mathbb{P}_n\left(\|Y'_{n+1}\| \leq \frac{1}{4}r_0\right) \geq \frac{1}{2}\hat{\alpha}(r_0/4, \hat{D}) > 0.$$

Then,

$$\begin{aligned}
\mathbb{P}(F(X_{\hat{\tau}_k+1}) \text{ for } k = 1, \ldots, K(\delta)) &= \mathbb{E}\left[\prod_{k=1}^{K(\delta)} \mathbb{P}_{\hat{\tau}_k}\left(F(X_{\hat{\tau}_k+1}) > \frac{1}{2}r_0\right)\right] \\
&\leq \left(1 - \frac{1}{2}\hat{\alpha}(r_0/4, \hat{D})\right)^{K(\delta)} < \frac{\delta}{3}.
\end{aligned}$$

From Lemma 19, by Markov inequality,

$$\begin{aligned}
&\mathbb{P}(\hat{\tau}_{K(\delta)} > N - k(\epsilon, \delta)) \\
&\leq \frac{\mathbb{E}[\exp(h\hat{\eta}\hat{C}_V\hat{\tau}_{K(\delta)})]}{\exp(h\hat{\eta}\hat{C}_V(N - k(\epsilon, \delta)))} \\
&\leq \frac{\exp(2K(\delta)h\hat{\eta}vC_V + K(\delta)\hat{\eta}\hat{R}_V)V(Y_0)}{\exp(h\hat{\eta}\hat{C}_V(N - k(\epsilon, \delta)))} \\
&\leq \frac{\delta}{3} \text{ since } N - k(\epsilon, \delta) > \frac{2K(\delta)h\hat{\eta}vC_V + K(\delta)\hat{\eta}\hat{R}_V + \log V(Y_0) - \log\delta + \log 3}{h\hat{\eta}\hat{C}_V}.
\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(T \le N - k(\epsilon,\delta)) &\ge \mathbb{P}\left(\hat{\tau}_{K(\delta)} \le N - k(\epsilon,\delta) \text{ and } F(\hat{X}_{\hat{\tau}_k+1}) \text{ for some } k = 1,\dots,K(\delta)\right) \\ &\ge \left(1 - \frac{\delta}{3}\right) + \left(1 - \frac{\delta}{3}\right) - 1 = 1 - \frac{2\delta}{3}.\end{aligned}$$

Lastly,

$$\begin{aligned}\mathbb{P}(F(X_N) \le \epsilon) &\ge \mathbb{P}(F(X_N) \le \epsilon, T \le N - k(\epsilon,\delta)) \\ &\ge \mathbb{E}[\mathbb{P}_T(F(X_N) \le \epsilon) | T \le N - k(\epsilon,\delta)]\mathbb{P}(T \le N - k(\epsilon,\delta)) \\ &\ge \left(1 - \frac{2\delta}{3}\right)\left(1 - \frac{\delta}{3}\right) \ge 1 - \delta.\end{aligned}$$

B.4 Proof of Theorem 5

Let

$$\hat{Y}_{n+1} = \hat{Y}_n - \nabla F(\hat{Y}_n)h - \xi_n^Y h + \sqrt{2\gamma h} Z_n.$$

This is to be differentiated with Y_n in Algorithm 2, which can swap position with X_n. We denote $\hat{\mu}_n$ as the distribution of $\hat{Y}_n$.

Lemma 24 *Under Assumptions 1, 5, and 6, for* $h < \frac{\beta}{4\sqrt{2}L^2}$,

$$KL(\hat{\mu}_n | \pi_\gamma) \le \exp(-\tfrac{1}{2}\beta\gamma h(n-1))\left(\frac{1}{\gamma}F(\hat{Y}_0) + \hat{M}_d\right) + \frac{6hL^2 d + \theta d/\gamma}{\beta\gamma},$$

where $\hat{M}_d = 2Lhd - \frac{d}{2}(\log(4\pi h) + 1) + \log(U_\gamma)$.

Proof For a given realization of ξ_0^Y and $\hat{Y}_0$,

$$\hat{Y}_1 \sim \mathcal{N}(m(\hat{Y}_0), 2\gamma h I_d), \text{ where } m(\hat{Y}_0) = \hat{Y}_0 - \nabla F(\hat{Y}_0)h - \xi_0^Y h.$$

Then, given $\tilde{Y}_0$,

$$\begin{aligned}\mathrm{KL}(\hat{\mu}_1 \| \pi_\gamma) &= \int \log\left(\frac{\hat{\mu}_1(y)}{\pi_\gamma(y)}\right)\hat{\mu}_1(y)dy \\ &= \int\left(-\frac{d}{2}\log(4\pi\gamma h) - \frac{\|y - m(\hat{Y}_0)\|^2}{4\gamma h} + \log U_\gamma + \frac{F(y)}{\gamma}\right)\hat{\mu}_1(y)dy \\ &= -\frac{d}{2}(\log(4\pi\gamma h) + 1) + \log U_\gamma + \frac{1}{\gamma}\mathbb{E}[F(\hat{Y}_1)|\hat{Y}_0] \qquad (17) \\ &\le -\frac{d}{2}(\log(4\pi\gamma h) + 1) + \log U_\gamma + \frac{1}{\gamma}(F(\hat{Y}_0) + 2Lh\|\xi_0^Y\|^2 + 2\gamma Lhd) \\ &= \frac{1}{\gamma}F(\hat{Y}_0) + \hat{M}_d + \frac{2Lh}{\gamma}\|\xi_0^Y\|^2,\end{aligned}$$

where the last inequality follows from (15).

Next, following the proof of Lemma 3 in (Vempala and Wibisono, 2019), for a given ξ_1^Y, with slight abuse of notation, consider the modified diffusion process:

$$dY_t = -\nabla F(\hat{Y}_1)dt - \xi_1^Y dt + \sqrt{2\gamma}dB_t, \quad Y_1 = \hat{Y}_t$$

for $t \geq 1$. Note that $\hat{Y}_2$ follows the same law as Y_{1+h}. Let μ_t denote the distribution of Y_t. Then

$$\begin{aligned}
&\frac{d}{dt}\mathrm{KL}(\mu_t\|\pi_\gamma) \\
=& -\gamma\int\left\|\nabla\log\frac{\mu_t(x)}{\pi_\gamma(x)}\right\|^2\mu_t(x)dx + \mathbb{E}\left[\left\langle\nabla F(Y_t) - \xi_1^Y - \nabla F(\hat{Y}_1), \nabla\log\frac{\mu_t(Y_t)}{\pi_\gamma(Y_t)}\right\rangle\right] \\
\leq& -\frac{3\gamma}{4}\int\left\|\nabla\log\frac{\mu_t(x)}{\pi_\gamma(x)}\right\|^2\mu_t(x)dx + \frac{1}{\gamma}\mathbb{E}\left[\left\|\nabla F(Y_t) - \xi_1^Y - \nabla F(\hat{Y}_1)\right\|^2\right] \\
\leq& -\frac{3\gamma}{4}\int\left\|\nabla\log\frac{\mu_t(x)}{\pi_\gamma(x)}\right\|^2\mu_t(x)dx + \frac{2}{\gamma}L^2\mathbb{E}[\|Y_t - \hat{Y}_1\|^2] + \frac{2}{\gamma}\|\xi_1^Y\|^2.
\end{aligned}$$

Because $Y_t \overset{d}{=} \hat{Y}_1 - (\nabla F(\hat{Y}_1) + \xi_1^Y)(t-1) + \sqrt{2\gamma(t-1)}Z_1$, where Z_1 is a standard d-dimensional Gaussian random vector,

$$\begin{aligned}
\mathbb{E}[\|Y_t - \hat{Y}_1\|^2] &= 2(t-1)^2\mathbb{E}[\|\nabla F(\hat{Y}_1)\|^2] + 2(t-1)^2\|\xi_1^Y\|^2 + 2\gamma(t-1)d \\
&\leq \frac{8(t-1)^2L^2}{\beta}\mathrm{KL}(\hat{\mu}_1\|\pi_\gamma) + 4(t-1)^2dL + 2(t-1)^2\|\xi_1^Y\|^2 + 2\gamma(t-1)d.
\end{aligned}$$

In addition, because,

$$\int\left\|\nabla\log\frac{\mu_t(x)}{\pi_\gamma(x)}\right\|^2\mu_t(x)dx \geq 2\beta\mathrm{KL}(\mu_t(x)\|\pi_\gamma),$$

we have

$$\begin{aligned}
&\frac{d}{dt}\mathrm{KL}(\mu_t\|\pi_\gamma) \\
\leq& -\frac{3\beta\gamma}{2}\mathrm{KL}(\mu_t\|\pi_\gamma) \\
&+ 2\frac{L^2}{\gamma}\left(\frac{8(t-1)^2L^2}{\beta}\mathrm{KL}(\hat{\mu}_1\|\pi_\gamma) + 4(t-1)^2dL + 2(t-1)^2\|\xi_1^Y\|^2 + 2\gamma(t-1)d\right) + \frac{2}{\gamma}\|\xi_1^Y\|^2 \\
\leq& -\frac{3\beta\gamma}{2}\mathrm{KL}(\mu_t\|\pi_\gamma) + \frac{16(t-1)^2L^4}{\beta\gamma}\mathrm{KL}(\mu_1\|\pi_\gamma) + \frac{1}{\gamma}(4(t-1)^2L^2 + 2)\|\xi_1^Y\|^2 \\
&+ 8(t-1)^2d\frac{L^3}{\gamma} + 4(t-1)dL^2,
\end{aligned}$$

which further implies that

$$\mathrm{KL}(\mu_{1+h}\|\pi_\gamma) \leq e^{-\frac{3}{2}\beta\gamma h}\mathrm{KL}(\mu_1\|\pi_\gamma) + \frac{16h^3L^4}{\beta\gamma}\mathrm{KL}(\mu_1\|\pi_\gamma) + \frac{4h^3L^2 + 2h}{\gamma}\|\xi_1^Y\|^2 + \frac{8h^3dL^3}{\gamma} + 4h^2dL^2$$

For $h < \frac{\beta\gamma}{4\sqrt{2}L^2}$, we have

$$\mathrm{KL}(\mu_{1+h}\|\pi_\gamma) \le (1-\beta h\gamma)\,\mathrm{KL}(\mu_1\|\pi_\gamma) + \frac{3h}{\gamma}\|\xi_1^Y\|^2 + 6h^2L^2d.$$

The above analysis implies that given ξ_k^Y's,

$$\begin{aligned}\mathrm{KL}(\hat{\mu}_n\|\pi_\gamma) &\le (1-\beta h\gamma)\,\mathrm{KL}(\hat{\mu}_{n-1}\|\pi_\gamma) + \frac{3h}{\gamma}\|\xi_1^Y\|^2 + 6h^2L^2d \\ &\le \exp(-\tfrac{1}{2}\beta h(n-1)\gamma)\mathrm{KL}(\hat{\mu}_1\|\pi_\gamma) + \frac{6hL^2d}{\beta\gamma} + \frac{3h}{\gamma}\sum_{k=1}^{n-1}(1-h\beta\gamma)^{n-k}\|\xi_k^Y\|^2.\end{aligned}$$

Plug in the bound for $\mathrm{KL}(\hat{\mu}_1\|\pi_\gamma)$ in (17) and taking the expectation with respect to ξ_k^Y's, we have

$$\begin{aligned}\mathrm{KL}(\hat{\mu}_n\|\pi_\gamma) &\le \exp(-\tfrac{1}{2}\beta h(n-1)\gamma)\left(\frac{1}{\gamma}F(\hat{Y}_0) + \hat{M}_d\right) + \frac{6hL^2d}{\beta\gamma} + \frac{1}{\gamma}h\sum_{k=0}^{n-1}(1-h\beta\gamma)^{n-k}\theta d \\ &\le \exp(-\tfrac{1}{2}\beta h(n-1)\gamma)\left(\frac{1}{\gamma}F(\hat{Y}_0) + \hat{M}_d\right) + \frac{6hL^2d + \theta d/\gamma}{\beta\gamma}.\end{aligned}$$

■

Based on Lemma 24, let

$$n_0 = \frac{4}{\beta\gamma}h^{-1}\log(1/h) + 1.$$

For $n \ge n_0$ and h small enough,

$$\begin{aligned}\mathrm{KL}(\hat{\mu}_n\|\pi_\gamma) &\le h^2\left(\frac{1}{\gamma}F(\hat{Y}_0) + \hat{M}_d\right) + \frac{6hL^2d + \theta d/\gamma}{\beta\gamma} \\ &\le hF(\hat{Y}_0) + \frac{8hL^2d + \theta d/\gamma}{\beta\gamma}\end{aligned} \tag{18}$$

Recall that

$$\hat{B}_0 = \{x : F(x) \le r_0/4\}.$$

We next draw connection between the bound (18) and the hitting time of Y_n to $\hat{B}_0$. For nSGDxSGLD, let $\hat{\phi} = \min_n\{Y_n \in \hat{B}_0\}$. With a slight abuse of notation, for SGDxSGLD, let $\hat{\phi} = \min_n\{X_n = Y_n' \text{ or } Y_n \in \hat{B}_0\}$.

Lemma 25 *For (n)SGDxSGLD, under Assumptions 1, 2, 5, and 6,*

$$\mathbb{P}(\hat{\phi} \le n_0) \ge \pi_\gamma(\hat{B}_0) - \sqrt{h}F(\hat{Y}_0) - 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} - \frac{\sqrt{\theta d}}{\sqrt{\beta\gamma}}.$$

Proof For both nSGDxSGLD and SGDxSGLD,

$$\mathbb{P}(\hat{\phi} \le n) \ge \mathbb{P}(\hat{Y}_n \in \hat{B}_0).$$

By Pinsker's inequality,

$$d_{tv}(\hat{\mu}_n, \pi_\gamma) \le \sqrt{\frac{1}{2}\mathrm{KL}(\hat{\mu}_n \| \pi_\gamma)} \le \sqrt{h}F(\hat{Y}_0) + 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} + \frac{\sqrt{\theta d}}{\sqrt{\beta}\gamma}$$

where the last inequality follows from (18). Then

$$\mathbb{P}(\hat{Y}_{n_0} \in \hat{B}_0) \le \pi_\gamma(\hat{B}_0) - \sqrt{h}F(\hat{Y}_0) - 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} - \frac{\sqrt{\theta d}}{\sqrt{\beta}\gamma}$$

■

Lemma 26 *For (n)GDxLD, fix any K and $T \ge (n_0+1)K$,*

$$\begin{aligned}\mathbb{P}_0(\hat{\phi} > T) \le & \exp(2(n_0+1)Kh\hat{\eta}\hat{C}_V + (n_0+1)K\hat{\eta}\hat{R}_V - h\hat{\eta}\hat{C}_V T)\hat{V}(Y_0) \\ & + \left(1 - \pi_\gamma(\hat{B}_0) + \sqrt{h}\hat{R}_V + 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} + \frac{\sqrt{\theta d}}{\sqrt{\beta}\gamma}\right)^K.\end{aligned}$$

Proof Recall the sequence of stoping times $\hat{\tau}_j = \inf\{n > \tau_{j-1} : F(Y_n) \le \hat{R}_V\}$. Applying Lemma 19, we have

$$\mathbb{E}[\exp(h\hat{\eta}\hat{C}_V\hat{\tau}_k)] \le \exp(2kh\hat{\eta}\hat{C}_V + k\hat{\eta}\hat{R}_V)\hat{V}(Y_0).$$

We next define a new sequence of stopping times:

$$\hat{\psi}_0 = \inf\left\{n : F(Y_n) \le \hat{R}_V\right\}, \quad \hat{\psi}_0' = \hat{\psi}_0 + \hat{n}_0,$$

and for $k = 1, \ldots,$

$$\hat{\psi}_k = \inf\left\{n \ge \hat{\psi}_{k-1}' + 1, F(Y_n) \le \hat{R}_V\right\}, \quad \hat{\psi}_k' = \hat{\psi}_k + \hat{n}_0 + 1.$$

Note that $\hat{\psi}_i$ always coincide with one of $\hat{\tau}_j$'s, and as $\hat{\tau}_j - \hat{\tau}_{j-1} \ge 1$,

$$\hat{\psi}_k \le \hat{\tau}_{(n_0+1)k}.$$

Thus,

$$\begin{aligned}\mathbb{P}_0(\hat{\phi} \ge T) \le & \mathbb{P}_0\left(\hat{\tau}_{(n_0+1)K} \ge T\right) + \mathbb{P}_0\left(Y_{\psi_k'} \notin B_0, \forall k \le K\right) \\ & \le \exp(2(n_0+1)Kh\hat{\eta}\hat{C}_V + (n_0+1)K\hat{\eta}\hat{R}_V - h\hat{\eta}\hat{C}_V T)\hat{V}(Y_0) \\ & \quad + \left(1 - \pi_\gamma(\hat{B}_0) + \sqrt{h}\hat{R}_V + 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} + \frac{\sqrt{\theta d}}{\sqrt{\beta}\gamma}\right)^K \text{ by Lemma 25.}\end{aligned}$$

■

We are now ready to prove Theorem 5. For a fixed N, let A_N denote the event

$$\left\{\xi_n^X - \xi_n^Y > \frac{1}{2}t_0 \text{ for some } n \le N\right\}$$

Note that

$$\mathbb{P}(A_N) \le \sum_{n=1}^{N} \mathbb{P}\left(\zeta_n^X - \zeta_n^Y > \frac{1}{2}t_0\right) \le N \exp\left(-\frac{t_0^2}{\theta}\right).$$

Next, set

$$h < \left(\frac{\pi_\gamma(\hat{B}_0)}{4\hat{R}_V + 8\sqrt{d}L/\sqrt{\beta\gamma}}\right)^2 \text{ and } \theta < \frac{\pi_\gamma(\hat{B}_0)^2\beta\gamma^2}{16d}.$$

Then,

$$1 - \pi_\gamma(\hat{B}_0) + \sqrt{h}\hat{R}_V + 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} + \frac{\sqrt{\theta d}}{\sqrt{\beta}\gamma} \le 1 - \pi_\gamma(\hat{B}_0)/2$$

Following the proof of Theorem 4, we also set

$$k(\epsilon, \delta) = \frac{\log(2\delta\epsilon/(9r_0))}{\log(1-mh)}$$

and

$$\theta(N,\delta) = \min\left\{\frac{1}{t_0^2}\log\left(\frac{\delta}{3N}\right), \frac{\delta\epsilon m}{9d}, \frac{r_0}{16h(dN/4 - \log(\delta/9))}, \frac{\pi_\gamma(\hat{B}_0)^2\beta\gamma^2}{16d}\right\}.$$

In what follows we assume $\theta \le \theta(N,\delta)$.

For nSGDxSGLD, define $J = 1$,

$$\hat{\phi}_0 = \min\{n \ge 0 : F(Y_n) \in \hat{B}_0\},$$

and for $k \ge 1$,

$$\hat{\phi}_k = \min\{n > 0 : F(Y_{\phi_k + n}) \in \hat{B}_0\}.$$

For SGDxSGLD, with a slight abuse of notation, define $J = \hat{R}_v/t_0$,

$$\hat{\phi}_0 = \min\{n \ge 0 : F(X_n) \le \hat{R}_V\},$$

and for $k \ge 1$,

$$\hat{\phi}_k = \min\{n : X_{\phi_{k-1}+n} = Y'_{\phi_{k-1}+n} \text{ or } Y_{\phi_{k-1}+n} \in \hat{B}_0\}.$$

We next note that

$$\begin{aligned}\mathbb{P}(F(X_N) > \epsilon) \le& \mathbb{P}(A_N) + \mathbb{P}(F(X_N) > \epsilon, \phi_J < N, A_N^c) \\ &+ \mathbb{P}(\phi_J > N, A_N^c)\end{aligned} \tag{19}$$

We shall look at the three terms on the left hand side of (19) one by one. First, with our choice of θ, $\mathbb{P}(A_N) \le \delta/3$.

Second, suppose there exist $n < N - k(\epsilon, \delta)$ such that $F(X_n) < r_0/2$. Let $\kappa_n = \inf\{k > 0 : F(X_n) > r_0\}$. Then,

$$\begin{aligned}
\mathbb{P}_n(F(X_N) > \epsilon) =& \mathbb{P}_n(F(X_N) > \epsilon, \kappa_n > N - n) + \mathbb{P}_n(F(X_N) > \epsilon, \kappa_n \le N - n) \\
\le& \mathbb{P}_n(F(X_N) 1_{\kappa_n > N-n} > \epsilon) + \mathbb{P}_n(\kappa_n \le N - n) \\
\le& \frac{1}{\epsilon}\left((1 - mh)^{N-n} F(X_n) + \frac{d\theta}{m}\right) + \exp\left(-\frac{r_0}{16h\theta} + \frac{1}{4}dN\right) \\
& \text{by Markov inequality, Lemma 23, and Lemma 22} \\
\le& \frac{1}{\epsilon}\left((1 - mh)^{k(\epsilon,\delta)} \frac{r_0}{2} + \frac{d\theta}{m}\right) + \exp\left(-\frac{r_0}{16h\theta} + \frac{1}{4}dN\right) \\
\le& \frac{1}{3}\delta \quad \text{by our choice of } \theta \text{ and } k(\epsilon, \delta).
\end{aligned}$$

In addition, from Lemma 21,

$$\mathbb{P}\left(F(X_n) \le \frac{1}{2} r_0 \middle| Y_n' \in \hat{B}_0\right) \ge \frac{1}{2}$$

Thus every time $Y_n' \in \hat{B}_0$, there is a chance larger than or equal to $\frac{1}{2}$ for X_n to visit $\{x : F(x) \le r_0/2\}$.

Lastly, we study how long it takes Y_n to visit $\hat{B}_0$. Conditional on A_N^c and $\phi_J < N$, there exists at least one ϕ_k, $k \le J$, such that $Y_{\phi_k} \in \hat{B}_0$. In addition, at all ϕ_k, $1 \le k \le J$, $F(Y_{\phi_k}) \le \hat{R}_V + t_0$. Let $\phi_{-1} \equiv 0$. Then, for $N > (J+1)T$

$$\begin{aligned}
& \mathbb{P}_0(\phi_J \ge (J+1)T | A_N^c) \\
\le& \mathbb{E}_0\left[\sum_{k=0}^{J} \mathbb{P}_{\phi_{k-1}}(\phi_{k+1} > T | A_N^c)\right] \\
\le& (J+1)\exp\left(2(n_0+1)Kh\hat{\eta}\hat{C}_V + (n_0+1)K\hat{\eta}\hat{R}_V - h\hat{\eta}\hat{C}_V T\right)(\hat{V}(Y_0) + \hat{R}_V + t_0) \\
& + (J+1)\left(1 - \pi_\gamma(\hat{B}_0) + \sqrt{h}\hat{R}_V + 2\sqrt{h}\frac{\sqrt{d}L}{\sqrt{\beta\gamma}} + \frac{\sqrt{\theta d}}{\sqrt{\beta}\gamma}\right)^K \text{ by Lemma 26} \\
\le& (J+1)\exp\left(2(n_0+1)Kh\hat{\eta}\hat{C}_V + (n_0+1)K\hat{\eta}\hat{R}_V - h\hat{\eta}\hat{C}_V T\right)(\hat{V}(Y_0) + \hat{R}_V + t_0) \\
& + (J+1)\left(1 - \pi_\gamma(\hat{B}_0)/2\right)^K \text{ by our choice of } h \text{ and } \theta.
\end{aligned}$$

Set

$$K(\delta) = \frac{\log(\delta) - \log(6(J+1))}{\log(1 - \pi_\gamma(\hat{B}_0)/2)}$$

and

$$\begin{aligned}
T(\beta, \delta) =& \frac{2(n_0+1)K(\delta)h\hat{\eta}\hat{C}_V + (n_0+1)K(\delta)\hat{\eta}\hat{R}_V}{h\hat{\eta}\hat{C}_V} \\
& + \frac{\log\left(\hat{V}(Y_0) + \hat{R}_V + t_0\right) - \log(\delta) + \log(6(J+1))}{h\hat{\eta}\hat{C}_V}.
\end{aligned}$$

We have for $K > K(\delta)$ and $T > T(\delta, \beta)$, $\mathbb{P}(\phi_J \geq T, A_N^{\mathrm{c}}) < \delta/3$.

Above all, if we set

$$N(\beta, \epsilon, \delta) = T(\beta, \delta) + k(\delta, \epsilon) = O(\log(1/\delta)/\beta) + O(1/\epsilon),$$

we have for $N \geq N(\beta, \epsilon, \delta)$, $\mathbb{P}(F(X_N) > \epsilon) \leq \delta$.

References

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems*, 2018.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, 2016.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

X. Chen, S. Du, and X. T. Tong. On stationary-point hitting time and ergodicity of stochastic gradient Langevin dynamics. arXiv:1904.13016, 2019a.

Y. Chen, J. Chen, J. Dong, J. Peng, and Z. Wang. Accelerating nonconvex learning via replica exchange Langevin diffusion. In *International Conference on Learning Representations*, 2019b.

Ran Cheng and Yaochu Jin. A competitive swarm optimizer for large scale optimization. *IEEE transactions on cybernetics*, 45(2):191–204, 2014.

X. Cheng, N.S. Chatterji, P.L. Bartlett, and M.I. Jordan. Underdamped langevin MCMC: A non-asymptotic analysis. In *Machine Learning Research*, volume 75, pages 1–24, 2018.

Tiangang Cui, James Martin, Youssef M Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014.

Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathrm{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2014.

A.S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. *J. R. Stat. Soc. B*, 79:651–676, 2017.

A.S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129:5278–5311, 2019.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *Proceedings of the International Conference on Machine Learning*, 2018.

Jing Dong and Xin T Tong. Spectral gap of replica exchange langevin diffusion on mixture distributions. *arXiv preprint arXiv:2006.16193*, 2020.

Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *Proceedings of the International Conference on Machine Learning*, 2018.

P. Dupuis, Y. Liu, N. Plattner, and J.D. Doll. On the infinite swapping limit for parallel tempering. *Multiscale Model Simulation*, 10(3):986–1022, 2012.

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

D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7:3910–3916, 2005.

Susana C Esquivel and CA Coello Coello. On the use of particle swarm optimization with multimodal functions. In *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, volume 2, pages 1130–1136. IEEE, 2003.

Roger Gämperle, Sibylle D Müller, and Petros Koumoutsakos. A parameter study for differential evolution. *Advances in intelligent systems, fuzzy systems, evolutionary computation*, 10(10):293–298, 2002.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the International Conference on Machine Learning*, 2017.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *Proceedings of the International Conference on Learning Representations*, 2018.

S.B. Gelfand and S.K. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Charles J. Geyer and Elizabeth A. Thompson. Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90 (431):909–920, 1995.

B. Gidas. Nonstationary markov chains and convergence of the annealing algorithm. *Journal of Statistical Physics*, 39(1/2), 1985.

Vincent Granville, Mirko Krivánek, and J-P Rasson. Simulated annealing: A proof of convergence. *IEEE transactions on pattern analysis and machine intelligence*, 16(6): 652–656, 1994.

M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 24(6):2455–2490, 2014.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the International Conference on Machine Learning*, 2017.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Proceedings of the Conference on Learning Theory*, 2018.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. 2019.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In *NeurIPS*, 2018.

Y. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. In *International Conference on Neural Information Processing Systems*, volume 2, pages 2917–2925, 2015.

Oren Mangoubi and Nisheeth K Vishnoi. Convex optimization with unbounded nonconvex oracles using simulated annealing. In *Conference On Learning Theory*, pages 1086–1124. PMLR, 2018.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto I Oliveira. Solving SDPs for synchronization and maxcut problems via the Grothendieck inequality. In *Proceedings of the Conference on Learning Theory*, 2017.

Georg Menz and André Schlichting. Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014.

Matthias Morzfeld, Xin T Tong, and Youssef M Marzouk. Localization for mcmc: sampling high-dimensional posterior distributions with local structure. *Journal of Computational Physics*, 380:1–28, 2019.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program., Ser. A*, 103: 127–152, 2005.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course.*, volume 87. Springer Science+Business Media, 2013.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, 2017.

M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Annals of Applied Probability*, pages 10–44, 1998.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the Conference on Learning Theory*, 2017.

Robert H Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

Xin T. Tong, Mathias Morzfeld, and Youssef M. Marzouk. Mala-within-gibbs samplers for high-dimensional distributions with sparse conditional structure. *SIAM Journal on Scientific Computing*, 42(3):A1765–A1788, 2020.

Santosh S Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *arXiv preprint arXiv:1903.08568*, 2019.

D. B. Woodard, S. C. Schmidler, and M. Huber. Conditions for rapid mixing of parallel and simulated termpering on multimodal distribution. *Annals of Applied Probability*, 19 (2):617–640, 2009.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2018.

Yaodong Yu, Pan Xu, and Quanquan Gu. Third-order smoothness helps: Faster stochastic optimization algorithms for finding local minima. In *Advances in Neural Information Processing Systems*, 2018.