Improved Convergence Rate of Stochastic Gradient Langevin Dynamics with Variance Reduction and its Application to Optimization

Yuri Kinoshita
Department of Mathematical Informatics,
Graduate School of Information Science and Technology,
The University of Tokyo, Tokyo, Japan
yuri-kinoshita111@g.ecc.u-tokyo.ac.jp

Taiji Suzuki
Department of Mathematical Informatics,
Graduate School of Information Science and Technology,
The University of Tokyo, Tokyo, Japan
Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan
taiji@mist.i.u-tokyo.ac.jp

Abstract

The stochastic gradient Langevin Dynamics is one of the most fundamental algorithms to solve sampling problems and non-convex optimization appearing in several machine learning applications. Especially, its variance reduced versions have nowadays gained particular attention. In this paper, we study two variants of this kind, namely, the Stochastic Variance Reduced Gradient Langevin Dynamics and the Stochastic Recursive Gradient Langevin Dynamics. We prove their convergence to the objective distribution in terms of KL-divergence under the sole assumptions of smoothness and Log-Sobolev inequality which are weaker conditions than those used in prior works for these algorithms. With the batch size and the inner loop length set to $\sqrt{n}$, the gradient complexity to achieve an ϵ-precision is $\tilde{O}((n + dn^{1/2}\epsilon^{-1})\gamma^2 L^2 \alpha^{-2})$, which is an improvement from any previous analyses. We also show some essential applications of our result to non-convex optimization.

1 Introduction

1.1 Background and Organization

Over the past decade, the gradient Langevin Dynamics (GLD) has gained particular attention for providing an effective tool for sampling from a Gibbs distribution, a fundamental task omnipresent in the field of machine learning and statistics, and for non-convex optimization, which is nowadays witnessing an unignorable empirical success. Notably, GLD is a stochastic differential equation (SDE) that can be viewed as the steepest descent flow of the Kullback-Leibler (KL) divergence towards the stationary Gibbs distribution in the space of measures endowed with the 2-Wasserstein metric (Jordan et al., 1998). As a consequence of the unique properties of GLD, its implementable discrete schemes and their ability to suitably track it have been the subject of a large number of studies.

The Euler-Maruyama scheme of GLD gives rise to an algorithm known as the Langevin Monte Carlo method (LMC). This algorithm is biased (Wibisono, 2018): that is, the distribution of the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

discrete scheme does not converge to the same as GLD. Nonetheless, it has been shown that this bias could be made arbitrarily small under certain assumptions by taking a sufficiently small step size (Dalalyan, 2017b; Vempala and Wibisono, 2019). Dalalyan (2017a,b) provided one of the first non-asymptotic rates of convergence of LMC for smooth log-concave distributions. Assumptions to obtain a non-asymptotic analysis and this controllable bias have been relaxed by further research to dissipativity and smoothness (Raginsky et al., 2017; Xu et al., 2018), and recently to Log-Sobolev inequality (LSI) and smoothness (Vempala and Wibisono, 2019). This relaxation of conditions is especially meaningful as the objective distribution nowadays tends to become more and more complicated beyond the classical assumption of log-concavity.

However, in the field of machine learning, the main function can often be formulated as the average of the loss function of an enormous number of training data points (Welling and Teh, 2011), which subsequently makes it difficult to calculate its full gradient. As a result, research on stochastic algorithms has been also conducted to avoid this computational burden (Chen et al., 2021; Dubey et al., 2016; Raginsky et al., 2017; Welling and Teh, 2011; Xu et al., 2018; Zou et al., 2018, 2019a,b, 2021). Welling and Teh (2011) introduced the concept of Stochastic Gradient Langevin Dynamics (SGLD) which combines the Stochastic Gradient Descent with LMC. This has been the subject of successful studies (Raginsky et al., 2017; Welling and Teh, 2011; Xu et al., 2018). Nevertheless, the variance of its stochastic gradient is too large, which leads to a slow convergence compared to LMC. Therefore, stochastic gradient Langevin Dynamics algorithms with variance reduction, such as the Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD), have been considered and their convergence has been thoroughly analyzed for both sampling (Dubey et al., 2016; Zou et al., 2018, 2019a, 2021) and optimization (Huang and Becker, 2021; Xu et al., 2018).

Dubey et al. (2016) first united SGLD with variance reduction techniques and proposed two new algorithms, namely, SVRG-LD and SAGA-LD. Chatterji et al. (2018) and Zou et al. (2018) proved the convergence rate of SVRG-LD to the target distribution in 2-Wasserstein distance for smooth log-concave distributions. Xu et al. (2018) showed the weak convergence of SVRG-LD under the smoothness and dissipativity conditions. They expanded the non-asymptotic analysis of Raginsky et al. (2017) to LMC and SVRG-LD, and improved the result for SGLD. Few years ago, Zou et al. (2019a) provided the gradient complexity of SVRG-LD to converge to the stationary distribution in 2-Wasserstein distance under the smoothness and dissipativity assumptions. This convergence can be even improved if we make a warm-start (Zou et al., 2021). While these works investigated algorithms with fixed hyperparameters, Huang and Becker (2021) additionally assumed a strict saddle and some other minor conditions to study SVRG-LD with a decreasing step size and improved its convergence in high probability to the second order stationary point. Zou et al. (2019b) also applied variance reduction techniques to the Hamiltonian Langevin Dynamics, or underdamped Langevin Dynamics in opposition to GLD also known as overdamped Langevin Dynamics. As we can observe, the current convergence analyses of the stochastic schemes with variance reduction are mostly restricted to log-concavity and dissipativity, and do not enjoy the same broad convergence guarantee with a concrete gradient complexity as LMC does under LSI and smoothness in terms of KL-divergence.

Therefore, in order to bridge this theoretical gap between LMC and stochastic gradient Langevin Dynamics with variance reduction, we study in this paper the convergence of the latter under the relaxed assumptions of smoothness and LSI. In Section 3, we study the convergence to the Gibbs distribution of SVRG-LD and the Stochastic Recursive Gradient Langevin Dynamics (SARAH-LD), another variant of stochastic gradient Langevin Dynamics with variance reduction inspired by the Stochastic Recursive Gradient algorithm (SARAH) of Nguyen et al. (2017a,b). On the other hand, optimization and sampling are only two sides of the same coin for GLD. That is why, in Section 4, we also investigate implications of Section 3 for non-convex optimization. We prove the convergence of SVRG-LD and SARAH-LD to the global minimum of dissipative functions and we provide their non-asymptotic rate of convergence. We also consider the additional Morse assumption and study its effect. Finally, we illustrate our main result with a simple experiment.

1.2 Contributions

The major contributions of this paper can be summarized as follows. We provide a non-asymptotic analysis of the convergence of SVRG-LD and SARAH-LD to the Gibbs distribution in terms of KL-divergence under smoothness and LSI which are weaker conditions than those used in prior works for these algorithms. KL-divergence is generally a stronger convergence criterion than both total

Table 1: Comparison of our main result with prior works (sampling). The first three works are about LMC. Compared to Vempala et al. (2019), with the same assumptions and criterion, the order of gradient complexity is improved from n to $\sqrt{n}$. The others are about SVRG-LD except the last one which is about the Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction. ϵ is the accuracy required on the criterion, d is the dimension of the input of the main function, n is the number of data points, and L is the smoothness constant. * 2-Wass. stands for "2-Wasserstein", and conv. stands for "convergence". ** $\mathrm{poly}(M, L)$ stands for a polynomial of M and L.

Method	Major Assumptions	Criterion*	Gradient Complexity**
Dalalyan (2017a)	Smooth, Log-concave (M)	2-Wass.	$\tilde{O}\big(\frac{nd}{\epsilon^2}\cdot\mathrm{poly}(M,L)\big)$
Xu et al. (2018)	Smooth, Dissipative	Weak conv.	$\tilde{O}\big(\frac{nd}{\epsilon}\big)\cdot\mathrm{e}^{\tilde{O}(d)}$
Vempala et al. (2019)	Smooth, Log-Sobolev (α)	KL	$\tilde{O}\big(\frac{n}{\epsilon}\cdot d\gamma^2L^2\alpha^{-2}\big)$
Zou et al. (2018)	Smooth, Log-concave (M)	2-Wass.	$\tilde{O}\Big(n+\frac{L^{3/2}n^{1/2}d^{1/2}}{M^{3/2}\epsilon}\Big)$
Zou et al. (2019a)	Smooth, Dissipative	2-Wass.	$\tilde{O}\Big(n+\frac{n^{3/4}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon^4}\Big)\cdot\mathrm{e}^{\tilde{O}(\gamma+d)}$
Zou et al. (2021)	Smooth, Dissipative, Warm-start	TV	$\tilde{O}\Big(\frac{\gamma^2}{\epsilon^2}\Big)\cdot\mathrm{e}^{\tilde{O}(d)}$
Zou et al. (2019b)	Smooth, Dissipative	2-Wass.	$\tilde{O}\Big((n+\frac{n^{1/2}}{\epsilon^2\mu_*^{3/2}})\wedge\frac{\mu_*^{-2}}{\epsilon^4}\Big)$
This paper	Smooth, Log-Sobolev (α)	KL	$\tilde{O}\Big(\Big(n+\frac{dn^{1/2}}{\epsilon}\Big)\cdot\gamma^2L^2\alpha^{-2}\Big)$

variation (TV) and 2-Wasserstein distance as they can be controlled by KL-divergence under the LSI condition. Notably, we prove that, with the batch size and inner loop length set to $\sqrt{n}$, the gradient complexity to achieve an ϵ-precision in terms of KL-divergence is $\tilde{O}((n + dn^{1/2}\epsilon^{-1})\gamma^2L^2\alpha^{-2})$, which is better than any previous analyses. See Table 1 for a comparison with previous research in terms of assumptions, criterion and gradient complexity. We also prove the convergence of SVRG-LD and SARAH-LD to the global minimum under an additional assumption of dissipativity with a gradient complexity of $\tilde{O}((n + n^{1/2}\epsilon^{-1}dL\alpha^{-1})\gamma^2L^2\alpha^{-2})$ which is better than previous work since it has almost all the time a dependence on n of $O(\sqrt{n})$ and does not require the batch size and the inner loop length to depend on the accuracy ϵ. On the other hand, we import the idea of Li and Erdogdu (2020) from product manifolds of spheres to the Euclidean space in order to show that under the additional assumption of Morse, the convergence in the Euclidean space can be accelerated by eliminating the exponential dependence on $1/\epsilon$.

1.3 Other Related Works

The theoretical study of GLD goes back to Chiang et al. (1987) who showed that global convergence could be achieved with a proper annealing schedule. This work did not specify how to implement this SDE, but Gelfand and Mitter (1991) filled this gap. Later, Borkar and Mitter (1999) proved an asymptotic convergence in terms of relative entropy for the discrete scheme of gradient Langevin Dynamics when the inverse temperature and the step size are kept constant.

The variance reduction technique, introduced to Langevin Dynamics by Dubey et al. (2016), was originally presented by Johnson and Zhang (2013) as Stochastic Variance Reduced Gradient (SVRG) to improve the convergence speed of Stochastic Gradient Descent. Other variance reduction techniques were also considered such as the Stochastic Recursive Gradient Langevin Dynamics (SARAH) from Nguyen et al. (2017a,b) which outperforms SVRG in non-convex optimization (Pham et al., 2020) and is used in many algorithms such as SSRGD (Li, 2019) and SpiderBoost (Wang et al., 2019).

Li and Erdogdu (2020) extended Vempala and Wibisono's result to Riemannian manifolds. One of the highlights of their work is that they showed the Log-Sobolev constant of the Gibbs distribution for a product manifold of spheres only depends on a polynomial of the inverse temperature under some particular conditions including Morse. We will adapt this result to our situation.

In the concurrent work of Balasubramanian et al. (2022) (especially Section 6), they also studied the convergence of stochastic schemes of GLD with more relaxed conditions than prior analyses. However, our contributions are not overshadowed by theirs, and we clarify the reasons. In Subsection

6.1 of their paper, Balasubramanian et al. (2022) focused on stochastic discrete schemes with finite variance and bias (which is not the case for SVRG-LD) and provided a first-order convergence guarantee in the space of measures equipped with the 2-Wasserstein distance. Subsection 6.2 proved a global convergence under some other conditions but most of these two analyses did not consider in particular the usual case in machine learning when F is the average of some other functions, which leads to a generally worse gradient complexity than ours. Concerning this finite sum setting, Balasubramanian et al. (2022) investigated the Variance Reduced LMC algorithm (slightly different from SVRG-LD in this paper) in Subsection 6.3 and gave a first-order convergence under the sole assumption of smoothness. When restrained in our problem setting, the gradient complexity of SVRG-LD and SARAH-LD we provide is still considerably better (see Section 3 for more details).

1.4 Notation

We denote deterministic vectors by a lower case symbol (e.g., x) and random variables by an upper case symbol (e.g., X). The Euclidean norm is denoted by $\|\cdot\|$ for vectors and the inner product by $\langle\cdot,\cdot\rangle$. For matrices, $\|\cdot\|$ is the norm induced by the Euclidean norm for vectors. We only treat distributions absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$ for simplicity. Especially, throughout the paper, ν refers to the probability measure with the density function $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$, where F is a function introduced below. $a \vee b$ is equivalent to $\max\{a, b\}$ and $a \wedge b$ to $\min\{a, b\}$. We also use the shorthand $\tilde{O}$ to hide logarithmic polynomials.

2 Preliminaries

In this section, we briefly explain the problem setting, necessary mathematical background and assumptions used in this paper.

2.1 Problem Setting and GLD

In Section 3, we consider sampling from a distribution written in the form $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$ where γ is a positive constant (which corresponds to the inverse temperature) and $F : \mathbb{R}^d \to \mathbb{R}$ is formulated as $F(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)$, the average of the loss function of n training data points $\{x^{(i)}\}_{i=1}^n$. Here, $f_i(x) := f(x, x^{(i)})$ can be regarded as the loss of data $x^{(i)}$. For instance, F can be the average of the negative log likelihood of n training data points. In Section 4, we consider the non-convex optimization (minimization) of the same F as above.

GLD can be described as the following stochastic differential equation (SDE):

$$\mathrm{d}X_t^{\mathrm{GLD}} = -\nabla F(X_t^{\mathrm{GLD}})\mathrm{d}t + \sqrt{2/\gamma}\mathrm{d}B(t), \tag{1}$$

where $\gamma > 0$ is called the inverse temperature parameter and $\{B(t)\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$. It can be used for sampling since under some reasonable assumptions of F, the distribution ρ_t^{GLD} of X_t^{GLD} governed by SDE (1) converges to the invariant stationary distribution $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$, also known as the Gibbs distribution (Chiang et al., 1987). Moreover, as previously mentioned, this convergence is efficient in the sense that SDE (1) corresponds to the steepest descent flow of the Kullback-Leibler (KL) divergence towards the stationary distribution in the space of measures endowed with the 2-Wasserstein metric (Jordan et al., 1998). Alternatively, GLD can be interpreted as the composite optimization problem of a negative entropy and an expected function value as follows (Wibisono, 2018):

$$\min_{q:\text{density}} \mathbb{E}_q[\gamma F] + \mathbb{E}_q[\log q].$$

The gradient flow is the well-known Fokker-Planck equation associated to SDE (1):

$$\frac{\partial \rho_t^{\mathrm{GLD}}}{\partial t} = \nabla\cdot(\rho_t^{\mathrm{GLD}}\nabla F) + \frac{1}{\gamma}\Delta\rho_t^{\mathrm{GLD}} = \frac{1}{\gamma}\nabla\cdot\left(\rho_t^{\mathrm{GLD}}\nabla\log\frac{\rho_t^{\mathrm{GLD}}}{\nu}\right). \tag{2}$$

This will be useful in our analysis. In addition to its potential for sampling, GLD can also be employed for non-convex optimization as the Gibbs distribution concentrates on the global minimum of F for sufficiently large values of γ (Hwang, 1980).

Algorithm 1: SVRG-LD / SARAH-LD

```
input: step size $\eta > 0$, batch size $B$, epoch length $m$, inverse temperature $\gamma \geq 1$
initialization: $X_0 = 0$, $X^{(0)} = X_0$
foreach $s = 0, 1, \ldots, (K/m)$ do
    $v_{sm} = \nabla F(X^{(s)})$
    randomly draw $\epsilon_{sm} \sim N(0, I_{d\times d})$
    $X_{sm+1} = X_{sm} - \eta v_{sm} + \sqrt{2\eta/\gamma}\epsilon_{sm}$
    foreach $l = 1, \ldots, m-1$ do
        $k = sm + l$
        randomly pick a subset $I_k$ from $\{1, \ldots, n\}$ of size $|I_k| = B$
        randomly draw $\epsilon_k \sim N(0, I_{d\times d})$
        if SVRG-LD then
            $v_k = \frac{1}{B}\sum_{i_k \in I_k}(\nabla f_{i_k}(X_k) - \nabla f_{i_k}(X^{(s)})) + v_{sm}$
        else if SARAH-LD then
            $v_k = \frac{1}{B}\sum_{i_k \in I_k}(\nabla f_{i_k}(X_k) - \nabla f_{i_k}(X_{k-1})) + v_{k-1}$
        end
        $X_{k+1} = X_k - \eta v_k + \sqrt{2\eta/\gamma}\epsilon_k$
    end
    $X^{(s+1)} = X_{(s+1)m}$
end
```

2.2 Algorithms of GLD

Applying the Euler-Maruyama scheme to (1), we obtain the Langevin Monte Carlo (LMC)

$$X_{k+1} = X_k - \eta \nabla F(X_k) + \sqrt{2\eta/\gamma}\epsilon_k,$$

where η is called the step size. This is similar to the gradient descent except the additional Gaussian noise $\sqrt{2\eta/\gamma}\epsilon_k$, where $\epsilon_k \sim N(0, I_{d\times d})$ and $I_{d\times d}$ is the $d \times d$ unit matrix. In the case n is huge and the computation of ∇F is too difficult, we are incited to use stochastic gradient methods in analogy to stochastic gradient optimization. This gives

$$X_{k+1} = X_k - \eta v(X_k) + \sqrt{2\eta/\gamma}\epsilon_k,$$

where $v(X_k)$ is the stochastic gradient. When $v(X_k)$ is defined as $\frac{1}{B}\sum_{i_k \in I_k} \nabla f_{i_k}(X_k)$, where B is called the batch size and I_k is a random subset uniformly chosen from $\{1, \ldots, n\}$ such that $|I_k| = B$, we obtain the Stochastic Gradient Langevin Dynamics (SGLD). As this method exhibits a slow convergence, it has been popular to use variance reduction methods such as the Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD) where $v(X_k) = \frac{1}{B}\sum_{i_k \in I_k}(\nabla f_{i_k}(X_k) - \nabla f_{i_k}(X^{(s)})) + \nabla F(X^{(s)})$. Details of this algorithm is stated in Algorithm 1. $X^{(s)}$ is a reference point updated every m steps so that $X_{sm} = X^{(s)}$. As we can observe in Lemma A.4, around the optimal point, the variance of the stochastic gradient is indeed decreased as $X^{(s)}$ and X_k are both close to each other. We can also easily extend some successful stochastic gradient algorithms to Langevin Dynamics. Hence, we are motivated to extend the Stochastic Recursive Gradient Algorithm (SARAH) to Langevin Dynamics since we can expect that some bottlenecks of the analysis of SVRG-LD can be removed in that of SARAH-LD as subtracting the previous stochastic gradient enables a stabler performance than SVRG-LD. This algorithm can be described as Algorithm 1 with $v(X_k) = \frac{1}{B}\sum_{i_k \in I_k}(\nabla f_{i_k}(X_k) - \nabla f_{i_k}(X_{k-1})) + v(X_{k-1})$.

Definition 1. *We define ρ_k as the distribution of X_k generated at the kth step of SVRG-LD, and similarly ϕ_k for SARAH-LD.*

2.3 Assumptions

The assumptions used throughout this paper can be summarized as follows.

Assumption 1. *For all $i = 1, \ldots, n$, ∇f_i is twice differentiable, and $\forall x, y \in \mathbb{R}^d$, $\|\nabla^2 f_i(x)\| \leq L$. In other words, f_i ($i = 1, \ldots, n$) and F are L-smooth.*

Assumption 2. *Distribution ν satisfies the Log-Sobolev inequality (LSI) with a constant α. That is, for all probability density functions ρ absolutely continuous with respect to ν, the following holds:*

$$H_\nu(\rho) \leq \frac{1}{2\alpha} J_\nu(\rho),$$

where $H_\nu(\rho) := \mathbb{E}_\rho \left[\log \frac{\rho}{\nu}\right]$ is the KL-divergence of ρ with respect to ν, and $J_\nu(\rho) := \mathbb{E}_\rho \left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right]$ is the relative Fisher information of ρ with respect to ν.

The recent work of Vempala and Wibisono (2019) motivates us to use the combination of smoothness and LSI for the analysis of SVRG-LD and SARAH-LD. Indeed, they showed that these conditions were enough to assure for the Euler-Maruyama scheme an exponentially fast convergence and a bias controllable by the step size. Under smoothness, LSI is not only the necessary condition of log-concavity and dissipativity, but is also robust to bounded perturbation and Lipschitz mapping, contrary to log-concavity (Vempala and Wibisono, 2019). For example, for any distribution $\mathrm{d}\nu$ that satisfies LSI and bounded function $B : \mathbb{R}^d \to \mathbb{R}$, $\mathrm{d}\tilde{\nu} \propto \mathrm{e}^B \mathrm{d}\nu$ satisfies LSI as well (Holley and Stroock, 1986). Moreover, while KL-divergence is not in general convex with regard to the Wasserstein geodesic, thanks to LSI, the Polyak-Łojaciewicz condition is satisfied. It is well-known that LSI suffices to realize an exponential convergence for the case of continuous time Langevin Dynamics (Vempala and Wibisono, 2019). That is why, it is actually both useful and natural to suppose LSI in this context. Note that under L-smoothness of F and LSI with constant α for $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$, it holds that $\alpha \leq \gamma L$ (Vempala and Wibisono, 2019).

As for optimization, we additionally use the following conditions.

Assumption 3. *F is (M, b)-dissipative. That is, there exist constants $M > 0$ and $b > 0$ such that for all $x \in \mathbb{R}^d$ the following holds: $\langle \nabla F(x), x\rangle \geq M\|x\|^2 - b$.*

Assumption 4 (Li and Erdogdu (2020), Assumption 3.3 adapted)**.** *F satisfies the Morse condition. That is, for all eigenvalues of the Hessian of stationary points, there exists a constant $\lambda^\dagger \in (0, 1]$ such that*

$$\lambda^\dagger \leq \inf \left\{\left|\lambda_i \left(\nabla^2 F(x)\right)\right| \mid \nabla F(x) = 0,\ i \in 1, \dots, d\right\}.$$

Furthermore, for the set $\mathcal{S}$ of stationary points that are not a global minimum, $\sup_{x\in\mathcal{S}} \lambda_{\min}\left(\nabla^2 F(x)\right) \leq -\lambda^\dagger$.

Assumption 5. *$\nabla^2 f_i$ is L'-Lipschitz and without loss of generality, we let $\min_{x\in\mathbb{R}^d} F(x) = 0$.*

Assumption 6. *F has a unique global minimum.*

Smoothness and dissipativity are a classical combination of assumptions for this kind of problem setting (Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2019a). We assume dissipativity instead of LSI for non-convex optimization in order to obtain an explicit value of the Log-Sobolev constant of $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$ in function of the inverse temperature parameter γ (see Property C.3), making a non-asymptotic analysis possible. Furthermore, Assumptions 4 to 6 can ameliorate the exponential dependence of the inverse of the Log-Sobolev constant on the inverse temperature parameter to a polynomial one (see Property C.4).

3 Main Results

In this section, we state our main results which prove that SVRG-LD and SARAH-LD (Algorithm 1) achieve an exponentially fast convergence to the Gibbs distribution and a controllable bias in terms of KL-divergence under the sole assumptions of LSI and smoothness. We provide their gradient complexity as well. The proofs can be found in Appendix A and B respectively.

3.1 Improved Convergence of SVRG-LD

Our analysis shows that the convergence of SVRG-LD to the stationary distribution $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$ can be formulated as the theorem below.

Theorem 1. *Under Assumptions 1 and 2, $0 < \eta < \frac{\alpha}{16\sqrt{6}L^2 m\gamma}$, $\gamma \geq 1$ and $B \geq m$, for all $k = 1, 2, \dots$, the following holds in the update of SVRG-LD where $\Xi = \frac{(n-B)}{B(n-1)}$:*

$$H_\nu(\rho_k) \leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\rho_0) + \frac{224\eta\gamma d L^2}{3\alpha}\left(2 + 3\Xi + 2m\Xi\right).$$

We observe that the bias term of the upper bound, which is the second term linearly dependent on η, can be easily controlled while the first term exponentially converges to 0 with $k \to \infty$. This is more precisely formulated in the following corollary.

Corollary 1.1. *Under the same assumptions as Theorem 1, for all $\epsilon \geq 0$, if we choose step size η such that $\eta \leq \frac{3\alpha\epsilon}{448\gamma d L^2}$, then a precision $H_\nu(\rho_k) \leq \epsilon$ is reached after $k \geq \frac{\gamma}{\alpha\eta} \log \frac{2H_\nu(\rho_0)}{\epsilon}$ steps. Especially, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \frac{\alpha}{16\sqrt{6}L^2\sqrt{n}\gamma} \wedge \frac{3\alpha\epsilon}{448dL^2\gamma}$, then the gradient complexity becomes*

$$\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right).$$

This gradient complexity is an improvement compared with prior works for three reasons. First of all, we provide a non-asymptotic analysis of the convergence of SVRG-LD under smoothness and Log-Sobolev inequality which are conditions weaker than those (e.g., log-concavity or dissipativity) used in prior works for these algorithms. Moreover, we prove it in terms of KL-divergence which is generally a stronger convergence criterion than both total variation (TV) and 2-Wasserstein distance as they can both be controlled by KL-divergence under the LSI condition. For instance, TV was used by Zou et al. (2021) and 2-Wasserstein distance by Dalalyan (2017a) and Zou et al. (2019a). KL-divergence makes it possible to unify these two different criteria. Finally, while prior research generally used Girsanov's theorem which generates a bias term that accumulates through the iteration (see for example Raginsky et al. (2017) and Xu et al. (2018)), we solve this issue by taking benefit of the exponential convergence of GLD to the Gibbs distribution under LSI and smoothness that enables us to forget about past bias. That way, with the batch size and inner loop set to $\sqrt{n}$, the gradient complexity to achieve an ϵ-precision in terms of KL-divergence becomes $\tilde{O}((n + dn^{1/2}\epsilon^{-1})\gamma^2 L^2 \alpha^{-2})$, which is better than previous analyses. For example, Vempala and Wibisono (2019) provided a gradient complexity of $\tilde{O}\left(n\epsilon^{-1}d\gamma^2 L^2 \alpha^{-2}\right)$ for LMC under Assumptions 1 and 2, and Zou et al. (2019a) a gradient complexity of $\tilde{O}(n + n^{3/4}\epsilon^{-2} + n^{1/2}\epsilon^{-4}) \cdot \mathrm{e}^{\tilde{O}(\gamma+d)}$ for SVRG-LD under Assumptions 1 and 3. Note that the dependence on the dimension d is not improved since α^{-1} may exponentially depend on d. Recently, Zou et al. (2019b) proposed the Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction with a gradient complexity of $\tilde{O}((n + n^{1/2}\epsilon^{-2}\mu_*^{-3/2}) \wedge \mu_*^{-2}\epsilon^{-4})$ in terms of 2-Wasserstein distance. Even though their algorithm is based on the underdamped Langevin Dynamics whose discrete schemes use to perform better than those of the overdamped Langevin Dynamics such as SVRG-LD, our gradient complexity, which applies to a broader family of distributions, is almost the same except for a small interval of ϵ, but we do not require the batch size B and the inner loop length m to depend on ϵ while Zou et al. (2019b) do, i.e., $B \lesssim B_0^{1/2}$, $m = O(B_0/B)$, where $B_0 = \tilde{O}\left(\epsilon^{-4}\mu_*^{-1} \wedge n\right)$. This strengthens the importance of our result since it shows that adapting this analysis to other stochastic schemes of GLD is promising and could lead to tighter bounds and relaxation of conditions. See Table 1 for a summary. Concerning the concurrent work of Balasubramanian et al. (2022), under the sole assumption of smoothness, they provided a gradient complexity of $O(L^2 d^2 n/\epsilon^2)$ for the Variance Reduced LMC algorithm that updates the stochastic gradient differently as SVRG-LD and SARAH-LD. This is almost the square of our result, and in some extent, our work can be interpreted as an acceleration of their result with a slightly stronger additional condition than Poincaré inequality.

Proof Sketch Proceeding in a similar way as Vempala and Wibisono (2019), we evaluate how $H_\nu(\rho_k)$ decreases at each step as shown in Theorem A.1 of Appendix A. This is realized by comparing the evolution of the continuous-time GLD for time η and one step of SVRG-LD. Since we use a stochastic gradient, we need at the same time to evaluate the variance of the stochastic gradient. Theorem 1 can be obtained by recursively solving the inequality derived in Theorem A.1.

3.2 Convergence Analysis of SARAH-LD

As for SARAH-LD, its convergence to the stationary distribution $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$ can be formulated as the theorem below. Interestingly, we obtain the same result as SVRG-LD (Theorem 1) but we do not require $B \geq m$ anymore.

Theorem 2. *Under Assumptions 1 and 2, $0 < \eta < \frac{\alpha}{16\sqrt{2}L^2 m\gamma}$ and $\gamma \geq 1$, for all $k = 1, 2, \ldots$, the following holds in the update of SARAH-LD where $\Xi = \frac{(n-B)}{B(n-1)}$:*

$$H_\nu(\phi_k) \leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\phi_0) + \frac{32\eta\gamma d L^2}{3\alpha}\left(2 + \Xi + 2m\Xi\right).$$

This is the first convergence guarantee of SARAH-LD in this problem setting so far, and it leads to the following gradient complexity.

Corollary 2.1. *Under the same assumptions as Theorem 2, for all $\epsilon \geq 0$, if we choose step size η such that $\eta \leq \frac{3\alpha\epsilon}{64\gamma d L^2}\left(2 + \Xi + 2m\Xi\right)^{-1}$, then a precision $H_\nu(\phi_k) \leq \epsilon$ is reached after $k \geq \frac{\gamma}{\alpha\eta}\log\frac{2H_\nu(\phi_0)}{\epsilon}$ steps. Especially, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \frac{\alpha}{16\sqrt{2}L^2\sqrt{n}\gamma} \wedge \frac{3\alpha\epsilon}{320 d L^2\gamma}$, then the gradient complexity becomes*

$$\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right).$$

The reason why we obtain the same gradient complexity for both SARAH-LD and SVRG-LD (except better coefficients for SARAH-LD) is that in our analysis, the Brownian noise added at each step of the Langevin Dynamics plays the role of a fundamental bottleneck that even SARAH-LD could not eliminate, and we still need to set $B = m = \sqrt{n}$. We can hypothesize that this order of gradient complexity might be tight for variance-reduced stochastic gradient Langevin Dynamics algorithms.

4 Some Applications to Non-Convex Optimization

Here, we apply our main results to non-convex optimization. Thanks to our analysis applicable to a broader family of probability distributions satisfying LSI, the additional conditions we pose in this section are mainly reflected in the concrete formulation of the Log-Sobolev constant, which keeps our study simple and clear. The proofs can be found in Appendix C. Since SVRG-LD and SARAH-LD exhibited almost the same performance in sampling, we can simultaneously analyse them. We first prove the convergence to the global minimum of SVRG-LD and SARAH-LD without clarifying the explicit formulation of the Log-Sobolev constant in function of γ.

Theorem 3. *Using SVRG-LD or SARAH-LD, under Assumptions 1 to 3, $0 < \eta < \frac{\alpha}{16\sqrt{6}L^2 m\gamma}$, $\gamma \geq \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2} \vee 1 \vee \frac{2}{M}$ and $B \geq m$, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \frac{\alpha}{16\sqrt{6}L^2\sqrt{n}\gamma} \wedge \frac{3}{1792}\frac{\alpha^2\epsilon}{L^2 d\gamma}$, the gradient complexity to reach a precision of*

$$\mathbb{E}_{X_k}[F(X_k)] - F(X^*) \leq \epsilon$$

is

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon} \cdot \frac{dL}{\alpha}\right)\frac{\gamma^2 L^2}{\alpha^2}\right),$$

where α is a function of γ, and X^* is the global minimum of F.

Remark 1. *Under Assumptions 1 and 3, Assumption 2 is negligible as shown in Property C.2.*

Under Assumptions 1 to 3 only, this leads to a gradient complexity which exponentially depends on the inverse of the precision level ϵ as shown in the next corollary since the inverse of the Log-Sobolev constant exponentially depends on γ.

Corollary 3.1. *Under the same assumptions as Theorem 3, taking $\gamma = i(\epsilon) := \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2} \vee 1 \vee \frac{2}{M}$, we obtain a gradient complexity of*

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon} \cdot \frac{dL}{C_1 i(\epsilon)}\mathrm{e}^{C_2 i(\epsilon)}\right) L^2 \mathrm{e}^{2C_2 i(\epsilon)}\right)$$

since $\alpha = \gamma C_1 \mathrm{e}^{-C_2\gamma}$ (Property C.3).

The second term with $n^{1/2}$ is almost all the time dominant since it has a factor that exponentially depends on $1/\epsilon$ and the first term not. This dependence on n of $O(n^{1/2})$ is the best so far for these algorithms. Moreover, comparing with the gradient complexity $\tilde{O}\left(n^{1/2}\lambda^{-4}\epsilon^{-5/2}\right) \cdot \mathrm{e}^{\tilde{O}(d)}$, also of order $n^{1/2}$, provided by Xu et al. (2018) who used SVRG-LD and the same assumptions, our gradient complexity is an improvement since their analysis required a batch size B and an inner loop length m that strongly depend on ϵ (i.e., $B = \sqrt{n}\epsilon^{-3/2}$, $m = \sqrt{n}\epsilon^{3/2}$) and ours does not. Note that the dependence of the gradient complexity of Xu et al. (2018) on $1/\epsilon$ is not necessarily better than ours as λ is actually the spectral gap of the discrete-time Markov chain generated by (1) and its inverse exponentially depends on $1/\epsilon$ as well. Although Xu et al. (2018) did not investigate the explicit nature of λ, this is supported by Raginsky et al. (2017) who proved this exponential dependence for the spectral gap of the continuous-time SDE and by Mattingly et al. (2002) who showed the spectral gap of continuous-time SDE and that of discrete-time version are almost the same in this context.

Analysis under the Morse condition Now, under the additional Assumptions 4 to 6, it is interesting to note that a *polynomial dependence* on $1/\epsilon$ is achieved as the following corollary shows.

Corollary 3.2. *Under the same assumptions as Theorem 3 and Assumptions 4 to 6, taking* $\gamma = j(\epsilon) := \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2} \vee 1 \vee \frac{2}{M} \vee C_\gamma$*, where* C_γ *is a constant independent of* ϵ *defined in Property C.4, we obtain a gradient complexity of*

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon} \cdot \frac{dL}{C_3} j(\epsilon)\right) C_3^2 j(\epsilon)^4 L^2\right),$$

since $\alpha = C_3/\gamma$ *(Property C.4).*

The crux of this corollary is Property C.4. To prove this, we show like Li and Erdogdu (2020) that ν satisfies the Poincaré inequality with a constant independent of γ. Since it is not hard to show this around the global minimum, we can step by step extend the set where this inequality holds by a Lyapunov argument (Theorems D.1 and D.2). The essential difference between this analysis and that of Li and Erdogdu (2020) is that we do not work on compact manifolds anymore. Some rather minor difficulties emerge as we cannot employ the compactness but they can be addressed by supposing dissipativity which assures a quadratic growth for large x.

Remark 2. *These results do not definitively assert that SARAH-LD and SVRG-LD show the exact same performance in terms of optimization. Indeed, suppose we are close enough to the global optimum. Then, a big noise is not necessary anymore since it is more important to stably converge to the global minimum. Here, we should be able to significantly decrease the noise* ϵ_k*, and the bottleneck from the noise should disappear. In this case, SARAH-LD would perform better than SVRG-LD as we approach the original non-convex optimization setting where SARAH outperforms SVRG.*

Remark 3. *We also investigated an annealed version of SVRG-LD and SARAH-LD but could not ameliorate the gradient complexity. The detailed analysis can be found in Appendix E.*

5 Experiment

In this section, we illustrate our main result with a simple experiment.[1] We follow the same problem setting as that of Welling and Teh (2011) in Section 5.1. That is, we aim to sample from a non-log-concave posterior distribution $p(\theta|x) \propto p(\theta)\prod_{i=1}^{n} p(x_i|\theta)$ where $\{x_i\}_{i=1}^{n}$ is sampled from $p(x|\theta)$, a distribution parameterized by $\theta = (\theta_1, \theta_2)$. The prior $p(\theta)$ and the distribution of x parametrized by θ are respectively defined as $\theta_1 \sim N(0, 10)$, $\theta_2 \sim N(0, 1)$ and $x \sim 1/2N(\theta_1, 2) + 1/2N(\theta_1 + \theta_2, 2)$. Here, we set $n = 10000$, $\theta_1 = 0$ and $\theta_2 = 1$. Using the obtained 10000 samples, we simulated 1000 points of SVRG-LD with the inner loop length $m = n/B$ and different batch sizes B, namely, 100, 1000 and 10000 so that $B \geq m$ as required in Theorem 1. Evolution of KL-divergence between the true posterior, estimated by the Metropolis-adjusted Langevin algorithm, and that simulated by SVRG-LD is plotted in Figure 1. KL-divergence was approximated following Pérez-Cruz (2008).

As we can observe, Figure 1 correctly reproduces the theoretical bound of Theorem 1, with an exponential convergence in the beginning and a persistent bias due to the use of a discrete scheme and mini-batches. The fastest convergence in terms of gradient complexity under the condition $B \geq m$

[1] Source code can be found in https://github.com/yuri-k111/NeurIPS2022_code

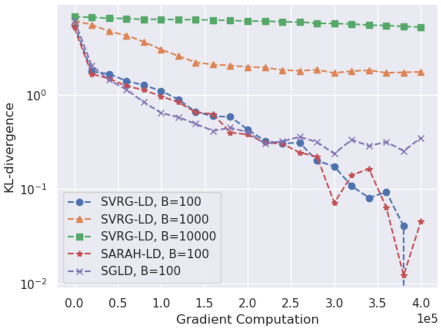

Figure 1: KL-divergence between the true and the simulated posterior. 1000 points were simulated for each algorithm with step size $\eta = 0.00001$. The inner loop length m for SVRG-LD was defined as n/B, and initial points were randomly drawn from the standard normal distribution. 1 gradient computation refers to one computation of ∇f_i.

is achieved by SVRG-LD with $B = \sqrt{n}$, which confirms our main theorem. Furthermore, with this best batch size, we also simulated 1000 points of SGLD and SARAH-LD as shown in Figure 1 as well. While SGLD and SVRG-LD have similar convergence speed in the beginning, the latter eventually achieves a higher precision thanks to the variance reduction method adopted in this scheme. SARAH-LD exhibits a similar performance as SVRG-LD, which agrees with Theorem 2.

6 Discussion and Conclusion

The main limitations of our work reside in the gap between practice and theory. Indeed, while our paper supposes assumptions quite standard in the literature of GLD, it cannot explain the whole empirical success that machine learning is currently experiencing. Some choices of parameters may also seem different than the practical use. However, compared to previous work, we succeeded in proving convergence of GLD with the popular stochastic gradient with relaxed conditions, and deleting the dependence of batch size and inner loop length on epsilon, which are all more realistic situations than prior work. The theoretical study in machine learning and deep learning precisely plays the role of filling as much as possible this large gap, and our work could be regarded as a further step forward to achieve this goal. Furthermore, in this paper, we focused on the pure sampling and optimization performance of the algorithms, and some of the drawbacks are simply due to this fact. For example, another limitation is that we did not investigate the generalization error in Section 4, but this was only outside the scope of this work.

In conclusion, we analysed the convergence rate of stochastic gradient Langevin Dynamics with variance reduction under smoothness and LSI and its application to optimization. In Section 3, we proved the convergence of SVRG-LD in terms of KL-divergence with more relaxed conditions (LSI and smoothness) and with a better gradient complexity than previous works. We also expanded SARAH to SARAH-LD and showed that this algorithm enjoyed the same advantages as SVRG-LD with only an improvement in the coefficients of the gradient complexity. These results led us to apply SVRG-LD and SARAH-LD to non-convex optimization in Section 4. We provided the global convergence and a non-asymptotic analysis of SVRG-LD and SARAH-LD. We obtained better conditions than prior works. Furthermore, we showed that under the additional assumption including Morse and Hessian Lipschitzness, the gradient complexity could be ameliorated, eliminating the exponential dependence on the inverse of the required error.

Acknowledgments and Disclosure of Funding

This study was partially supported by Japan Digital Design and JST CREST. We would like to thank Mufan Li and Murat A. Erdogdu for their valuable comments to fix some issues in the proof in Appendix D and their suggestion to improve it. We also thank anonymous reviewers for their feedback.

References

D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and M. Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. *arXiv preprint arXiv:2202.05214*, 2022.

V. S. Borkar and S. K. Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of Optimization Theory and Applications*, 100(3):499–513, 1999.

A. Bovier and F. Den Hollander. *Metastability: a Potential-Theoretic Approach*, volume 351. Springer, 2016.

P. Cattiaux, A. Guillin, and L.-M. Wu. A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. *Probability Theory and Related Fields*, 148(1):285–304, 2010.

N. Chatterji, N. Flammarion, Y. Ma, P. Bartlett, and M. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

P. Chen, J. Lu, and L. Xu. Approximation to stochastic variance reduced gradient Langevin dynamics by stochastic delay differential equations. *arXiv preprint arXiv:2106.04357*, 2021.

T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a.

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

K. A. Dubey, S. J Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing. Variance reduction in stochastic gradient Langevin dynamics. *Advances in Neural Information Processing Systems*, 29:1154–1162, 2016.

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

R. Holley and D. W. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. 1986.

Z. Huang and S. Becker. Stochastic gradient Langevin dynamics with variance reduction. In *2021 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2021.

C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182, 1980.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26:315–323, 2013.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

M. B. Li and M. A. Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *arXiv preprint arXiv:2010.11176v4*, 2020.

Z. Li. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. *Advances in Neural Information Processing Systems*, 32, 2019.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2): 185–232, 2002.

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

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017a.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

F. Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. *Advances in neural information processing systems*, 21, 2008.

N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in Neural Information Processing Systems*, 32:8094–8106, 2019.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

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

D. Zou, P. Xu, and Q. Gu. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 2936–2945. PMLR, 2019a.

D. Zou, P. Xu, and Q. Gu. Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction. *Advances in Neural Information Processing Systems*, 32:3835–3846, 2019b.

D. Zou, P. Xu, and Q. Gu. Faster convergence of stochastic gradient Langevin Dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pages 1152–1162. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All the assumptions are summarized in Subsection 2.3 and refered when used.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the footnote.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Theorem 1 and Corollary 1.1

In this Section, to clearly differentiate from SARAH-LD, we redefine the random variable generated at the k-th step of SVRG-LD (Algorithm 1) as Y_k and the stochastic gradient as $v_k^{(Y)}$. The distribution of Y_k is ρ_k.

A.1 Preparation for the Proof

We first prepare some lemmas.

Lemma A.1. *Under Assumption 1,*

$$\mathbb{E}_\nu[\|\nabla F\|^2] \leq dL/\gamma.$$

Proof. As $\mathrm{d}\nu = \mathrm{e}^{-\gamma F}\mathrm{d}x$, using integration by parts, we obtain

$$\mathbb{E}_\nu[\|\nabla(\gamma F)\|^2] = \mathbb{E}_\nu[\|\Delta(\gamma F)\|].$$

Now, since F is L-smooth by Assumption 1, $\nabla^2 F \preceq LI$ holds, which implies $\Delta F \leq dL$. As a result,

$$\mathbb{E}_\nu[\|\nabla F\|^2] = \frac{1}{\gamma}\mathbb{E}_\nu[\|\Delta F\|] \leq \frac{dL}{\gamma}.$$

$Q.E.D$

The relation between 2-Wasserstein distance and KL-divergence is given by the following inequality.

Lemma A.2. *Under Assumption 2, ν satisfies the following Talagrand's inequality with the same Log-Sobolev constant α:*

$$\frac{\alpha}{2}W_2(\rho,\nu)^2 \leq H_\nu(\rho). \tag{3}$$

Remark A.1. *See Theorem 1 of Otto and Villani (2000) for a proof of Lemma A.2.*

The following two lemmas that bound $\mathbb{E}[\|v_k^{(\mathrm{Y})}\|^2]$ and the variance of the stochastic gradient $v_k^{(\mathrm{Y})}$ with the KL-divergences $H_\nu(\rho_k), H_\nu(\rho_{k-1}), \ldots$ are crucial in our proof.

Lemma A.3. *Under Assumption 1, suppose Talagrand's inequality (3) holds for ν with a constant α, then for all $k = sm + r$, where $s \in \mathbb{N} \cup \{0\}$ and $r = 0, \ldots, m-1$, the following holds in the update of SVRG-LD:*

$$\mathbb{E}_{Y_k, I_k, Y^{(s)}}[\|v_k^{(\mathrm{Y})}\|^2] \leq \Lambda' H_\nu(\rho_{sm+r}) + T + \sum_{i=0}^{r-1} S(S+1)^{r-i-1}\left(\Lambda' H_\nu(\rho_{sm+i}) + T\right),$$

where $\Lambda = \left(1 + \frac{2(n-B)}{B(n-1)}\right)$, $\Xi = \frac{(n-B)}{B(n-1)}$,

$$\Lambda' = \frac{4L^2}{\alpha}\Lambda,$$

$$S = 4L^2 m\eta^2\Xi,$$

and

$$T = \frac{2dL}{\gamma}\Lambda + \frac{8\eta m d L^2}{\gamma}\Xi.$$

Proof. Let $v_i^{(1)}(Y_k) := \nabla f_i(Y_k) - \nabla f_i(Y^{(s)}) + \nabla F(Y^{(s)})$, then

$$\begin{aligned}
\mathbb{E}_{Y_k, I_k, Y^{(s)}}[\|v_k^{(\mathrm{Y})}\|^2] =& \mathbb{E}_{Y_k, I_k, Y^{(s)}}\left[\left\|\frac{1}{B}\sum_{i \in I_k} v_i^{(1)}(Y_k)\right\|^2\right] \\
=& \frac{1}{B^2}\mathbb{E}_{Y_k, I_k, Y^{(s)}}\left[\sum_{i \neq i', \{i, i'\} \in I_k} \left\langle v_i^{(1)}(Y_k), v_{i'}^{(1)}(Y_k)\right\rangle\right] \\
&+ \frac{1}{B^2}\mathbb{E}_{Y_k, I_k, Y^{(s)}}\left[\sum_{i \in I_k} \|v_i^{(1)}(Y_k)\|^2\right] \\
=& \frac{B-1}{Bn(n-1)}\mathbb{E}_{Y_k, Y^{(s)}}\left[\sum_{i \neq i'} \left\langle v_i^{(1)}(Y_k), v_{i'}^{(1)}(Y_k)\right\rangle\right] \\
&+ \frac{1}{B}\mathbb{E}_{Y_k, i, Y^{(s)}}\left[\|v_i^{(1)}(Y_k)\|^2\right] \\
&(i \text{ follows the uniform distribution under } \{1, \ldots, n\}) \\
=& \frac{B-1}{Bn(n-1)}\mathbb{E}_{Y_k, Y^{(s)}}\left[\sum_{i, i'} \left\langle v_i^{(1)}(Y_k), v_{i'}^{(1)}(Y_k)\right\rangle\right] \\
&- \frac{B-1}{B(n-1)}\mathbb{E}_{Y_k, i, Y^{(s)}}\left[\|v_i^{(1)}(Y_k)\|^2\right] \\
&+ \frac{1}{B}\mathbb{E}_{Y_k, i, Y^{(s)}}\left[\|v_i^{(1)}(Y_k)\|^2\right] \\
=& \frac{(B-1)n}{B(n-1)}\mathbb{E}_{Y_k}[\|\nabla F(Y_k)\|^2] + \frac{n-B}{B(n-1)}\mathbb{E}_{Y_k, i, Y^{(s)}}[\|v_i^{(1)}(Y_k)\|^2],
\end{aligned}$$

where we used $\frac{1}{n}\sum_{i=1}^n v_i^{(1)}(Y_k) = \nabla F(Y_k)$ for the last equality.

As a result, taking into account $\frac{(B-1)n}{B(n-1)} - 1 = \frac{B-n}{B(n-1)} \leq 0$,

$$\mathbb{E}_{Y_k, I_k, Y^{(s)}}[\|v_k^{(\mathrm{Y})}\|^2] = \mathbb{E}_{Y_k}[\|\nabla F(Y_k)\|^2] + \frac{n-B}{B(n-1)}\mathbb{E}_{Y_k, i, Y^{(s)}}[\|v_i^{(1)}(Y_k)\|^2]. \tag{4}$$

Choosing an optimal coupling $Y_k \sim \rho_k$ and $Y^* \sim \nu$ so that $\mathbb{E}[\|Y_k - Y^*\|^2] = W_2(\rho_k, \nu)^2$, we obtain

$$\begin{aligned}
\mathbb{E}_{Y_k}[\|\nabla F(Y_k)\|^2] &\leq 2\mathbb{E}_{Y_k, Y^*}[\|\nabla F(Y_k) - \nabla F(Y^*)\|^2] + 2\mathbb{E}_{Y^*}[\|\nabla F(Y^*)\|^2] \\
&\leq 2L^2\mathbb{E}[\|Y_k - Y^*\|^2] + 2dL/\gamma \\
&= 2L^2 W_2(\rho_k, \nu)^2 + 2dL/\gamma \\
&\leq \frac{4L^2}{\alpha} H_\nu(\rho_k) + 2dL/\gamma, \qquad (5)
\end{aligned}$$

where we used Lemma A.1 and the smoothness of F for the second inequality, the definition of W_2 for the equality and Talagrand's inequality (Lemma A.2) for the last inequality.

Moreover,

$$\begin{aligned}
\mathbb{E}_{Y_k,i,Y^{(s)}}[\|v_i^{(1)}(Y_k)\|^2] =& \mathbb{E}_{Y_k,i,Y^{(s)}}\left[\left\|\nabla f_i(Y_k) - \nabla f_i(Y^{(s)}) + \nabla F(Y^{(s)})\right\|^2\right] \\
\leq& 2\mathbb{E}\left[\left\|(\nabla f_i(Y_k) - \nabla f_i(Y^{(s)})) - \left(\nabla F(Y_k) - \nabla F(Y^{(s)})\right)\right\|^2\right] \\
&+ 2\mathbb{E}[\|\nabla F(Y_k)\|^2] \\
\leq& 2\mathbb{E}\left[\left\|\nabla f_i(Y_k) - \nabla f_i(Y^{(s)})\right\|^2\right] + 2\mathbb{E}[\|\nabla F(Y_k)\|^2] \\
\leq& 2L^2\mathbb{E}\left[\left\|Y_k - Y^{(s)}\right\|^2\right] + \frac{8L^2}{\alpha}H_\nu(\rho_k) + 4dL/\gamma \\
=& 2L^2\mathbb{E}\left[\left\|\sum_{i=1}^{r}(Y_{sm+i} - Y_{sm+i-1})\right\|^2\right] \\
&+ \frac{8L^2}{\alpha}H_\nu(\rho_k) + 4dL/\gamma \\
=& 2L^2\mathbb{E}\left[\left\|\sum_{i=1}^{r}\left(-\eta v_{sm+i-1}^{(\mathrm{Y})} + \sqrt{\frac{2\eta}{\gamma}}\epsilon_{sm+i-1}\right)\right\|^2\right] \\
&+ \frac{8L^2}{\alpha}H_\nu(\rho_k) + 4dL/\gamma \\
\leq& 4L^2\mathbb{E}\left[\left\|\sum_{i=1}^{r}\eta v_{sm+i-1}^{(\mathrm{Y})}\right\|^2\right] + 4L^2\mathbb{E}\left[\left\|\sum_{i=1}^{r}\left(\sqrt{\frac{2\eta}{\gamma}}\epsilon_{sm+i-1}\right)\right\|^2\right] \\
&+ \frac{8L^2}{\alpha}H_\nu(\rho_k) + 4dL/\gamma \\
\leq& 4r\eta^2L^2\sum_{i=1}^{r}\mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Y})}\|^2] + \frac{8\eta L^2}{\gamma}\sum_{i=1}^{r}\mathbb{E}[\|\epsilon_{sm+i-1}\|^2] \\
&+ \frac{8L^2}{\alpha}H_\nu(\rho_k) + 4dL/\gamma \\
\leq& 4m\eta^2L^2\sum_{i=1}^{r}\mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Y})}\|^2] + \frac{8\eta mL^2d}{\gamma} + \frac{8L^2}{\alpha}H_\nu(\rho_k) + 4dL/\gamma.
\end{aligned}$$

We used $\mathbb{E}[\|y - \mathbb{E}[y]\|^2] \leq \mathbb{E}[\|y\|^2]$ for the second inequality, smoothness of F and equation (5) for the third inequality and $r < m$ for the last inequality.

Plugging these to equation (4), we conclude

$$\begin{aligned}
\mathbb{E}_{Y_k,I_k,Y^{(s)}}[\|v_k^{(\mathrm{Y})}\|^2] \leq& \left(1 + \frac{2(n-B)}{B(n-1)}\right)\left(\frac{4L^2}{\alpha}H_\nu(\rho_k) + \frac{2dL}{\gamma}\right) \\
&+ \frac{(n-B)}{B(n-1)}\left(4m\eta^2L^2\sum_{i=1}^{r}\mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Y})}\|^2] + \frac{8\eta mL^2d}{\gamma}\right).
\end{aligned}$$

Therefore, setting

$$\Lambda' = \frac{4L^2}{\alpha}\left(1 + \frac{2(n-B)}{B(n-1)}\right),$$

$$S = 4L^2m\eta^2\frac{(n-B)}{B(n-1)},$$

and

$$T = \frac{2dL}{\gamma}\left(1 + \frac{2(n-B)}{B(n-1)}\right) + \frac{8\eta mdL^2}{\gamma}\frac{(n-B)}{B(n-1)},$$

we can rearrange this so that

$$\mathbb{E}_{Y_k,I_k,Y^{(s)}}[\|v_k^{(\mathrm{Y})}\|^2] \le \sum_{i=1}^{r} S\mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Y})}\|^2] + \Lambda' H_\nu(\rho_{sm+r}) + T. \tag{6}$$

Now, we are ready to prove by mathematical induction that the inequality of the statement holds for all $r = 0, \ldots, m-1$. When $r = 0$, the inequality holds from equation (5) as follows:

$$\begin{aligned}
\mathbb{E}_{Y_k,I_k,Y^{(s)}}[\|v_k^{(\mathrm{Y})}\|^2] &= \mathbb{E}_{Y_{sm}}[\|v_{sm}^{(\mathrm{Y})}\|^2] \\
&\le \mathbb{E}_{Y_{sm}}[\|\nabla F(Y_{sm})\|^2] + \frac{n-B}{B(n-1)}\mathbb{E}_{Y_{sm}}[\|v_i^{(1)}(Y_{sm})\|^2] \\
&= \left(1 + \frac{n-B}{B(n-1)}\right)\mathbb{E}_{Y_{sm}}[\|\nabla F(Y_{sm})\|^2] \\
&\le \left(1 + \frac{n-B}{B(n-1)}\right)\left(\frac{4L^2}{\alpha}H_\nu(\rho_{sm}) + \frac{2dL}{\gamma}\right) \\
&\le \Lambda' H_\nu(\rho_{sm}) + T,
\end{aligned}$$

where for the second equality we used $v_i^{(1)}(Y_{sm}) = \nabla F(Y_{sm})$.

Next, let us assume that the inequality of the lemma holds for $r \le l$. Then, from equation (6), we obtain

$$\begin{aligned}
\mathbb{E}[\|v_{sm+l+1}^{(\mathrm{Y})}\|^2] \le& \sum_{i=0}^{l} S\mathbb{E}[\|v_{sm+i}^{(\mathrm{Y})}\|^2] + \Lambda' H_\nu(\rho_{sm+l+1}) + T \\
\le& \sum_{i=0}^{l} S\left(\Lambda' H_\nu(\rho_{sm+i}) + T + \sum_{j=0}^{i-1} S(S+1)^{i-j-1}\left(\Lambda' H_\nu(\rho_{sm+j}) + T\right)\right) \\
&+ \Lambda' H_\nu(\rho_{sm+l+1}) + T \\
=& \sum_{i=0}^{l} S\left(\Lambda' H_\nu(\rho_{sm+i}) + T\right)\left(1 + \sum_{j=0}^{l-i-1} S(S+1)^j\right) \\
&+ \Lambda' H_\nu(\rho_{sm+l+1}) + T \\
=& \sum_{i=0}^{l} S\left(\Lambda' H_\nu(\rho_{sm+i}) + T\right)\left(1 + S\frac{(S+1)^{l-i} - 1}{(S+1) - 1}\right) \\
&+ \Lambda' H_\nu(\rho_{sm+l+1}) + T \\
=& \Lambda' H_\nu(\rho_{sm+l+1}) + T + \sum_{i=0}^{l} S(S+1)^{l+1-i-1}\left(\Lambda' H_\nu(\rho_{sm+i}) + T\right).
\end{aligned}$$

In the second inequality, we used the hypothesis of mathematical induction. This is equivalent to using Gronwall's lemma. This concludes the proof.

Q.E.D

Lemma A.4. *Under Assumption 1, for all $k = sm + r$, where $s \in \mathbb{N} \cup \{0\}$ and $r = 0, \ldots, m-1$, the following holds in the update of SVRG-LD:*

$$\mathbb{E}_{Y_k,I_k,Y^{(s)}}[\|v_k^{(\mathrm{Y})} - \nabla F(Y_k)\|^2] \le \frac{L^2(n-B)}{B(n-1)}\mathbb{E}_{Y_k,I_k,Y^{(s)}}[\|Y_k - Y^{(s)}\|^2].$$

Proof. Let $v_i^{(2)}(Y_k) = \nabla f_i(Y_k) - \nabla f_i(Y^{(s)}) + \nabla F(Y^{(s)}) - \nabla F(Y_k)$. Then,

$$
\begin{aligned}
\mathbb{E}_{Y_k, I_k, Y^{(s)}}[\|v_k^{(Y)} - \nabla F(Y_k)\|^2] =& \mathbb{E}_{Y_k, I_k, Y^{(s)}}\left[\left\|\frac{1}{B}\sum_{i \in I_k} v_i^{(2)}(Y_k)\right\|^2\right] \\
=& \frac{1}{B^2}\mathbb{E}_{Y_k, I_k, Y^{(s)}}\left[\sum_{i \neq i', \{i,i'\} \in I_k} \left\langle v_i^{(2)}(Y_k), v_{i'}^{(2)}(Y_k)\right\rangle\right] \\
&+ \frac{1}{B^2}\mathbb{E}_{Y_k, I_k, Y^{(s)}}\left[\sum_{i \in I_k} \|v_i^{(2)}(Y_k)\|^2\right] \\
=& \frac{B-1}{Bn(n-1)}\mathbb{E}_{Y_k, Y^{(s)}}\left[\sum_{i \neq i'} \left\langle v_i^{(2)}(Y_k), v_{i'}^{(2)}(Y_k)\right\rangle\right] \\
&+ \frac{1}{B}\mathbb{E}_{Y_k, i, Y^{(s)}}\left[\|v_i^{(2)}(Y_k)\|^2\right] \\
&(i \text{ follows the uniform distribution under } \{1, \ldots, n\}) \\
=& \frac{B-1}{Bn(n-1)}\mathbb{E}_{Y_k, Y^{(s)}}\left[\sum_{i, i'} \left\langle v_i^{(2)}(Y_k), v_{i'}^{(2)}(Y_k)\right\rangle\right] \\
&- \frac{B-1}{B(n-1)}\mathbb{E}_{Y_k, i, Y^{(s)}}\left[\|v_i^{(2)}(Y_k)\|^2\right] \\
&+ \frac{1}{B}\mathbb{E}_{Y_k, i, Y^{(s)}}\left[\|v_i^{(2)}(Y_k)\|^2\right] \\
=& \frac{n-B}{B(n-1)}\mathbb{E}_{Y_k, i, Y^{(s)}}[\|v_i^{(2)}(Y_k)\|^2].
\end{aligned}
$$

In the last equality, we used $\frac{1}{n}\sum_{i=1}^n v_i^{(2)}(Y_k) = 0$.

Now, since

$$
\begin{aligned}
\mathbb{E}_{Y_k, i, Y^{(s)}}[\|v_i^{(2)}(Y_k)\|^2] &= \mathbb{E}_{Y_k, i, Y^{(s)}}[\|\nabla f_i(Y_k) - \nabla f_i(Y^{(s)}) + \nabla F(Y^{(s)}) - \nabla F(Y_k)\|^2] \\
&= \mathbb{E}[\|\nabla f_i(Y_k) - \nabla f_i(Y^{(s)}) - \mathbb{E}[\nabla f_i(Y_k) - \nabla f_i(Y^{(s)})]\|^2] \\
&\leq \mathbb{E}[\|\nabla f_i(Y_k) - \nabla f_i(Y^{(s)})\|^2] \\
&\leq L^2\mathbb{E}[\|Y_k - Y^{(s)}\|^2],
\end{aligned}
$$

we obtain the desired result.

Q.E.D

A.2 Main Proof

We are now ready to prove the main results. The main idea of the following proofs is due to Vempala and Wibisono (2019). We first evaluate how $H_\nu(\rho_k)$ decreases compared with the previous steps.

Theorem A.1. *Under Assumptions 1 and 2, $0 < \eta < \frac{\alpha}{16\sqrt{6}L^2 m\gamma}$, $\gamma \geq 1$ and $B \geq m$, for all $k = sm + r$, where $s \in \mathbb{N} \cup \{0\}$ and $r = 0, \ldots, m-1$, the following holds in the update of SVRG-LD:*

$$
\begin{aligned}
H_\nu(\rho_{k+1}) \leq& \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1 + \frac{\alpha}{4\gamma}\eta\right)H_\nu(\rho_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta}H_\nu(\rho_{sm+i}) \\
&+ 8\eta^2 dL^2\Upsilon,
\end{aligned}
$$

where $\Lambda = \left(1 + \frac{2(n-B)}{B(n-1)}\right)$, $\Xi = \frac{(n-B)}{B(n-1)}$ and $\Upsilon = (\Lambda + \Xi + 1 + 2m\Xi)$.

Proof. Note that from Lemma A.2, Talagrand's inequality is satisfied with constant α.

One step of SVRG-LD can be formulated as follows:

$$Y_{sm+r+1} \leftarrow Y_{sm+r} - \eta v^{(\mathrm{Y})}_{sm+r} + \sqrt{2\eta/\gamma}\epsilon_{sm+r}.$$

This can be further interpreted as the output at time $t=\eta$ of the following SDE:

$$\mathrm{d}\tilde{Y}_t = -v^{(\mathrm{Y})}_{sm+r}\mathrm{d}t + \sqrt{2/\gamma}\mathrm{d}B_t,\ \tilde{Y}_0 = Y_{sm+r}. \tag{7}$$

In this context, the distribution $\tilde{\rho}_t$ of $\tilde{Y}_t$ depends on both Y_{sm+r} and

$$\beta^{(\mathrm{Y})}_{sm+r} := (I_{sm+r}, Y^{(s)}).$$

Let us define their joint distribution $\tilde{\rho}_{rt\beta^{(\mathrm{Y})}_{sm+r}}$ as follows:

$$\begin{aligned}\mathrm{d}\tilde{\rho}_{rt\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}, \tilde{Y}_t, \beta^{(\mathrm{Y})}_{sm+r}) &= \mathrm{d}\tilde{\rho}_{r\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})\mathrm{d}\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(\tilde{Y}_t|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})\\ &= \mathrm{d}\tilde{\rho}_{t\beta^{(\mathrm{Y})}_{sm+r}}(\tilde{Y}_t, \beta^{(\mathrm{Y})}_{sm+r})\mathrm{d}\tilde{\rho}_{r|t\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}|\tilde{Y}_t, \beta^{(\mathrm{Y})}_{sm+r}).\end{aligned}$$

Then, the Fokker-Planck equation (2) when Y_{sm+r} and $\beta^{(\mathrm{Y})}_{sm+r}$ are fixed becomes

$$\begin{aligned}\frac{\partial\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(\tilde{Y}_t|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})}{\partial t} &= \nabla\cdot(\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(\tilde{Y}_t|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})v^{(\mathrm{Y})}_{sm+r})\\ &\quad+\frac{1}{\gamma}\Delta\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(\tilde{Y}_t|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r}).\end{aligned} \tag{8}$$

Therefore, the following holds about the distribution $\tilde{\rho}_t$ of $\tilde{Y}_t$ governed by equation (7):

$$\begin{aligned}\frac{\partial\tilde{\rho}_t(y)}{\partial t} &= \int\frac{\partial\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(y|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})}{\partial t}\tilde{\rho}_{r\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})\mathrm{d}Y_{sm+r}\mathrm{d}\beta^{(\mathrm{Y})}_{sm+r}\\ &= \int\left(\nabla\cdot(\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(y|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})v^{(\mathrm{Y})}_{sm+r}) + \frac{1}{\gamma}\Delta\tilde{\rho}_{t|r\beta^{(\mathrm{Y})}_{sm+r}}(y|Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})\right)\\ &\qquad\cdot\tilde{\rho}_{r\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}, \beta^{(\mathrm{Y})}_{sm+r})\mathrm{d}Y_{sm+r}\mathrm{d}\beta^{(\mathrm{Y})}_{sm+r}\\ &= \int\nabla\cdot(\tilde{\rho}_{rt\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}, y, \beta^{(\mathrm{Y})}_{sm+r})v^{(\mathrm{Y})}_{sm+r})\mathrm{d}Y_{sm+r}\mathrm{d}\beta^{(\mathrm{Y})}_{sm+r}\\ &\quad+\int\frac{1}{\gamma}\Delta\tilde{\rho}_{rt\beta^{(\mathrm{Y})}_{sm+r}}(Y_{sm+r}, y, \beta^{(\mathrm{Y})}_{sm+r})\mathrm{d}Y_{sm+r}\mathrm{d}\beta^{(\mathrm{Y})}_{sm+r}\\ &=\nabla\cdot\left(\tilde{\rho}_t(y)\int\tilde{\rho}_{r\beta^{(\mathrm{Y})}_{sm+r}|t}v^{(\mathrm{Y})}_{sm+r}\mathrm{d}Y_{sm+r}\mathrm{d}\beta^{(\mathrm{Y})}_{sm+r}\right) + \frac{1}{\gamma}\Delta\tilde{\rho}_t(y)\\ &=\nabla\cdot\left(\tilde{\rho}_t(y)\mathbb{E}_{\tilde{\rho}_{r\beta^{(\mathrm{Y})}_{sm+r}|t}}[v^{(\mathrm{Y})}_{sm+r}|\tilde{Y}_t = y]\right) + \frac{1}{\gamma}\Delta\tilde{\rho}_t(y),\end{aligned}$$

where for the second equation we used equation (8).

Plugging this to

$$\frac{\mathrm{d}}{\mathrm{d}t}H_\nu(\tilde{\rho}_t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^n}\tilde{\rho}_t\log\frac{\tilde{\rho}_t}{\nu}\mathrm{d}y = \int_{\mathbb{R}^n}\frac{\partial\tilde{\rho}_t}{\partial t}\log\frac{\tilde{\rho}_t}{\nu}\mathrm{d}y,$$

we obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H_\nu(\tilde\rho_t) =& \int_{\mathbb{R}^d}\left(\nabla\cdot\left(\tilde\rho_t(y)\mathbb{E}_{\tilde\rho_{r\beta^{(\mathrm{Y})}_{sm+r}|t}}[v^{(\mathrm{Y})}_{sm+r}|\tilde Y_t = y]\right) + \frac{1}{\gamma}\Delta\tilde\rho_t(y)\right)\log\frac{\tilde\rho_t}{\nu}\mathrm{d}y \\
=& \int\left(\nabla\cdot\left(\tilde\rho_t(y)\left(\frac{1}{\gamma}\nabla\log\frac{\tilde\rho_t(y)}{\nu(y)} + \mathbb{E}_{\tilde\rho_{r\beta^{(\mathrm{Y})}_{sm+r}|t}}[v^{(\mathrm{Y})}_{sm+r}|\tilde Y_t = y] - \nabla F(y)\right)\right)\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot\log\frac{\tilde\rho_t(y)}{\nu(y)}\mathrm{d}y \\
=& -\int\tilde\rho_t(y)\left\langle\frac{1}{\gamma}\nabla\log\frac{\tilde\rho_t(y)}{\nu(y)} + \mathbb{E}_{\tilde\rho_{r\beta^{(\mathrm{Y})}_{sm+r}|t}}[v^{(\mathrm{Y})}_{sm+r}|\tilde Y_t = y] - \nabla F, \nabla\log\frac{\tilde\rho_t}{\nu}\right\rangle\mathrm{d}y \\
=& -\int\tilde\rho_t(y)\frac{1}{\gamma}\left\|\log\frac{\tilde\rho_t(y)}{\nu(y)}\right\|^2\mathrm{d}y \\
&+\int\tilde\rho_t(y)\left\langle\nabla F(y) - \mathbb{E}_{\tilde\rho_{r\beta^{(\mathrm{Y})}_{sm+r}|t}}[v^{(\mathrm{Y})}_{sm+r}|\tilde Y_t = y], \nabla\log\frac{\tilde\rho_t(y)}{\nu(y)}\right\rangle\mathrm{d}y \\
=& -\frac{1}{\gamma}J_\nu(\tilde\rho_t) \\
&+\int\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}\left\langle\nabla F - v^{(\mathrm{Y})}_{sm+r}, \nabla\log\frac{\tilde\rho_t}{\nu}\right\rangle\mathrm{d}Y_{sm+r}\mathrm{d}y\mathrm{d}\beta^{(\mathrm{Y})}_{sm+r} \\
=& -\frac{1}{\gamma}J_\nu(\tilde\rho_t) + \mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}\left[\left\langle\nabla F(\tilde Y_t) - v^{(\mathrm{Y})}_{sm+r}, \nabla\log\frac{\tilde\rho_t(\tilde Y_t)}{\nu(\tilde Y_t)}\right\rangle\right].
\end{aligned}
$$

Now, let us define the second term of the right-hand side of the very last equality as Ⓐ. Applying $\langle a, b\rangle \leq \gamma\|a\|^2 + \frac{1}{4\gamma}\|b\|^2$ to this, we obtain

$$
\begin{aligned}
\text{Ⓐ} \leq& \gamma\mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}\left[\|\nabla F(\tilde Y_t) - v^{(\mathrm{Y})}_{sm+r}\|^2\right] + \frac{1}{4\gamma}\mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}\left[\left\|\nabla\log\frac{\tilde\rho_t(\tilde Y_t)}{\nu(\tilde Y_t)}\right\|^2\right] \\
\leq& 2\gamma\mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}\left[\|\nabla F(\tilde Y_t) - \nabla F(Y_{sm+r})\|^2\right] \\
&+ 2\gamma\mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}\left[\|\nabla F(Y_{sm+r}) - v^{(\mathrm{Y})}_{sm+r}\|^2\right] \\
&+ \frac{1}{4\gamma}J_\nu(\tilde\rho_t) \\
\leq& 2\gamma L^2\mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}[\|\tilde Y_t - Y_{sm+r}\|^2] + \frac{2\gamma L^2(n-B)}{B(n-1)}\mathbb{E}_{\tilde\rho_{rt\beta^{(\mathrm{Y})}_{sm+r}}}\left[\|Y_{sm+r} - Y_{sm}\|^2\right] \\
&+ \frac{1}{4\gamma}J_\nu(\tilde\rho_t),
\end{aligned}
$$

where for the last inequality we used the smoothness of F and Lemma A.4.

As $\tilde Y_t = Y_{sm+r} - tv^{(\mathrm{Y})}_{sm+r} + \sqrt{2t/\gamma}\epsilon_{sm+r}$ ($\epsilon_{sm+r} \sim N(0, I)$), from Lemma A.3, we have

$$
\begin{aligned}
\mathbb{E}[\|\tilde Y_t - Y_{sm+r}\|^2] =& \mathbb{E}[\| - tv^{(\mathrm{Y})}_{sm+r} + \sqrt{2t/\gamma}\epsilon_{sm+r}\|^2] \\
=& t^2\mathbb{E}[\|v^{(\mathrm{Y})}_{sm+r}\|^2] + 2td/\gamma \\
\leq& t^2\left(\Lambda' H_\nu(\rho_{sm+r}) + T + \sum_{i=0}^{r-1}S(S+1)^{r-i-1}\left(\Lambda' H_\nu(\rho_{sm+i}) + T\right)\right) \\
&+ 2td/\gamma.
\end{aligned}
$$

Furthermore, by the proof of Lemma A.3 we know that the following holds:

$$\begin{aligned}\mathbb{E}\left[\|Y_{sm+r}-Y_{sm}\|^2\right] \leq & 2m\eta^2\sum_{i=0}^{r-1}\mathbb{E}[\|v^{(\mathrm{Y})}_{sm+i}\|^2]+4\eta md/\gamma \\ \leq & 2m\eta^2\sum_{i=0}^{r-1}(S+1)^{r-i-1}(\Lambda' H_\nu(\rho_{sm+i})+T)+4\eta md/\gamma.\end{aligned}$$

As a result, taking into account that we are only concerned about the time interval $0 \leq t \leq \eta$, applying $t \leq \eta$, we conclude

$$\begin{aligned}\textcircled{A} \leq & 2\gamma L^2\eta^2\left(\Lambda' H_\nu(\rho_{sm+r})+T+\sum_{i=0}^{r-1}S(S+1)^{r-i-1}\left(\Lambda' H_\nu(\rho_{sm+i})+T\right)\right) \\ & +4\eta dL^2+4\gamma L^2\eta^2 m\Xi\sum_{i=0}^{r-1}(S+1)^{r-i-1}(\Lambda' H_\nu(\rho_{sm+i})+T)+8\eta mdL^2\Xi \\ & +\frac{1}{4\gamma}J_\nu(\rho_t) \\ \leq & 2\gamma L^2\eta^2\Lambda' H_\nu(\rho_{sm+r})+\sum_{i=0}^{r-1}4\gamma L^2\eta^2(S+1)^{r-i}\Lambda' H_\nu(\rho_{sm+i}) \\ & +4\gamma L^2\eta^2\sum_{i=0}^{r}(S+1)^{r-i}T+4\eta dL^2(1+2m\Xi)+\frac{1}{4\gamma}J_\nu(\rho_t) \\ \leq & 2\gamma L^2\eta^2\Lambda' H_\nu(\rho_{sm+r})+\sum_{i=0}^{r-1}4\gamma L^2\eta^2(S+1)^{r}\Lambda' H_\nu(\rho_{sm+i}) \\ & +4\gamma L^2\eta^2\sum_{i=0}^{r}(S+1)^{r}T+4\eta dL^2(1+2m\Xi)+\frac{1}{4\gamma}J_\nu(\rho_t) \\ \leq & 2\gamma L^2\eta^2\Lambda' H_\nu(\rho_{sm+r})+\sum_{i=0}^{r-1}4\gamma L^2\eta^2(S+1)^{m}\Lambda' H_\nu(\rho_{sm+i}) \\ & +4\gamma L^2\eta^2 m(S+1)^{m}T+4\eta dL^2(1+2m\Xi)+\frac{1}{4\gamma}J_\nu(\rho_t),\end{aligned}$$

where for the second inequality we used $m\Xi \leq 1$ and for the last inequality $r < m$. Here, as $\Xi \leq 1$ and $\eta \leq \frac{1}{4mL}$ by $\alpha \leq \gamma L$,

$$(S+1)^m \leq \mathrm{e}^{Sm}=\mathrm{e}^{4L^2m^2\eta^2\Xi} \leq \mathrm{e}^{1/4} \leq 2.$$

Therefore,

$$
\begin{aligned}
\textcircled{A} &\le 2\gamma L^2\eta^2\Lambda' H_\nu(\rho_{sm+r}) + \sum_{i=0}^{r-1} 8\gamma L^2\eta^2\Lambda' H_\nu(\rho_{sm+i}) \\
&\quad + 8\gamma L^2\eta^2 mT + 4\eta dL^2(1+2m\Xi) + \frac{1}{4\gamma}J_\nu(\rho_t) \\
&= \frac{8\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+r}) + \sum_{i=0}^{r-1}\frac{32\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+i}) \\
&\quad + 8\gamma L^2\eta^2 m\left(\frac{2dL}{\gamma}\Lambda + \frac{8\eta m dL^2}{\gamma}\Xi\right) + 4\eta dL^2(1+2m\Xi) + \frac{1}{4\gamma}J_\nu(\rho_t) \\
&= \frac{8\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+r}) + \sum_{i=0}^{r-1}\frac{32\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+i}) \\
&\quad + 4\eta dL^2\left(4\eta mL\Lambda + 16\eta^2 m^2L^2\Xi + 1 + 2m\Xi\right) + \frac{1}{4\gamma}J_\nu(\rho_t) \\
&\le \frac{8\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+r}) + \sum_{i=0}^{r-1}\frac{32\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+i}) \\
&\quad + 4\eta dL^2\left(\Lambda + \Xi + 1 + 2m\Xi\right) + \frac{1}{4\gamma}J_\nu(\rho_t),
\end{aligned}
$$

where for the first equality, we used $\Lambda' = \frac{4L^2}{\alpha}\Lambda$ and $T = \left(\frac{2dL}{\gamma}\Lambda + \frac{8\eta m dL^2}{\gamma}\Xi\right)$, and for the last inequality $\eta \le \frac{1}{4mL}$. Thus, setting $\Upsilon = \Lambda + \Xi + 1 + 2m\Xi$, we obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H_\nu(\tilde\rho_t) \le&\ -\frac{3}{4\gamma}J_\nu(\tilde\rho_t) + \frac{8\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+r}) + \sum_{i=0}^{r-1}\frac{32\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+i}) \\
&+ 4\eta dL^2\Upsilon.
\end{aligned}
$$

According to Assumption 2,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H_\nu(\tilde\rho_t) \le&\ -\frac{3\alpha}{2\gamma}H_\nu(\tilde\rho_t) + \frac{8\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+r}) + \sum_{i=0}^{r-1}\frac{32\gamma L^4\eta^2}{\alpha}\Lambda H_\nu(\rho_{sm+i}) \\
&+ 4\eta dL^2\Upsilon.
\end{aligned}
$$

Grouping the second to fourth terms as $U^{(\Upsilon)}_{sm+r}$ and multiplying both sides by $\mathrm{e}^{\frac{3\alpha}{2\gamma}t}$, we can write the above equation as

$$
\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{\frac{3\alpha}{2\gamma}t}H_\nu(\tilde\rho_t)\right) \le \mathrm{e}^{\frac{3\alpha}{2\gamma}t}U^{(\Upsilon)}_{sm+r}.
$$

Integrating both sides from $t = 0$ to $t = \eta$ and using $\tilde\rho_\eta = \rho_{sm+r+1}$, we obtain

$$
\begin{aligned}
\mathrm{e}^{\frac{3\alpha}{2\gamma}\eta}H_\nu(\rho_{sm+r+1}) - H_\nu(\rho_{sm+r}) &\le \frac{2\gamma(\mathrm{e}^{\frac{3\alpha}{2\gamma}\eta}-1)}{3\alpha}U^{(\Upsilon)}_{sm+r} \\
&\le 2\eta U^{(\Upsilon)}_{sm+r}.
\end{aligned}
$$

Here, for the last inequality, we used that $\mathrm{e}^c \le 1 + 2c$ $(0 < c = \frac{3\alpha}{2\gamma}\eta \le 1)$ holds since $0 < \eta \le \frac{\alpha}{16\sqrt{6}L^2m^\gamma} \le \frac{2\gamma}{3\alpha}$, where we used $1/L \le \gamma/\alpha$ and $m \ge 1$. Rearranging this, we obtain

$$
\begin{aligned}
H_\nu(\rho_{sm+r+1}) \le&\ \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1 + \frac{16\gamma L^4\eta^3}{\alpha}\Lambda\right)H_\nu(\rho_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{64\gamma L^4\eta^3}{\alpha}\Lambda H_\nu(\rho_{sm+i}) \\
&+ \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}8\eta^2dL^2\Upsilon. \qquad (9)
\end{aligned}
$$

Furthermore, since $\eta \le \frac{\alpha}{16\sqrt{6}mL^2\gamma} \le \frac{\alpha}{8\sqrt{3}L^2\gamma}$, $\mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta} \le 1$ and $\Lambda \le 3$

$$H_\nu(\rho_{sm+r+1}) \le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)H_\nu(\rho_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{8\gamma m}\eta H_\nu(\rho_{sm+i}) + 8\eta^2 dL^2\Upsilon.$$

On the other hand, since $\eta \le \frac{\alpha}{8mL^2\gamma}$ and $\alpha \le \gamma L$ holds,

$$\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta} \ge \mathrm{e}^{-\frac{\alpha m}{\gamma}\cdot\frac{\alpha}{8mL^2\gamma}} = \mathrm{e}^{-\frac{\alpha^2}{8L^2\gamma^2}} \ge \mathrm{e}^{-1/8} \ge 0.88 \ge \frac{1}{2},$$

which further implies

$$\begin{aligned}H_\nu(\rho_{sm+r+1}) \le{}& \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)H_\nu(\rho_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta}H_\nu(\rho_{sm+i})\\ &+ 8\eta^2 dL^2\Upsilon.\end{aligned}$$

Q.E.D

Finally, let us prove Theorem 1 and Corollary 1.1.

Theorem A.2 (Theorem 1 restated)**.** *Under Assumptions 1 and 2,* $0 < \eta < \frac{\alpha}{16\sqrt{6}L^2m\gamma}$*,* $\gamma \ge 1$ *and* $B \ge m$*, for all* $k \ge 1$*, the following holds in the update of SVRG-LD:*

$$H_\nu(\rho_k) \le \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k}H_\nu(\rho_0) + \frac{224\eta\gamma dL^2}{3\alpha}\Upsilon,$$

where $\Xi = \frac{(n-B)}{B(n-1)}$ *and* $\Upsilon = (\Lambda + \Xi + 1 + 2m\Xi)$*.*

Proof. Let us first prove by mathematical induction that the following inequality holds for all $k = 1, 2\ldots$:

$$H_\nu(\rho_k) \le \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k}H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}. \qquad \ldots\ (*)$$

(I) When $k = 1$, from Theorem A.1, since $Y^{(s)} = Y_0$,

$$\begin{aligned}H_\nu(\rho_1) &\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)H_\nu(\rho_0) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta}H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\\ &\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta+\frac{\alpha}{4m\gamma}\eta\right)H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\\ &\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{2\gamma}\eta\right)H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\\ &\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\mathrm{e}^{\frac{\alpha}{2\gamma}\eta}H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\\ &= \mathrm{e}^{-\frac{\alpha}{\gamma}\eta}H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\\ &\le \mathrm{e}^{-\frac{\alpha}{\gamma}\eta}H_\nu(\rho_0) + 8\eta^2 dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}.\end{aligned}$$

Here, for the second and last inequality, we used $\mathrm{e}^{-\frac{\alpha m\eta}{\gamma}} \le \mathrm{e}^{-\frac{\alpha\eta}{\gamma}} \le 1$. Thus, $(*)$ holds for $k = 1$.
(II) Now, let us assume that $(*)$ holds for all $k \le l$. Letting r and s the remainder and quotient of the Euclidian division of l by m respectively, when $k = l+1$ we obtain from Theorem A.1,

$$\begin{aligned}H_\nu(\rho_{sm+r+1}) \le{}& \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)H_\nu(\rho_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta}H_\nu(\rho_{sm+i})\\ &+ 8\eta^2 dL^2\Upsilon.\end{aligned}$$

From the hypothesis of mathematical induction,

$$\begin{aligned}
&H_\nu(\rho_{sm+r+1})\\
&\qquad \le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\qquad\quad + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta}\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+i)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\qquad\quad + 8\eta^2dL^2\Upsilon.
\end{aligned}$$

Since $\mathrm{e}^{-\frac{\alpha}{\gamma}\eta m} \le \mathrm{e}^{-\frac{\alpha}{\gamma}\eta r} \le \mathrm{e}^{-\frac{\alpha}{\gamma}\eta(r-i)}$ when $0 \le i < r < m$,

$$\begin{aligned}
H_\nu(\rho_{sm+r+1}) &\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\quad + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(r-i)}\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+i)}H_\nu(\rho_0)\right)\\
&\quad + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(r-i)}\left(8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\quad + 8\eta^2dL^2\Upsilon\\
&\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\quad + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\quad + 8\eta^2dL^2\Upsilon\\
&\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{2\gamma}\eta\right)\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\quad + 8\eta^2dL^2\Upsilon\\
&\le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\mathrm{e}^{\frac{\alpha}{2\gamma}\eta}\left(\mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)\\
&\quad + 8\eta^2dL^2\Upsilon\\
&= \mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r+1)}H_\nu(\rho_0)+\left(1+\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}\right)8\eta^2dL^2\Upsilon\\
&= \mathrm{e}^{-\frac{\alpha\eta}{\gamma}(sm+r+1)}H_\nu(\rho_0)+8\eta^2dL^2\Upsilon\cdot\left(1-\mathrm{e}^{-\frac{\alpha\eta}{\gamma}}\right)^{-1}.
\end{aligned}$$

Therefore, $(*)$ holds for all $k \ge 1$.

Now, using the inequality $1-\mathrm{e}^{-c} \ge \frac{3}{4}c$ for $0 < c = \frac{\alpha\eta}{\gamma} \le \frac{1}{4}$ (since $y = 1-\mathrm{e}^{-x}$ and $y = \frac{3}{4}x$ are both concave increasing functions intersecting at $x = 0$ and $1-\mathrm{e}^{-1/4} \ge \frac{3}{4}\times\frac{1}{4}$), which holds here because $\eta \le \frac{\alpha}{16\sqrt{6}L^2\gamma} \le \frac{\gamma}{4\alpha}$ since $1/L \le \gamma/\alpha$ and $m \ge 1$, we conclude

$$\begin{aligned}
H_\nu(\rho_k) &\le \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k}H_\nu(\rho_0)+\frac{32\eta\gamma dL^2}{3\alpha}\Upsilon\\
&\le \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k}H_\nu(\rho_0)+\frac{224\eta\gamma dL^2}{3\alpha},
\end{aligned}$$

which is the desired result. Here, for the last inequality, we used $\Upsilon = \Lambda + \Xi + 1 + 2m\Xi \le 3 + 1 + 1 + 2 = 7$.

Q.E.D

Corollary A.2.1 (Corollary 1.1 restated). *Under the same assumptions as Theorem A.2, for all $\epsilon \ge 0$, if we choose step size η such that*

$$\eta \le \frac{3\alpha\epsilon}{448\gamma dL^2},$$

then a precision $H_\nu(\rho_k) \le \epsilon$ is reached after

$$k \ge \frac{\gamma}{\alpha\eta} \log \frac{2H_\nu(\rho_0)}{\epsilon}$$

steps. Especially, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \frac{\alpha}{16\sqrt{6}L^2\sqrt{n}\gamma} \wedge \frac{3\alpha\epsilon}{448dL^2\gamma}$, then the gradient complexity becomes

$$\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right)\frac{\gamma^2L^2}{\alpha^2}\right).$$

Proof. Let $\epsilon > 0$. Then, by additionally requiring

$$\eta \le \frac{3\alpha\epsilon}{448\gamma dL^2},$$

we obtain

$$H_\nu(\rho_k) \le \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\rho_0) + \frac{\epsilon}{2}.$$

Thus, $H_\nu(\rho_k) \le \epsilon$ can be reached for

$$k \ge \frac{\gamma}{\alpha\eta} \log \frac{2H_\nu(\rho_0)}{\epsilon}.$$

As a result, if $0 < \epsilon \le \frac{28d}{3\sqrt{6}m}$ and we select the largest permissible step size, the gradient complexity becomes

$$O\left(k \cdot B + \frac{k}{m} \cdot n\right) = \tilde{O}\left(\left(\frac{B + n/m}{\epsilon}\right)\frac{d\gamma^2L^2}{\alpha^2}\right),$$

and the optimal complexity is

$$\tilde{O}\left(\frac{dn^{1/2}\gamma^2L^2}{\epsilon\alpha^2}\right)$$

with $B = \sqrt{n}$ and $m = \sqrt{n}$.

On the other hand, if $\epsilon \ge \frac{28d}{3\sqrt{6}m}$ and we select the largest permissible step size, the gradient complexity becomes

$$O\left(k \cdot B + \frac{k}{m} \cdot n\right) = \tilde{O}\left((mB + n)\frac{\gamma^2L^2}{\alpha^2}\right),$$

and the optimal complexity is

$$\tilde{O}\left(\frac{n\gamma^2L^2}{\alpha^2}\right)$$

with $B = \sqrt{n}$ and $m = \sqrt{n}$

Therefore, for all $\epsilon \ge 0$, the gradient complexity is

$$\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right)\frac{\gamma^2L^2}{\alpha^2}\right).$$

Q.E.D

B Proof of Theorem 2 and Corollary 2.1

In this Section, to clearly differentiate from SVRG-LD, we redefine the random variable generated at the k-th step of SARAH-LD (Algorithm 1) as Z_k and the stochastic gradient as $v_k^{(\mathrm{Z})}$. The distribution of Z_k is ϕ_k.

B.1 Preparation for the Proof

Let us first provide an upper bound of $\mathbb{E}[\|v_k^{(\mathrm{Z})}\|^2]$ and the variance of the stochastic gradient $v_k^{(\mathrm{Z})}$ using the KL-divergences $H_\nu(\phi_k), H_\nu(\phi_{k-1}), \ldots$.

Lemma B.1. *Under Assumption 1, for all $k = sm + r$, where $s \in \mathbb{N} \cup \{0\}$ and $r = 0, \ldots, m-1$, the following holds in the update of SARAH-LD:*

$$\mathbb{E}[\|\nabla F(Z_k) - v_k^{(\mathrm{Z})}\|^2] = \sum_{i=1}^{r} \mathbb{E}[\|v_{sm+i}^{(\mathrm{Z})} - v_{sm+i-1}^{(\mathrm{Z})}\|^2] - \sum_{i=1}^{r} \mathbb{E}[\|\nabla F(Z_{sm+i}) - \nabla F(Z_{sm+i-1})\|^2].$$

Proof. Let us define

$$\mathcal{F}_r = \sigma\left(Z^{(s)}, \epsilon_{sm}, I_{sm+1}, \epsilon_{sm+1}, I_{sm+2}, \epsilon_{sm+2}, \ldots, I_{sm+r-1}, \epsilon_{sm+r-1}\right),$$

which is the σ-$algebra$ generated by

$$Z^{(s)}, \epsilon_{sm}, I_{sm+1}, \epsilon_{sm+1}, I_{sm+2}, \epsilon_{sm+2}, \ldots, I_{sm+r-1}, \text{ and } \epsilon_{sm+r-1}.$$

When $r = 0$, the statement clearly holds. In the remainder of the proof, we assume $r \geq 1$. Then,

$$\begin{aligned}
\mathbb{E}[\|\nabla F(Z_k) - v_k^{(\mathrm{Z})}\|^2 \mid \mathcal{F}_r] &= \mathbb{E}[\|\nabla F(Z_{k-1}) - v_{k-1}^{(\mathrm{Z})} + \nabla F(Z_k) - \nabla F(Z_{k-1}) \\
&\qquad\qquad - (v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})})\|^2 \mid \mathcal{F}_r] \\
&= \|\nabla F(Z_{k-1}) - v_{k-1}^{(\mathrm{Z})}\|^2 + \|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 \\
&\quad + \mathbb{E}[\|v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})}\|^2 \mid \mathcal{F}_r] \\
&\quad + 2\left\langle \nabla F(Z_{k-1}) - v_{k-1}^{(\mathrm{Z})}, \nabla F(Z_k) - \nabla F(Z_{k-1}) \right\rangle \\
&\quad - 2\left\langle \nabla F(Z_{k-1}) - v_{k-1}^{(\mathrm{Z})}, \mathbb{E}[v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})} \mid \mathcal{F}_r] \right\rangle \\
&\quad - 2\left\langle \nabla F(Z_k) - \nabla F(Z_{k-1}), \mathbb{E}[v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})} \mid \mathcal{F}_r] \right\rangle \\
&= \|\nabla F(Z_{k-1}) - v_{k-1}^{(\mathrm{Z})}\|^2 - \|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 \\
&\quad + \mathbb{E}[\|v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})}\|^2 \mid \mathcal{F}_r].
\end{aligned}$$

Here in the last equality, we used that the following holds:

$$\begin{aligned}
\mathbb{E}[v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})} \mid \mathcal{F}_r] =& \mathbb{E}\left[\frac{1}{B}\sum_{i \in I_k} \nabla f_i(Z_k) - \nabla f_i(Z_{k-1}) \mid \mathcal{F}_r\right] \\
=& \nabla F(Z_k) - \nabla F(Z_{k-1}).
\end{aligned}$$

Taking expectation, we obtain

$$\begin{aligned}
\mathbb{E}[\|\nabla F(Z_k) - v_k^{(\mathrm{Z})}\|^2] = & \mathbb{E}[\|\nabla F(Z_{k-1}) - v_{k-1}^{(\mathrm{Z})}\|^2] - \mathbb{E}[\|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2] \\
& + \mathbb{E}[\|v_k^{(\mathrm{Z})} - v_{k-1}^{(\mathrm{Z})}\|^2].
\end{aligned}$$

Since this equation holds for all $k = sm + r$ ($r = 1, \ldots m-1$), recalling that

$$\mathbb{E}[\|\nabla F(Z_{sm}) - v_{sm}^{(\mathrm{Z})}\|^2] = 0,$$

and recursively applying this, we conclude that

$$\mathbb{E}[\|\nabla F(Z_k) - v_k^{(\mathrm{Z})}\|^2] = \sum_{i=1}^{r} \mathbb{E}[\|v_{sm+i}^{(\mathrm{Z})} - v_{sm+i-1}^{(\mathrm{Z})}\|^2] - \sum_{i=1}^{r} \mathbb{E}[\|\nabla F(Z_{sm+i}) - \nabla F(Z_{sm+i-1})\|^2].$$

$Q.E.D$

Lemma B.2. *Under Assumption 1, for all $k = sm + r$, where $s \in \mathbb{N} \cup \{0\}$ and $r = 0, \ldots, m-1$, the following holds in the update of SARAH-LD:*

$$\mathbb{E}[\|\nabla F(Z_k) - v_k^{(\mathrm{Z})}\|^2] \leq \sum_{i=1}^{r} \Xi L^2 \eta^2 \mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Z})}\|^2] + \frac{2\eta m d L^2}{\gamma}\Xi,$$

where $\Xi = \frac{n-B}{B(n-1)}$.

Proof. When $r = 0$, the statement clearly holds. In the remainder of the proof, we assume $r \geq 1$. Since $v_k^{(\mathrm{z})} - v_{k-1}^{(\mathrm{z})} = \frac{1}{B}\sum_{j\in I_k}(\nabla f_j(Z_k) - \nabla f_j(Z_{k-1}))$, defining

$$w_j := \nabla f_j(Z_k) - \nabla f_j(Z_{k-1}),$$

we obtain

$$\begin{aligned}
\mathbb{E}[\|v_k^{(\mathrm{z})} - v_{k-1}^{(\mathrm{z})}\|^2 \mid \mathcal{F}_k] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{j\in I_k} w_j\right\|^2 \mid \mathcal{F}_k\right] \\
&= \frac{1}{B^2}\mathbb{E}\left[\sum_{j\neq j', \{j,j'\}\in I_k} \langle w_j, w_{j'}\rangle \mid \mathcal{F}_k\right] + \frac{1}{B^2}\mathbb{E}\left[\sum_{j\in I_k}\|w_j\|^2 \mid \mathcal{F}_k\right] \\
&= \frac{B-1}{Bn(n-1)}\mathbb{E}\left[\sum_{j\neq j'}\langle w_j, w_{j'}\rangle \mid \mathcal{F}_k\right] + \frac{1}{B}\mathbb{E}\left[\|w_j\|^2 \mid \mathcal{F}_k\right] \\
&\quad (j \text{ follows a uniform distribution under } \{1,\ldots,n\}) \\
&= \frac{B-1}{Bn(n-1)}\mathbb{E}\left[\sum_{j,j'}\langle w_j, w_{j'}\rangle \mid \mathcal{F}_k\right] - \frac{B-1}{B(n-1)}\mathbb{E}\left[\|w_j\|^2 \mid \mathcal{F}_k\right] \\
&\quad + \frac{1}{B}\mathbb{E}\left[\|w_j\|^2 \mid \mathcal{F}_k\right] \\
&= \frac{(B-1)n}{B(n-1)}\mathbb{E}[\|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 \mid \mathcal{F}_k] \\
&\quad + \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_j(Z_k) - \nabla f_j(Z_{k-1})\|^2 \mid \mathcal{F}_k] \\
&\leq \mathbb{E}[\|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 | \mathcal{F}_k] \\
&\quad + \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_j(Z_k) - \nabla f_j(Z_{k-1})\|^2 \mid \mathcal{F}_k],
\end{aligned}$$

where for the fifth equation we used $\frac{1}{n}\sum_{j=1}^n w_j = \nabla F(Z_k) - \nabla F(Z_{k-1})$ and for the inequality, $\frac{(B-1)n}{B(n-1)} \leq 1$.

As a result,

$$\begin{aligned}
\mathbb{E}[\|v_k^{(\mathrm{z})} - v_{k-1}^{(\mathrm{z})}\|^2 \mid \mathcal{F}_k] &\leq \|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 \\
&\quad + \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_j(Z_k) - \nabla f_j(Z_{k-1})\|^2 \mid \mathcal{F}_k] \\
&\leq \|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 + L^2\Xi\mathbb{E}[\|Z_k - Z_{k-1}\|^2 \mid \mathcal{F}_k] \\
&= \|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2 + L^2\Xi\left\|-\eta v_{k-1}^{(\mathrm{z})} + \sqrt{\frac{2\eta}{\gamma}}\epsilon_{k-1}\right\|^2.
\end{aligned}$$

Taking expectation, we obtain

$$\begin{aligned}
\mathbb{E}[\|v_k^{(\mathrm{z})} - v_{k-1}^{(\mathrm{z})}\|^2] - \mathbb{E}[\|\nabla F(Z_k) - \nabla F(Z_{k-1})\|^2] &\leq L^2\Xi\mathbb{E}\left[\left\|-\eta v_{k-1}^{(\mathrm{z})} + \sqrt{\frac{2\eta}{\gamma}}\epsilon_{k-1}\right\|^2\right] \\
&= L^2\Xi\left(\eta^2\mathbb{E}[\|v_{k-1}^{(\mathrm{z})}\|^2] + \frac{2\eta d}{\gamma}\right).
\end{aligned}$$

Since this equation holds for all $k = sm + r$ $(r = 1, \dots m-1)$, from Lemma B.1,

$$\begin{aligned}
\mathbb{E}[\|\nabla F(Z_k) - v_k^{(\mathrm{Z})}\|^2] \leq & \sum_{i=1}^{r} \mathbb{E}[\|v_{sm+i}^{(\mathrm{Z})} - v_{sm+i-1}^{(\mathrm{Z})}\|^2] \\
& - \sum_{i=1}^{r} \mathbb{E}[\|\nabla F(Z_{sm+i}) - \nabla F(Z_{sm+i-1})\|^2] \\
& \leq \Xi L^2 \eta^2 \sum_{i=1}^{r} \mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Z})}\|^2] + \frac{2\eta r d L^2}{\gamma} \Xi \\
& \leq \Xi L^2 \eta^2 \sum_{i=1}^{r} \mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Z})}\|^2] + \frac{2\eta m d L^2}{\gamma} \Xi.
\end{aligned}$$

Q.E.D

Lemma B.3. *Under Assumption 1, suppose Talagrand's inequality holds for ν with a constant α, then for all $k = sm + r$, where $s \in \mathbb{N} \cup \{0\}$ and $r = 0, \dots, m-1$, the following holds in the update of SARAH-LD:*

$$\mathbb{E}[\|v_k^{(\mathrm{Z})}\|^2] \leq \frac{8L^2}{\alpha} H_\nu(\phi_{sm+r}) + P + \sum_{i=0}^{r-1} Q(Q+1)^{r-i-1} \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P \right),$$

where

$$\Xi = \frac{(n-B)}{B(n-1)},$$

$$P = \frac{4dL}{\gamma} + \frac{4\eta m d L^2}{\gamma} \Xi,$$

and

$$Q = 2\Xi L^2 \eta^2.$$

Proof. First, from Lemma B.2, we have

$$\begin{aligned}
\mathbb{E}[\|v_k^{(\mathrm{Z})}\|^2] &\leq 2\mathbb{E}[\|v_k^{(\mathrm{Z})} - \nabla F(Z_k)\|^2] + 2\mathbb{E}[\|\nabla F(Z_k)\|^2] \\
&\leq 2 \left(\sum_{i=1}^{r} \Xi L^2 \eta^2 \mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Z})}\|^2] + \frac{2\eta m d L^2}{\gamma} \Xi \right) + 2\mathbb{E}[\|\nabla F(Z_k)\|^2].
\end{aligned}$$

Choosing an optimal coupling $Z_k \sim \phi_k$ and $Z^* \sim \nu$ so that $\mathbb{E}[\|Z_k - Z^*\|^2] = W_2(\phi_k, \nu)^2$, we obtain

$$\begin{aligned}
\mathbb{E}_{Z_k}[\|\nabla F(Z_k)\|^2] &\leq 2\mathbb{E}_{Z_k, Z^*}[\|\nabla F(Z_k) - \nabla F(Z^*)\|^2] + 2\mathbb{E}_{Z^*}[\|\nabla F(Z^*)\|^2] \\
&\leq 2L^2 \mathbb{E}[\|Z_k - Z^*\|^2] + 2dL/\gamma \\
&= 2L^2 W_2(\phi_k, \nu)^2 + 2dL/\gamma \\
&\leq \frac{4L^2}{\alpha} H_\nu(\phi_k) + 2dL/\gamma, \qquad (10)
\end{aligned}$$

where, we used the smoothness of F and Lemma A.1 for the second inequality, the definition of W_2 for the equality and Talagrand's inequality for the last inequality.

As a result,

$$\begin{aligned}
\mathbb{E}[\|v_k^{(\mathrm{Z})}\|^2] \leq & 2 \left(\sum_{i=1}^{r} \Xi L^2 \eta^2 \mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Z})}\|^2] + \frac{2\eta m d L^2}{\gamma} \Xi \right) + \frac{8L^2}{\alpha} H_\nu(\phi_k) + 4dL/\gamma \\
= & \sum_{i=1}^{r} Q \mathbb{E}[\|v_{sm+i-1}^{(\mathrm{Z})}\|^2] + \frac{8L^2}{\alpha} H_\nu(\phi_k) + P. \qquad (11)
\end{aligned}$$

Here, we set

$$P = \frac{4dL}{\gamma} + \frac{4\eta m d L^2}{\gamma}\Xi,$$

and

$$Q = 2\Xi L^2 \eta^2.$$

Now, let us prove by mathematical induction that the inequality of the statement holds for all $r = 0, \dots, m-1$. When $r = 0$, the inequality holds from equation (10) as follows:

$$\begin{aligned}\mathbb{E}[\|v_{sm}^{(\mathrm{Z})}\|^2] &= \mathbb{E}[\|\nabla F(Z_{sm})\|^2] \\ &\le \frac{4L^2}{\alpha} H_\nu(\phi_{sm}) + 2dL/\gamma \\ &\le \frac{8L^2}{\alpha} H_\nu(\phi_{sm}) + P.\end{aligned}$$

Next, let us assume that the inequality of the lemma holds for $r \le l$. Then, from equation (11), we obtain

$$\begin{aligned}&\mathbb{E}[\|v_{sm+l+1}^{(\mathrm{Z})}\|^2] \\ &\quad \le \sum_{i=0}^{l} Q\mathbb{E}[\|v_{sm+i}^{(\mathrm{Z})}\|^2] + \frac{8L^2}{\alpha} H_\nu(\phi_{sm+l+1}) + P \\ &\quad \le \sum_{i=0}^{l} Q\left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P + \sum_{j=0}^{i-1} Q(Q+1)^{i-j-1}\left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+j}) + P\right)\right) \\ &\qquad + \frac{8L^2}{\alpha} H_\nu(\phi_{sm+l+1}) + P \\ &\quad = \frac{8L^2}{\alpha} H_\nu(\phi_{sm+l+1}) + P + \sum_{i=0}^{l} Q\left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P\right)\left(1 + \sum_{j=0}^{l-i-1} Q(Q+1)^j\right) \\ &\quad = \frac{8L^2}{\alpha} H_\nu(\phi_{sm+l+1}) + P + \sum_{i=0}^{l} Q\left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P\right)\left(1 + Q\frac{(Q+1)^{l-i} - 1}{(Q+1) - 1}\right) \\ &\quad = \frac{8L^2}{\alpha} H_\nu(\phi_{sm+l+1}) + P + \sum_{i=0}^{l} Q(Q+1)^{l+1-i-1}\left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P\right).\end{aligned}$$

In the second inequality, we used the hypothesis of mathematical induction. This is equivalent to using Gronwall's lemma. This concludes the proof.

$Q.E.D$

B.2 Main Proof

We are now ready to prove the main results. The main idea of the following proofs is due to Vempala and Wibisono (2019). We first evaluate how $H_\nu(\phi_k)$ decreases compared with the previous steps.

Theorem B.1. *Under Assumptions 1 and 2,* $0 < \eta < \frac{\alpha}{16\sqrt{2}L^2 m\gamma}$ *and* $\gamma \ge 1$*, for all* $k = sm + r$*, where* $s \in \mathbb{N} \cup \{0\}$ *and* $r = 0, \dots, m-1$*, the following holds in the update of SARAH-LD:*

$$\begin{aligned}H_\nu(\phi_{sm+r+1}) \le \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1 + \frac{\alpha}{4\gamma}\eta\right) H_\nu(\phi_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta} \sum_{i=0}^{r-1} \frac{\alpha}{4m\gamma}\eta \mathrm{e}^{-\frac{\alpha m}{\gamma}\eta} H_\nu(\phi_{sm+i}) \\ + 8\eta^2 dL^2\left(2 + \Xi + 2m\Xi\right),\end{aligned}$$

where $\Xi = \frac{(n-B)}{B(n-1)}$.

Proof. Note that from Lemma A.2, Talagrand's inequality is satisfied with constant α.

One step of SVRG-LD can be formulated as follows:

$$Z_{sm+r+1} \leftarrow Z_{sm+r} - \eta v^{(\mathrm{z})}_{sm+r} + \sqrt{2\eta/\gamma}\epsilon_{sm+r}.$$

This can be further interpreted as the output at time $t = \eta$ of the following SDE:

$$\mathrm{d}\tilde{Z}_t = -v^{(\mathrm{z})}_{sm+r}\mathrm{d}t + \sqrt{2/\gamma}\mathrm{d}B_t,\ \tilde{Z}_0 = Z_{sm+r}. \tag{12}$$

In this context, the distribution $\tilde{\phi}_t$ of $\tilde{Z}_t$ depends on both Z_{sm+r} and

$$\beta^{(\mathrm{z})}_{sm+r} := (v^{(\mathrm{z})}_{sm+r-1}, I_{sm+r}).$$

Let us define their joint distribution as follows:

$$\begin{aligned} \mathrm{d}\tilde{\phi}_{rt\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}, \tilde{Z}_t, \beta^{(\mathrm{z})}_{sm+r}) &= \mathrm{d}\tilde{\phi}_{r\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})\mathrm{d}\tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(\tilde{Z}_t|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r}) \\ &= \mathrm{d}\tilde{\phi}_{t\beta^{(\mathrm{z})}_{sm+r}}(\tilde{Z}_t, \beta^{(\mathrm{z})}_{sm+r})\mathrm{d}\tilde{\phi}_{r|t\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}|\tilde{Z}_t, \beta^{(\mathrm{z})}_{sm+r}). \end{aligned}$$

Then, the Fokker-Planck equation (2) when Z_{sm+r} and $\beta^{(\mathrm{z})}_{sm+r}$ are fixed becomes

$$\begin{aligned} \frac{\partial \tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(\tilde{Z}_t|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})}{\partial t} &= \nabla \cdot (\tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(\tilde{Z}_t|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})v^{(\mathrm{z})}_{sm+r}) \\ &\quad + \frac{1}{\gamma}\Delta\tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(\tilde{Z}_t|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r}). \end{aligned} \tag{13}$$

Therefore, the following holds about the distribution $\tilde{\phi}_t$ of $\tilde{Z}_t$ governed by equation (12),

$$\begin{aligned} \frac{\partial \tilde{\phi}_t(z)}{\partial t} &= \int \frac{\partial \tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(z|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})}{\partial t} \tilde{\phi}_{r\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})\mathrm{d}Z_{sm+r}\mathrm{d}\beta^{(\mathrm{z})}_{sm+r} \\ &= \int \left(\nabla \cdot (\tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(z|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})v^{(\mathrm{z})}_{sm+r}) + \frac{1}{\gamma}\Delta\tilde{\phi}_{t|r\beta^{(\mathrm{z})}_{sm+r}}(z|Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r}) \right) \\ &\qquad \cdot \tilde{\phi}_{r\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}, \beta^{(\mathrm{z})}_{sm+r})\mathrm{d}Z_{sm+r}\mathrm{d}\beta^{(\mathrm{z})}_{sm+r} \\ &= \int \nabla \cdot (\tilde{\phi}_{rt\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}, z, \beta^{(\mathrm{z})}_{sm+r})v^{(\mathrm{z})}_{sm+r})\mathrm{d}Z_{sm+r}\mathrm{d}\beta^{(\mathrm{z})}_{sm+r} \\ &\quad + \int \frac{1}{\gamma}\Delta\tilde{\phi}_{rt\beta^{(\mathrm{z})}_{sm+r}}(Z_{sm+r}, z, \beta^{(\mathrm{z})}_{sm+r})\mathrm{d}Z_{sm+r}\mathrm{d}\beta^{(\mathrm{z})}_{sm+r} \\ &= \nabla \cdot \left(\tilde{\phi}_t(z) \int \tilde{\phi}_{r\beta^{(\mathrm{z})}_{sm+r}|t} v^{(\mathrm{z})}_{sm+r}\mathrm{d}Z_{sm+r}\mathrm{d}\beta^{(\mathrm{z})}_{sm+r} \right) + \frac{1}{\gamma}\Delta\tilde{\phi}_t(z) \\ &= \nabla \cdot \left(\tilde{\phi}_t(z)\mathbb{E}_{\tilde{\phi}_{r\beta^{(\mathrm{z})}_{sm+r}|t}}[v^{(\mathrm{z})}_{sm+r}|\tilde{Z}_t = z] \right) + \frac{1}{\gamma}\Delta\tilde{\phi}_t(z), \end{aligned}$$

where for the second equation we used equation (13).

Plugging this to

$$\frac{\mathrm{d}}{\mathrm{d}t}H_\nu(\tilde{\phi}_t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^n} \tilde{\phi}_t \log\frac{\tilde{\phi}_t}{\nu}\mathrm{d}z = \int_{\mathbb{R}^n} \frac{\partial \tilde{\phi}_t}{\partial t}\log\frac{\tilde{\phi}_t}{\nu}\mathrm{d}z,$$

we obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} H_\nu(\tilde{\phi}_t) &= \int_{\mathbb{R}^n} \left(\nabla \cdot \left(\tilde{\phi}_t(z) \mathbb{E}_{\tilde{\phi}_{rZ|t}} [v^{(\mathrm{Z})}_{sm+r} | \tilde{Z}_t = z] \right) + \frac{1}{\gamma} \Delta \tilde{\phi}_t(z) \right) \log \frac{\tilde{\phi}_t}{\nu} \mathrm{d}z \\
&= \int \left(\nabla \cdot \left(\tilde{\phi}_t \left(\frac{1}{\gamma} \nabla \log \frac{\tilde{\phi}_t}{\nu} + \mathbb{E}_{\tilde{\phi}_{r\beta^{(\mathrm{Z})}_{sm+r}|t}} [v^{(\mathrm{Z})}_{sm+r} | \tilde{Z}_t = z] - \nabla F \right) \right) \right) \log \frac{\tilde{\phi}_t}{\nu} \mathrm{d}z \\
&= - \int \tilde{\phi}_t \left\langle \frac{1}{\gamma} \nabla \log \frac{\tilde{\phi}_t}{\nu} + \mathbb{E}_{\tilde{\phi}_{r\beta^{(\mathrm{Z})}_{sm+r}|t}} [v^{(\mathrm{Z})}_{sm+r} | \tilde{Z}_t = z] - \nabla F, \nabla \log \frac{\tilde{\phi}_t}{\nu} \right\rangle \mathrm{d}z \\
&= - \int \tilde{\phi}_t \frac{1}{\gamma} \left\| \log \frac{\tilde{\phi}_t}{\nu} \right\|^2 \mathrm{d}z \\
&\quad + \int_{\mathbb{R}^n} \tilde{\phi}_t \left\langle \nabla F - \mathbb{E}_{\tilde{\phi}_{r\beta^{(\mathrm{Z})}_{sm+r}|t}} [v^{(\mathrm{Z})}_{sm+r} | \tilde{Z}_t = z], \nabla \log \frac{\tilde{\phi}_t}{\nu} \right\rangle \mathrm{d}z \\
&= - \frac{1}{\gamma} J_\nu(\tilde{\phi}_t) \\
&\quad + \int \tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}} \left\langle \nabla F - v^{(\mathrm{Z})}_{sm+r}, \nabla \log \frac{\tilde{\phi}_t}{\nu} \right\rangle \mathrm{d}Z_{sm+r} \mathrm{d}z \mathrm{d}\beta^{(\mathrm{Z})}_{sm+r} \\
&= - \frac{1}{\gamma} J_\nu(\tilde{\phi}_t) + \mathbb{E}_{\tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}}} \left[\left\langle \nabla F(\tilde{Z}_t) - v^{(\mathrm{Z})}_{sm+r}, \nabla \log \frac{\tilde{\phi}_t(\tilde{Z}_t)}{\nu(\tilde{Z}_t)} \right\rangle \right].
\end{aligned}
$$

Now, let us define the second term of the right-hand side of the very last equality as Ⓑ. Applying $\langle a, b \rangle \leq \gamma \|a\|^2 + \frac{1}{4\gamma} \|b\|^2$ to this, we obtain

$$
\begin{aligned}
\text{Ⓑ} &\leq \gamma \mathbb{E}_{\tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}}} \left[\|\nabla F(\tilde{Z}_t) - v^{(\mathrm{Z})}_{sm+r}\|^2 \right] + \frac{1}{4\gamma} \mathbb{E}_{\tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}}} \left[\left\| \nabla \log \frac{\tilde{\phi}_t(\tilde{Z}_t)}{\nu(\tilde{Z}_t)} \right\|^2 \right] \\
&\leq 2\gamma \mathbb{E}_{\tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}}} \left[\|\nabla F(\tilde{Z}_t) - \nabla F(Z_{sm+r})\|^2 \right] + 2\gamma \mathbb{E}_{\tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}}} [\|\nabla F(Z_{sm+r}) - v^{(\mathrm{Z})}_{sm+r}\|^2] \\
&\quad + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t) \\
&\leq 2\gamma L^2 \mathbb{E}_{\tilde{\phi}_{rt\beta^{(\mathrm{Z})}_{sm+r}}} [\|\tilde{Z}_t - Z_{sm+r}\|^2] + \sum_{i=1}^{r} 2\gamma \Xi L^2 \eta^2 \mathbb{E}[\|v^{(\mathrm{Z})}_{sm+i-1}\|^2] + 4\eta m d L^2 \Xi \\
&\quad + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t),
\end{aligned}
$$

where for the last inequality, we used the smoothness of F and Lemma B.2.

As $\tilde{Z}_t = Z_{sm+r} - t v^{(\mathrm{Z})}_{sm+r} + \sqrt{2t/\gamma} \epsilon_{sm+r}$ ($\epsilon_{sm+r} \sim N(0, I)$), from Lemma B.3, we have

$$
\begin{aligned}
\mathbb{E}[\|\tilde{Z}_t - Z_{sm+r}\|^2] &= \mathbb{E}[\| - t v^{(\mathrm{Z})}_{sm+r} + \sqrt{2t/\gamma} \epsilon_{sm+r}\|^2] \\
&= t^2 \mathbb{E}[\|v^{(\mathrm{Z})}_{sm+r}\|^2] + 2td/\gamma \\
&\leq t^2 \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+r}) + P \right) \\
&\quad + t^2 \sum_{i=0}^{r-1} Q(Q+1)^{r-i-1} \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P \right) \\
&\quad + 2td/\gamma.
\end{aligned}
$$

Furthermore, by the proof of Lemma B.3, we know that the following holds:

$$
\sum_{i=1}^{r} \mathbb{E}[\|v^{(\mathrm{Z})}_{sm+i-1}\|^2] \leq \sum_{i=0}^{r-1} (Q+1)^{r-i-1} \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P \right).
$$

As a result, taking into account that we are only concerned about the time interval $0 \le t \le \eta$, applying $t \le \eta$, we conclude

$$\begin{aligned}
\textcircled{B} &\le 2\gamma L^2\eta^2 \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+r}) + P + \sum_{i=0}^{r-1} Q(Q+1)^{r-i-1} \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P \right) \right) \\
&\quad + 4\eta d L^2 + 2\gamma L^2 \eta^2 \Xi \sum_{i=0}^{r-1} (Q+1)^{r-i-1} \left(\frac{8L^2}{\alpha} H_\nu(\phi_{sm+i}) + P \right) + 4\eta m d L^2 \Xi \\
&\quad + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t) \\
&\le \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} (Q+1)^{r-i} \frac{16\gamma L^4 \eta^2}{\alpha} H_\nu(\phi_{sm+i}) + 2\gamma L^2 \eta^2 \sum_{i=0}^{r} (Q+1)^{r-i} P \\
&\quad + 4\eta d L^2 (1 + 2m\Xi) + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t) \\
&\le \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} \frac{16L^4\gamma\eta^2}{\alpha} (Q+1)^r H_\nu(\phi_{sm+i}) + 2\gamma L^2 \eta^2 \sum_{i=0}^{r} (Q+1)^r P \\
&\quad + 4\eta d L^2 (1 + 2m\Xi) + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t) \\
&\le \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} \frac{16L^4\gamma\eta^2}{\alpha} (Q+1)^m H_\nu(\phi_{sm+i}) + 2\gamma L^2 \eta^2 m (Q+1)^m P \\
&\quad + 4\eta d L^2 (1 + 2m\Xi) + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t).
\end{aligned}$$

where for the second inequality we used $\Xi \le 1$ and for the last inequality $r < m$.

Here, as $\Xi \le 1$ and $\eta \le \frac{1}{4mL}$ by $\alpha \le \gamma L$,

$$(Q+1)^m \le \mathrm{e}^{Qm} = \mathrm{e}^{2L^2 m \eta^2 \Xi} \le \mathrm{e}^{1/4} \le 2.$$

Therefore,

$$\begin{aligned}
\textcircled{B} &\le \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} \frac{32L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+i}) \\
&\quad + 4\gamma L^2 \eta^2 m P + 4\eta d L^2 (1 + 2m\Xi) + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t) \\
&\le \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} \frac{32L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+i}) \\
&\quad + 4\gamma L^2 \eta^2 m \left(\frac{4dL}{\gamma} + \frac{4\eta m d L^2}{\gamma} \Xi \right) + 4\eta d L^2 (1 + 2m\Xi) + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t) \\
&\le \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} \frac{32L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+i}) \\
&\quad + 4\eta d L^2 \left(2 + \Xi + 2m\Xi\right) + \frac{1}{4\gamma} J_\nu(\tilde{\phi}_t).
\end{aligned}$$

where for the last inequality, we used $\eta \le \frac{1}{4mL}$.

Thus,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} H_\nu(\tilde{\phi}_t) &\le -\frac{3}{4\gamma} J_\nu(\tilde{\phi}_t) + \frac{16L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1} \frac{32L^4\gamma\eta^2}{\alpha} H_\nu(\phi_{sm+i}) \\
&\quad + 4\eta d L^2 \left(2 + \Xi + 2m\Xi\right).
\end{aligned}$$

According to Assumption 2,

$$\frac{\mathrm{d}}{\mathrm{d}t}H_\nu(\tilde{\phi}_t) \leq -\frac{3\alpha}{2\gamma}H_\nu(\tilde{\phi}_t) + \frac{16L^4\gamma\eta^2}{\alpha}H_\nu(\phi_{sm+r}) + \sum_{i=0}^{r-1}\frac{32L^4\gamma\eta^2}{\alpha}H_\nu(\phi_{sm+i}) \\ + 4\eta dL^2\left(2+\Xi+2m\Xi\right).$$

Grouping the second to fourth terms as $U^{(\mathrm{Z})}_{sm+r}$ and multiplying both sides by $\mathrm{e}^{\frac{3\alpha}{2\gamma}t}$, we can write the above equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{\frac{3\alpha}{2\gamma}t}H_\nu(\tilde{\phi}_t)\right) \leq \mathrm{e}^{\frac{3\alpha}{2\gamma}t}U^{(\mathrm{Z})}_{sm+r}.$$

Integrating both sides from $t=0$ to $t=\eta$ and using $\tilde{\phi}_\eta = \phi_{sm+r+1}$, we obtain

$$\mathrm{e}^{\frac{3\alpha}{2\gamma}\eta}H_\nu(\phi_{sm+r+1}) - H_\nu(\phi_{sm+r}) \leq \frac{2\gamma(\mathrm{e}^{\frac{3\alpha}{2\gamma}\eta}-1)}{3\alpha}U^{(\mathrm{Z})}_{sm+r} \\ \leq 2\eta U^{(\mathrm{Z})}_{sm+r}.$$

Here, for the last inequality, we used $\mathrm{e}^c \leq 1+2c$ ($0<c=\frac{3\alpha}{2\gamma}\eta\leq 1$) holds since $0<\eta\leq \frac{\alpha}{16\sqrt{2}L^2m\gamma}\leq\frac{2\gamma}{3\alpha}$, where we used $1/L\leq\gamma/\alpha$ and $m\geq 1$. Rearranging this, we obtain

$$H_\nu(\phi_{sm+r+1}) \leq \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{32\gamma L^4\eta^3}{\alpha}\right)H_\nu(\phi_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{64\gamma L^4\eta^3}{\alpha}H_\nu(\phi_{sm+i}) \\ + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}8\eta^2dL^2\left(2+\Xi+2m\Xi\right). \tag{14}$$

Furthermore, since $\eta\leq\frac{\alpha}{16\sqrt{2m}L^2\gamma}\leq\frac{\alpha}{8\sqrt{3}L^2\gamma}$ and $\mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\leq 1$,

$$H_\nu(\phi_{sm+r+1}) \leq \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)H_\nu(\phi_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{8\gamma m}\eta H_\nu(\phi_{sm+i}) \\ + 8\eta^2dL^2\left(2+\Xi+2m\Xi\right).$$

On the other hand, since $\eta\leq\frac{\alpha}{8mL^2\gamma}$ and $\alpha\leq\gamma L$ holds,

$$\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta} \geq \mathrm{e}^{-\frac{\alpha m}{\gamma}\cdot\frac{\alpha}{8mL^2\gamma}} = \mathrm{e}^{-\frac{\alpha^2}{8L^2\gamma^2}} \geq \mathrm{e}^{-1/8} \geq 0.88 \geq \frac{1}{2},$$

which further implies

$$H_\nu(\phi_{sm+r+1}) \leq \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\left(1+\frac{\alpha}{4\gamma}\eta\right)H_\nu(\phi_{sm+r}) + \mathrm{e}^{-\frac{3\alpha}{2\gamma}\eta}\sum_{i=0}^{r-1}\frac{\alpha}{4m\gamma}\eta\mathrm{e}^{-\frac{\alpha m}{\gamma}\eta}H_\nu(\phi_{sm+i}) \\ + 8\eta^2dL^2\left(2+\Xi+2m\Xi\right).$$

$Q.E.D$

Finally, let us prove Theorem 2 and Corollary 2.1.

Theorem B.2 (Theorem 2 restated)**.** *Under Assumptions 1 and 2,* $0<\eta<\frac{\alpha}{16\sqrt{2}L^2m\gamma}$ *and* $\gamma\geq 1$*, for all* k*, the following holds in the update of SARAH-LD:*

$$H_\nu(\phi_k) \leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k}H_\nu(\phi_0) + \frac{32\eta\gamma dL^2}{3\alpha}\left(2+\Xi+2m\Xi\right),$$

where $\Xi=\frac{(n-B)}{B(n-1)}$.

Proof. Same as Theorem A.2.

$Q.E.D$

Corollary B.2.1 (Corollary 2.1 restated)**.** *Under the same assumptions as Theorem B.2, for all $\epsilon \geq 0$, if we choose step size η such that*

$$\eta \leq \frac{3\alpha\epsilon}{64\gamma dL^2} \left(2 + \Xi + 2m\Xi\right)^{-1},$$

then a precision $H_\nu(\phi_k) \leq \epsilon$ is reached after

$$k \geq \frac{\gamma}{\alpha\eta} \log \frac{2H_\nu(\phi_0)}{\epsilon}$$

steps. Especially, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \frac{\alpha}{16\sqrt{2}L^2\sqrt{n}\gamma} \wedge \frac{3\alpha\epsilon}{320dL^2\gamma}$, then the gradient complexity becomes

$$\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right).$$

Proof. The first half of the statement is the same as Corollary A.2.1.

When $B \geq m$, from Theorem B.2, we obtain

$$\begin{aligned} H_\nu(\phi_k) &\leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\phi_0) + \frac{32\eta\gamma dL^2}{3\alpha}\left(2 + \Xi + 2m\Xi\right) \\ &\leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\phi_0) + \frac{160\eta\gamma dL^2}{3\alpha}. \end{aligned}$$

Proceeding in the same way as Corollary A.2.1, we obtain the optimal gradient complexity of

$$\tilde{O}\left(\left(n + \frac{dn^{\frac{1}{2}}}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right)$$

with $B = m = \sqrt{n}$ and $\eta = \frac{\alpha}{16\sqrt{2}L^2\sqrt{n}\gamma} \wedge \frac{3\alpha\epsilon}{320dL^2\gamma}$.

Now, when $B \leq m$, from Theorem B.2, we obtain

$$\begin{aligned} H_\nu(\phi_k) &\leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\phi_0) + \frac{32\eta\gamma dL^2}{3\alpha}\left(2 + \Xi + 2m\Xi\right) \\ &\leq \mathrm{e}^{-\frac{\alpha\eta}{\gamma}k} H_\nu(\phi_0) + \frac{160\eta\gamma dL^2}{3\alpha}\frac{m}{B}. \end{aligned}$$

This leads to a gradient complexity of

$$\tilde{O}\left(\left(n + \frac{d(m + n/B)}{\epsilon}\right) \cdot \frac{\gamma^2 L^2}{\alpha^2}\right)$$

with $\eta = \frac{\alpha}{16\sqrt{2}L^2 m\gamma} \wedge \frac{3\alpha\epsilon}{320dL^2\gamma}\frac{B}{m}$, which is optimal with $B = m = \sqrt{n}$ again.

Q.E.D

C Proof of Theorem 3, Corollaries 3.1 and 3.2

We define X_k like Algorithm 1 in order to simultaneously represent Y_k and Z_k.

C.1 Preparation for the Proof

C.1.1 Link between Sampling and Optimization

Since

$$\mathbb{E}_{X_k}[F(X_k)] - F(X^*)$$

can be separated into the discretisation error

$$\mathbb{E}_{X_k}[F(X_k)] - \mathbb{E}_{X\sim\nu}[F(X)]$$

and the approximation error due to sampling

$$\mathbb{E}_{X\sim\nu}[F(X)] - F(X^*),$$

in this subsection, we analyse the upper bound of these two terms.

Property C.1. *Under Assumption 1, the following holds:*

$$\forall x \in \mathbb{R}^d,\ F(x) - F(x^*) \geq \frac{1}{2L}\|\nabla F(x)\|,$$

where x^* is the global minimum of F.

Proof. Let us define $G(x) := F(x) - F(x^*)$. Since G is also L-smooth,

$$G\left(x - \frac{1}{L}\nabla G(x)\right) \leq G(x) - \frac{1}{L}\|\nabla G(x)\|^2 + \frac{1}{2L}\|\nabla G(x)\|^2 = G(x) - \frac{1}{2L}\|\nabla G(x)\|^2,$$

where for the inequality, we used that the following holds for a L-smooth function H:

$$\forall x, y \in \mathbb{R}^d,\ H(y) \leq H(x) + \langle \nabla H(x), y - x\rangle + \frac{L}{2}\|y - x\|^2.$$

Now, since $G \geq 0$, we obtain

$$F(x) - F(x^*) \geq \frac{1}{2L}\|\nabla F(x)\|,$$

which concludes the proof.

Q.E.D

Theorem C.1. *Under Assumption 1, the following holds for distributions ρ_k and ν:*

$$\mathbb{E}_{X_k \sim \rho_k}[F(X_k)] - \mathbb{E}_{X \sim \nu}[F(X)] \leq LW_2^2(\rho_k, \nu) + \mathbb{E}_{X \sim \nu}[F(X)] - F(X^*),$$

where X^* is the global minimum of F. The same statement holds with $X_k \sim \phi_k$.

Proof. Let $X_k \sim \rho_k$ and $X \sim \nu$ be an optimal coupling so that $\mathbb{E}[\|X_k - X\|^2] = W_2(\rho_k, \nu)^2$. The following holds only from the smoothness of F:

$$\begin{aligned}
F(X_k) - F(X) &= \int_0^1 \langle X_k - X, \nabla F\left((1-t)X + tX_k\right)\rangle \mathrm{d}t \\
&\leq \int_0^1 \|X_k - X\|\|\nabla F\left((1-t)X + tX_k\right)\|\mathrm{d}t \\
&\leq \int_0^1 \|X_k - X\|\|\nabla F\left((1-t)X + tX_k\right) - \nabla F(X)\| \\
&\qquad\qquad + \|X_k - X\|\|\nabla F(X)\|\mathrm{d}t \\
&\leq \int_0^1 Lt\|X_k - X\|^2 + \frac{L}{2}\|X_k - X\|^2 + \frac{1}{2L}\|\nabla F(X)\|^2\mathrm{d}t \\
&\leq L\|X_k - X\|^2 + F(X) - F(X^*).
\end{aligned}$$

For the first inequality, we used the Cauchy-Schwarz inequality, for the third inequality we used the smoothness of F on the first term, and for the fourth inequality, Property C.1.

Hence, taking expectation of both sides, we obtain the desired result.

Q.E.D

Corollary C.1.1. *Under the same assumptions as Theorem C.1, the following holds:*

$$\mathbb{E}_{X_k \sim \rho_k}[F(X_k)] - F(X^*) \leq LW_2^2(\rho_k, \nu) + 2\left(\mathbb{E}_{X \sim \nu}[F(X)] - F(X^*)\right).$$

The same statement holds with $X_k \sim \phi_k$.

An important feature of this theorem is that the square of the 2-Wasserstein metric appears. Thanks to this and Talagrand's inequality, we can directly use the results from sampling (e.g., Corollaries 1.1 and 2.1)

The approximation error can be bounded thanks to the following theorem from Raginsky et al. (2017).

Theorem C.2 (Raginsky et al. (2017), Proposition 11)**.** *Under Assumptions 1 and 3, for all* $\gamma \geq \frac{2}{M}$

$$\mathbb{E}_{X\sim\nu}[F(X)] - F(X^*) \leq \frac{d}{2\gamma} \log\left(\frac{\mathrm{e}L}{M}\left(\frac{b\gamma}{d}+1\right)\right).$$

Corollary C.2.1. *Under the same assumptions as Theorem C.2, for all* $\epsilon > 0$*, if we additionally require* $\gamma \geq \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2}$*, then*

$$\mathbb{E}_{X\sim\nu}[F(X)] - F(X^*) \leq \frac{\epsilon}{4}.$$

Proof. Since

$$\frac{d}{2\gamma}\log\left(\frac{\mathrm{e}L}{M}\left(\frac{b\gamma}{d}+1\right)\right) = \frac{d}{2\gamma}\log\frac{\mathrm{e}L}{M} + \frac{d}{2\gamma}\log\left(\frac{b\gamma}{d}+1\right),$$

it suffices to have $\frac{d}{2\gamma}\log\frac{\mathrm{e}L}{M} \leq \frac{\epsilon}{8}$ and $\frac{d}{2\gamma}\log\left(\frac{b\gamma}{d}+1\right) \leq \frac{\epsilon}{8}$. Furthermore, since for all $x \geq 0$

$$\frac{\log(x+1)}{x} \leq \frac{1}{\sqrt{x+1}} \leq \frac{1}{\sqrt{x}}$$

holds, we only need to require $\frac{d}{2\gamma}\log\frac{\mathrm{e}L}{M} \leq \frac{\epsilon}{8}$ and $\frac{b}{2}\frac{1}{\sqrt{b\gamma/d}} \leq \frac{\epsilon}{8}$. Solving these two inequalities according to γ leads to the desired result.

Q.E.D

Remark C.1. *The lower bound* $\frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2}$ *is only calculated to acquire a concrete condition on* γ*. A more involved analysis could find a better lower bound.*

C.1.2 Explicit Formulation of the Log-Sobolev Constant

In this subsection, we give an explicit formulation of the Log-Sobolev constant of $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$ in function of γ for two cases: under Assumptions 1 and 3, and under Assumptions 1, 3 and 4 to 6. The second case is roughly the first combined with the Morse condition.

When we only assume dissipativity and smoothness, we can obtain a Log-Sobolev constant whose inverse exponentially depends on the inverse temperature γ. This employs the following result from Raginsky et al. (2017).

Property C.2 (Raginsky et al. (2017), Proposition 9)**.** *Under Assumptions 1 and 3, for all* $\gamma \geq \frac{2}{M}$*,* ν *satisfies Log-Sobolev inequality with a constant* α *such that*

$$\frac{1}{\alpha} \leq \frac{2M^2+2L^2}{M^2L\gamma} + \frac{1}{\lambda_*}\left(\frac{6L(d+\gamma)}{M}+2\right),$$

where

$$\frac{1}{\lambda_*} \leq \frac{1}{M\gamma(d+b\gamma)} + \frac{2C_*(d+b\gamma)}{M\gamma}\exp\left(\frac{2}{M}(L+B_*)(b\gamma+d)+\gamma(A_*+B_*)\right).$$

Here, $A_* = \max_i\{|f_i(0)|\}$*,* $B_* = \max_i\{|\nabla f_i(0)|\}$ *and* C_* *is a universal constant that does not depend on* F*.*

From this, we immediately have the following property.

Property C.3. *Under Assumptions 1 and 3, for all* $\gamma \geq \frac{2}{M}$*, we can take a Log-Sobolev constant* α *of* ν *which can be written with constants* C_1 *and* $C_2 > 0$ *independent of* γ *as follows:*

$$\alpha = \gamma C_1 \mathrm{e}^{-C_2\gamma},$$

where

$$C_1 = \left(\frac{2M^2+2L^2}{M^2L} + \left(\frac{6Ld}{M}+2\right)\left(\frac{1}{Md} + \frac{2C_*d}{M}\mathrm{e}^{\frac{2d}{M}(L+B_*)}\right)\right)^{-1},$$

and

$$C_2 = \frac{2b}{M}(L+B_*) + (A_*+B_*) + b + 1.$$

Proof. From Proposition C.2,

$$\frac{\gamma}{\alpha} \leq \frac{2M^2 + 2L^2}{M^2 L} + \frac{\gamma}{\lambda_*}\left(\frac{6L(d+\gamma)}{M} + 2\right),$$

and

$$\frac{\gamma}{\lambda_*} \leq \frac{1}{M(d+b\gamma)} + \frac{2C_*(d+b\gamma)}{M}\exp\left(\frac{2}{M}(L+B_*)(b\gamma+d) + \gamma(A_*+B_*)\right).$$

Roughly bounding these inequalities, we obtain

$$\begin{aligned}\frac{\gamma}{\lambda_*} &\leq \frac{1}{Md} + \frac{2C_*(d+b\gamma)}{M}\mathrm{e}^{\frac{2d}{M}(L+B_*)}\mathrm{e}^{\left(\frac{2b}{M}(L+B_*)+(A_*+B_*)\right)\gamma} \\ &\leq \frac{1}{Md} + \frac{2C_*d}{M}\mathrm{e}^{\frac{2d}{M}(L+B_*)}\mathrm{e}^{\left(\frac{2b}{M}(L+B_*)+(A_*+B_*)+b\right)\gamma} \\ &\leq \left(\frac{1}{Md} + \frac{2C_*d}{M}\mathrm{e}^{\frac{2d}{M}(L+B_*)}\right)\mathrm{e}^{\left(\frac{2b}{M}(L+B_*)+(A_*+B_*)+b\right)\gamma},\end{aligned}$$

where for the second inequality we used $d + b\gamma \geq d\mathrm{e}^{b\gamma}$ for all $\gamma > 0$ when $d \geq 1$. Thus,

$$\begin{aligned}\frac{\gamma}{\alpha} &\leq \frac{2M^2+2L^2}{M^2L} + \frac{\gamma}{\lambda_*}\left(\frac{6L(d+\gamma)}{M}+2\right) \\ &\leq \frac{2M^2+2L^2}{M^2L}\mathrm{e}^{\left(\frac{2b}{M}(L+B_*)+(A_*+B_*)+b+1\right)\gamma} + \frac{\gamma}{\lambda_*}\left(\frac{6Ld}{M}+2\right)\mathrm{e}^{\gamma} \\ &\leq \left(\frac{2M^2+2L^2}{M^2L} + \left(\frac{6Ld}{M}+2\right)\left(\frac{1}{Md} + \frac{2C_*d}{M}\mathrm{e}^{\frac{2d}{M}(L+B_*)}\right)\right)\mathrm{e}^{\left(\frac{2b}{M}(L+B_*)+(A_*+B_*)+b+1\right)\gamma}.\end{aligned}$$

Finally, taking into account that a lower bound of a Log-Sobolev constant automatically satisfies the Log-Sobolev inequality, we obtain the desired result.

$Q.E.D$

On the other hand, under the additional condition of Morse, Lipschitzness of $\nabla^2 F$ and other minor assumptions, we can obtain a far better Log-Sobolev constant whose inverse depends only linearly on γ as follows. This is a straightforward adaptation of Li and Erdogdu's result (Li and Erdogdu, 2020). We provide a proof in Appendix D.

Property C.4. *Under Assumptions 1, 3 and 4 to 6, with $\gamma \geq 1$ such that*

$$\gamma \geq C_\gamma := \max\left(1, \left(\frac{24dL}{C_F^2}\right)^2 \frac{4dL'^2}{\lambda^{\dagger 2}}, 4L'^2\left(\frac{24dL}{C_F^2}\right)^6\right),$$

where C_F is defined in Lemma D.5, ν satisfies the Log-Sobolev inequality with constant α such that

$$\frac{1}{\alpha} = \frac{\gamma}{C_3},$$

where

$$C_3 := \left(\frac{2M^2+8L^2}{M^2L} + \left(\frac{6L(d+1))}{M}+2\right)\frac{35}{\lambda^\dagger}\right).$$

C.2 Main Proof

Theorem C.3 (Theorem 3 restated)**.** *Using SVRG-LD or SARAH-LD, under Assumptions 1 to 3, $0 < \eta < \frac{\alpha}{16\sqrt{6}L^2 m\gamma}$, $\gamma \geq \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2} \vee 1 \vee \frac{2}{M}$ and $B \geq m$, if we take $B = m = \sqrt{n}$ and the largest permissible step size $\eta = \frac{\alpha}{16\sqrt{6}L^2\sqrt{n}\gamma} \wedge \frac{3}{1792}\frac{\alpha^2\epsilon}{L^2 d\gamma}$, the gradient complexity to reach a precision of*

$$\mathbb{E}_{X_k}[F(X_k)] - F(X^*) \leq \epsilon$$

is

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon}\cdot\frac{dL}{\alpha}\right)\frac{\gamma^2L^2}{\alpha^2}\right),$$

where α is a function of γ.

Proof. It is sufficient to consider the case of SVRG-LD with $X_k \sim \rho_k$. From Corollary C.1.1, the sufficient condition for

$$\mathbb{E}_{X_k}[F(X_k)] - F(X^*) \leq \epsilon$$

is $LW_2^2(\rho_k, \nu) \leq \epsilon/2$ and $\mathbb{E}_{X\sim\nu}[F(X)] - F(X^*) \leq \epsilon/4$. From Corollary C.2.1, the latter condition is satisfied when $\gamma \geq \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8bd}{\epsilon^2} \vee 1 \vee \frac{2}{M}$. Moreover, concerning the former, from Talagrand's inequality

$$W_2^2(\rho_k, \nu) \leq \frac{2}{\alpha} H_\nu(\rho_k),$$

it suffices to have

$$H_\nu(\rho_k) \leq \frac{\alpha\epsilon}{4L}.$$

Thus, from Corollaries A.2.1 and B.2.1, under the same conditions, we obtain a gradient complexity of

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon} \cdot \frac{dL}{\alpha}\right)\frac{\gamma^2 L^2}{\alpha^2}\right).$$

Q.E.D

This leads to the following corollaries.

Corollary C.3.1 (Corollary 3.1 restated)**.** *Under the same assumptions as Theorem C.3, taking*

$$\gamma = i(\epsilon) := \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2} \vee 1 \vee \frac{2}{M},$$

we obtain a gradient complexity of

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon} \cdot \frac{dL}{C_1 i(\epsilon)}\mathrm{e}^{C_2 i(\epsilon)}\right) L^2 \mathrm{e}^{2C_2 i(\epsilon)}\right)$$

since $\alpha = \gamma C_1 \mathrm{e}^{-C_2\gamma}$ (Property C.3).

Proof. The proof follows from Property C.3.

Q.E.D

Corollary C.3.2 (Corollary 3.2 restated)**.** *Under the same assumptions as Theorem 11 and Assumptions 4 to 6, taking*

$$\gamma = j(\epsilon) := \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8db}{\epsilon^2} \vee 1 \vee \frac{2}{M} \vee C_\gamma,$$

where C_γ is a constant independent of ϵ defined in Property C.4, we obtain a gradient complexity of

$$\tilde{O}\left(\left(n + \frac{n^{\frac{1}{2}}}{\epsilon} \cdot \frac{dL}{C_3} j(\epsilon)\right) C_3^2 j(\epsilon)^4 L^2\right)$$

since $\alpha = C_3/\gamma$ (Property C.4).

Proof. The proof follows from Property C.4.

Q.E.D

D Proof of Property C.4

D.1 Overview and Main Result

In this appendix, we prove Property C.4 which is only a slight adaptation of Theorem 3.4 from Li and Erdogdu (2020), which builds its foundation from prior work such as Cattiaux et al. (2010) and Menz and Schlichting (2014). We show that with additional Morse, and smoothness assumptions to dissipativity, we can obtain a Log-Sobolev constant of $\mathrm{d}\nu \propto \mathrm{e}^{-\gamma F}\mathrm{d}x$ whose inverse only depends linearly on the inverse temperature parameter. Property C.4 is reminded below in a more precise form.

Theorem 4. *Under Assumptions 1, 3 and 4 to 6, with $a > 0$ and $\gamma \geq 1$ such that*

$$a^2 \geq \frac{24dL}{C_F^2},$$

and

$$\gamma \geq \max\left(1, a^2 \frac{4dL'^2}{\lambda^{\dagger 2}}, 4L'^2 a^6\right),$$

ν satisfies the Log-Sobolev inequality with constant α such that

$$\frac{1}{\alpha} = \left(\frac{2M^2 + 8L^2}{M^2 L} + \left(\frac{6L(d+1))}{M} + 2\right)\frac{35}{\lambda^\dagger}\right)\gamma.$$

This theorem shows that the strict saddle node assumption is almost sufficient to obtain in the Euclidean space for dissipative distributions a Log-Sobolev constant whose inverse does not exponentially depend on the inverse temperature, which was the case without this assumption.

D.2 Preliminaries

We first clarify some definitions.

Definition D.1. *We say a probability measure ν satisfies the Poincaré inequality with a constant κ if for all smooth $g : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}_\nu[g^2] - \mathbb{E}_\nu[g]^2 \leq \frac{1}{\kappa}\mathbb{E}_\nu[\|\nabla g\|^2].$$

Definition D.2. *A probability measure ν on $\mathbb{R}^d$ restricted on a set $\mathcal{Z} \subset \mathbb{R}^d$ is defined as*

$$\nu|_{\mathcal{Z}} := \frac{\nu(x)}{\int_{\mathcal{Z}} \nu(y)dy} \mathbb{1}_{\mathcal{Z}}(x).$$

Definition D.3. *We define the following sets:*

$$\mathcal{B} := \left\{x \in \mathbb{R}^d \mid d(x, \mathcal{S})^2 < \frac{a^2}{\gamma}\right\},$$
$$\mathcal{U} := \left\{x \in \mathbb{R}^d \mid d(x, \mathcal{X})^2 < \frac{a^2}{\gamma}\right\},$$
$$\mathcal{A} := \left\{x \in \mathbb{R}^d \mid d(x, \mathcal{S} \cup \mathcal{X})^2 \geq \frac{a^2}{4\gamma}\right\},$$

where $\mathcal{X}$ is the set of global minima and $\mathcal{S}$ is the set of stationary points except the global minima. Note that $\mathcal{B} \cup \mathcal{U} \cup \mathcal{A} = \mathbb{R}^d$.

Here, the distance from a point $x \in \mathbb{R}^d$ and a set $\mathcal{Z} \subset \mathbb{R}^d$ is defined as

$$d(x, \mathcal{Z}) := \inf_{z \in \mathcal{Z}} \|x - z\|.$$

In this appendix, we only consider the following generator $\mathcal{L}$.

Definition D.4. *We define $\mathcal{L}$ such that*

$$\mathcal{L}f := \langle -\nabla f, \nabla F\rangle + \frac{1}{\gamma}\Delta f, \ \ \forall f \in C^2(\mathbb{R}^d)$$

which is the generator of the gradient Langevin Dynamics (1).

We will need some lemmas proved by Li and Erdogdu (2020).

Lemma D.1 (Li and Erdogdu (2020))**.** *Under Assumptions 1 and 5, suppose $y \in \mathbb{R}^d$ is a stationary point of F. Then, with*

$$H(x) := \nabla^2 F(0) \cdot x$$

defined in the coordinate centered at y, we obtain for all $x \in \mathbb{R}^d$,

$$\|\nabla F - H(x)\| \leq L'\|x\|^2.$$

Proof. From the mean value theorem, there exist a $\hat{x}$ on the line between 0 and x such that

$$\nabla F(x) = \nabla F(0) + \nabla^2 F(\hat{x}) \cdot x = \nabla^2 F(\hat{x}) \cdot x,$$

where for the last equality, we used that 0 was a stationary point. Therefore, we obtain

$$\begin{aligned} \|\nabla F(x) - H(x)\| &= \|\nabla^2 F(\hat{x}) \cdot x - \nabla^2 F(0) \cdot x\| \\ &\leq \|\nabla^2 F(\hat{x}) - \nabla^2 F(0)\| \|x\| \\ &\leq L' \|\hat{x}\| \|x\| \\ &\leq L' \|x\|^2, \end{aligned}$$

where for the second inequality we used the L'-Lipschitzness of $\nabla^2 F$, and for the last inequality we used $\|\hat{x}\| \leq \|x\|$.

$Q.E.D$

Lemma D.2 (Li and Erdogdu (2020), Proposition E.5)**.** *Let W_t and $\tilde{W}_t$ be weak solutions on some filtered probability space of the following one dimensional SDE's:*

$$\begin{aligned} \mathrm{d}W_t &= \Phi(W_t)\mathrm{d}t + \sigma \mathrm{d}B_t, \\ \mathrm{d}\tilde{W}_t &= \tilde{\Phi}(\tilde{W}_t)\mathrm{d}t + \sigma \mathrm{d}B_t, \end{aligned}$$

where $W_0 = \tilde{W}_0$ a.s. and $\sigma > 0$ is a constant. We further assume that for all $T \geq 0$,

$$\int_0^T |\Phi(W_t)| + |\tilde{\Phi}(\tilde{W}_t)| \mathrm{d}t < \infty, \ a.s.$$

If $\Phi(W_t) \geq \tilde{\Phi}(\tilde{W}_t)$ for all $x \in \mathbb{R}$, then $W_t \geq \tilde{W}_t$ a.s.

Lemma D.3 (Li and Erdogdu (2020), Corollary D.6)**.** *Consider the following Cox-Ingersoll-Ross process defined as*

$$\mathrm{d}W_t = \left(2\lambda^\dagger W_t + \frac{1}{2\gamma}\right)\mathrm{d}t + \frac{2}{\sqrt{\gamma}}\sqrt{W_t}\mathrm{d}B_t, \ \ W_0 = w_0 \geq 0,$$

where $\lambda^\dagger > 0$, $\gamma > 0$ and $\{B_t\}_{t \geq 0}$ is a standard one dimensional Brownian motion. Then for its unique strong solution W_t, we have the following density function:

$$f(w;t) = 2^{-\frac{5}{4}} \left(\frac{w}{w_0}\right)^{-\frac{1}{4}} \frac{\lambda^\dagger \gamma}{\mathrm{e}^{\frac{\lambda^\dagger t}{2}} \sinh(\lambda^\dagger t)} \exp\left(\frac{\lambda^\dagger \gamma (w\mathrm{e}^{-2\lambda^\dagger t} - \frac{w_0}{2})}{1 - \mathrm{e}^{-2\lambda^\dagger t}}\right) I_{-\frac{1}{2}} \left(\frac{\lambda^\dagger \gamma}{\sinh(\lambda^\dagger t)} \sqrt{\frac{w w_0}{2}}\right)$$

for $w > 0$ and $f(w;t) = 0$ for $w = 0$, where $I_{-\frac{1}{2}}$ is the modified Bessel function of the first kind of degree $-\frac{1}{2}$. Thus, $W_t > 0$ a.s.

Lemma D.4 (Li and Erdogdu (2020), Lemma C.7)**.** *For the density function $f(w;t)$ defined in Lemma D.3, we have for $w \leq R$ and $t \geq 0$,*

$$f(w;t) \leq C\mathrm{e}^{-2\lambda^\dagger t},$$

where $C := C(R, \lambda^\dagger, \gamma) > 0$ is a constant independent of t and w_0.

Finally, the next two theorems will be highly useful to establish the Poincaré inequality with an explicit constant.

Theorem D.1 (Bakry et al. (2008), Theorem 1.4 adapted)**.** *Suppose $\nu|_{\mathcal{Z}}$ ($\mathcal{Z} \subset \mathbb{R}^d$) satisfies the Poincaré inequality with constant $\kappa_{\mathcal{Z}}$ and there exists a Lyapunov function $V \in C^2(\mathcal{Z}')$, where $\mathcal{Z} \subset \mathcal{Z}'$. That is, $V \geq 1$ and there exist constants $\theta > 0$ and $b \geq 0$ such that*

$$\mathcal{L}V = \langle -\nabla F, \nabla V \rangle + \frac{1}{\gamma}\Delta V \leq -\theta V + b\mathbb{1}_{\mathcal{Z}}.$$

Then $\nu|_{\mathcal{Z}'}$ satisfies the Poincaré inequality with constant

$$\kappa = \frac{\theta}{1 + b/\kappa_{\mathcal{Z}}}.$$

Theorem D.2 (Li and Erdogdu (2020), Lemma B.14 adapted)**.** *Set the following neighbourhood of saddle points*

$$\mathcal{B}_r = \{x \in \mathbb{R}^d \mid d(x, \mathcal{S}) < r\}.$$

Let $r > \tilde{r} > 0$ and suppose $\nu \mid_{\mathcal{B}_{\tilde{r}}}$ satisfies the Poincaré inequality with constant $\tilde{\kappa}$ and there exist a Lyapunov function $1 \leq V \in C^2(\mathcal{B})$ and a constant $\theta > 0$ such that

$$\mathcal{L}V \leq -\theta V.$$

Then, ν satisfies the Poincaré inequality with constant κ such that

$$\frac{1}{\kappa} = \frac{4}{\theta} + \left(\frac{4}{\theta\gamma(r-\tilde{r})^2} + 2\right)\frac{1}{\tilde{\kappa}}.$$

From these theorems, we will be able to find a Poincaré constant for ν consecutively from $\mathcal{U}$ to $\mathcal{U} \cup \mathcal{A}$ and then to $\mathbb{R}^d$.

D.3 Lyapunov Function for $\mathcal{B}$

The following theorem gives a sufficient condition to find a Lyapunov function for $\mathcal{B}$ in the sense of Theorem D.2. This is actually a combination of Bovier and Den Hollander's Theorem 7.15 (Bovier and Den Hollander, 2016) and Wainwright's Theorem 2.13 (Wainwright, 2019). See Li and Erdogdu (2020) for details.

Theorem D.3 (Li and Erdogdu (2020), Proposition 9.5)**.** *If there exist constants $c_1 > 0$ and $c_2 > 0$ such that*

$$P\left(\tau_{\mathcal{B}^c} \geq t \mid X_0 = x\right) \leq c_1 \mathrm{e}^{-c_2 t}, \;\; \forall t \geq 0, \forall x \in \mathcal{B},$$

where $\tau_{\mathcal{B}^c} := \inf\{t \geq 0 \mid X_t \notin \mathcal{B}\}$, then by defining $V(x) := \mathbb{E}[\mathrm{e}^{c_2 \tau_{\mathcal{B}^c}/2} \mid X_0 = x]$, the following holds:

$$\mathcal{L}V \leq -\frac{c_2}{2}V.$$

It is thus enough to establish an exponentially decaying tail bound for $\tau_{\mathcal{B}^c}$ as shown in the next theorem.

Theorem D.4 (Li and Erdogdu (2020), Proposition 9.6 adapted)**.** *Let $\{X_t\}_{t\geq 0}$ be the Langevin diffusion defined in (1). Under Assumptions 1, 4 and 5, with $a > 0$ and $\gamma > 0$ such that*

$$\gamma \geq \max\left(a^2, 4{L'}^2 a^6\right),$$

the following holds:

$$P\left(\tau_{\mathcal{B}^c} \geq t \mid X_0 = x\right) \leq c_1 \mathrm{e}^{-\lambda^\dagger t/2}, \;\; \forall t \geq 0, \forall x \in \mathcal{B},$$

where $c_1 := c_1(a, \gamma, \lambda^\dagger)$ is a constant independent of t and x. Hence, $V(x) := \mathbb{E}[\mathrm{e}^{\theta\tau_{\mathcal{B}^c}} \mid X_0 = x]$ is a Lyapunov function on $\mathcal{B}$ in the sense of Theorem D.2 with parameter $\theta = \frac{\lambda^\dagger}{4}$.

Proof. For each $y \in \mathcal{S}$, we define v_y as the unit eigenvector of $\nabla^2 F(y)$ that corresponds to the minimum eigenvalue of $\nabla^2 F(y)$. From Assumption 4, we have that $\left\langle v_y, \nabla^2 F(y) v_y \right\rangle \leq -\lambda^\dagger$. Now, let us fix a $x \in \mathcal{B}$ and take a $y \in \mathcal{S}$ such that $\|x - y\|^2 < \frac{a^2}{\gamma}$. In the remainder of this proof, we will work in coordinates centered at this y. Without loss of generality, we can thus set $y = 0$.

Let $r(x) := \|x\|$. Then

$$\nabla r(x) = \frac{x}{\|x\|}.$$

We also define $P : \mathcal{B} \to \tilde{\mathcal{B}}$ where $\tilde{\mathcal{B}} := \left\{ t v_0 \mid |t| < \frac{a^2}{\gamma} \right\}$ such that

$$Px := \langle v_0, x \rangle v_0,$$

and $\tilde{r}(x) := |\langle v_0, x\rangle|$.[2] As a result,

$$\nabla \tilde{r}(x) = \frac{Px}{\|Px\|} = \operatorname{sign}(\langle v_0, x\rangle) v_0.$$

Using Itô's formula, we obtain

$$\begin{aligned} \mathrm{d}\left(\frac{1}{2}\tilde{r}(X_t)^2\right) = & \left(\langle -\nabla F(X_t), \tilde{r}(X_t)\nabla\tilde{r}(X_t)\rangle + \frac{1}{\gamma}\left(\|\nabla\tilde{r}(X_t)\|^2 + \tilde{r}(X_t)\Delta\tilde{r}(X_t)\right)\right)\mathrm{d}t \\ & + \frac{2}{\gamma}\langle \tilde{r}(X_t)\nabla\tilde{r}(X_t), \mathrm{d}W_t\rangle. \end{aligned}$$

Since $\nabla\tilde{r}(x)$ is a unit vector, we can consider $\langle \nabla\tilde{r}(X_t), dW_t\rangle$ as a standard one-dimensional Brownian motion independent of X_t that we denote as $\mathrm{d}B_t$.

Next, considering

$$H(x) := \nabla^2 F(0)\cdot x,$$

it follows from Lemma D.1 that

$$\|\nabla F(x) - H(x)\| \le L'\|x\|^2.$$

Therefore,

$$\begin{aligned} \langle -\nabla F(X_t), \nabla\tilde{r}(X_t)\rangle &= \langle -H(X_t), \nabla\tilde{r}(X_t)\rangle - \langle F(X_t) - \nabla H(X_t), \nabla\tilde{r}(X_t)\rangle \\ &\ge \langle -H(X_t), \nabla\tilde{r}(X_t)\rangle - \|H(X_t) - \nabla F(X_t)\|\|\nabla\tilde{r}(X_t)\| \\ &\ge \langle -H(X_t), \nabla\tilde{r}(X_t)\rangle - \|H(X_t) - \nabla F(X_t)\| \\ &\ge \langle -H(X_t), \nabla\tilde{r}(X_t)\rangle - L'\|X_t\|^2, \end{aligned}$$

where for the second inequality, we used that $\nabla\tilde{r}(x)$ is a unit vector.

Using the definition of $H(x)$, we can further write

$$\langle -H(x), \nabla\tilde{r}(x)\rangle = -\left\langle \nabla^2 F(0)\cdot x, \frac{Px}{\|Px\|}\right\rangle \ge \lambda^\dagger \tilde{r}(x),$$

where for the inequality, we used that v_0 is an eigenvector of $\nabla^2 F(0)$ and Assumption 4.

Since $\Delta\tilde{r}(x) \ge 0$ we have that $\|\nabla\tilde{r}(x)\|^2 + \tilde{r}(x)\Delta\tilde{r}(x) \ge 1$. Therefore, from Lemma D.2, $\frac{1}{2}\tilde{r}(X_t)^2$ is lower bounded by the stochastic process $\frac{1}{2}\left(r_t^{(1)}\right)^2$ defined as

$$\mathrm{d}\left(\frac{1}{2}\left(r_t^{(1)}\right)^2\right) = \left(\lambda^\dagger\left(r_t^{(1)}\right)^2 - L'\|X_t\|^2 r_t^{(1)} + \frac{1}{\gamma}\right)\mathrm{d}t + \sqrt{\frac{2}{\gamma}}r_t^{(1)}\mathrm{d}B_t.$$

Since we are only concerned with $X_t \in \mathcal{B}$, the following holds:

$$\|X_t\|^2 \le \frac{a^2}{\gamma},$$

and

$$r_t^{(1)} \le \tilde{r}(X_t) \le \sqrt{\frac{a^2}{\gamma}}.$$

We can again use Lemma D.2 to obtain a lower bound of $r_t^{(1)}$ defined as

$$\begin{aligned} \mathrm{d}\left(\frac{1}{2}\left(r_t^{(2)}\right)^2\right) &= \left(\lambda^\dagger\left(r_t^{(2)}\right)^2 - L'\frac{a^3}{\gamma^{3/2}} + \frac{1}{\gamma}\right)\mathrm{d}t + \sqrt{\frac{2}{\gamma}}r_t^{(2)}\mathrm{d}B_t \\ &= \left(\lambda^\dagger\left(r_t^{(2)}\right)^2 + \frac{1}{\gamma}\left(1 - L'\frac{a^3}{\gamma^{1/2}}\right)\right)\mathrm{d}t + \sqrt{\frac{2}{\gamma}}r_t^{(2)}\mathrm{d}B_t. \end{aligned}$$

[2] Note that, $\tilde{r}(x)$ is not differentiable for x such that $\tilde{r}(x) = 0$. For these points, we redefine $\nabla\tilde{r}(x)$ and $\Delta\tilde{r}(x)$ to be some constant $C_r > 0$, but this case can be ignored as shown later.

Since $\gamma \geq 4L'^2 a^6$, $1 - L' \frac{a^3}{\gamma^{1/2}} \geq \frac{1}{2}$. This gives us a further lower bound $r_t^{(3)}$ defined as

$$\mathrm{d}\left(\frac{1}{2}\left(r_t^{(3)}\right)^2\right) = \left(2\lambda^\dagger \frac{1}{2}\left(r_t^{(3)}\right)^2 + \frac{1}{2\gamma}\right)\mathrm{d}t + \sqrt{\frac{2}{\gamma}} r_t^{(3)} \mathrm{d}B_t.$$

This is a Cox-Ingersoll-Ross process with $W_t = \frac{1}{2}\left(r_t^{(3)}\right)^2$. Consequently, from Lemmas D.3 and D.4, when $x \leq \frac{a^2}{2\gamma}$, the density function $f(w;t)$ satisfies

$$f(w;t) \leq C\mathrm{e}^{-\lambda^\dagger t/2},$$

for all $t \geq 0$ and some constant $C := C(a, \gamma, \lambda^\dagger)$ independent of t and x. Furthermore, since $\tilde{r}(X_t)$ is lower-bounded by $r_t^{(3)}$ which is almost surely positive by Lemma D.3, the case $\tilde{r}(X_t) = 0$ can be ignored.

As a result, we obtain

$$\begin{aligned}
P\left(\tau_{\mathcal{B}^c} \geq t \mid X_0 = x\right) &= P\left(\sup_{s\in[0,t]} \frac{1}{2} r(X_s)^2 \leq \frac{a^2}{2\gamma} \mid X_0 = x\right) \\
&\leq P\left(\frac{1}{2} r(X_t)^2 \leq \frac{a^2}{2\gamma} \mid X_0 = x\right) \\
&\leq P\left(\frac{1}{2}\left(r_t^{(3)}\right)^2 \leq \frac{a^2}{2\gamma} \mid r_0^{(3)} = \tilde{r}(x)\right) \\
&= \int_0^{\frac{a^2}{2\gamma}} f(w;t) dw \\
&\leq \frac{a^2}{2\gamma} C\mathrm{e}^{-\lambda^\dagger t/2},
\end{aligned}$$

where for the second equality, we used Lemma D.3 and for the last inequality we used Lemma D.4. This gives the desired result.

Q.E.D

D.4 Lyapunov Function for $\mathcal{A}$

In this section, we prepare statements for the Lyapunov function on $\mathcal{A}$.

Lemma D.5 (Li and Erdogdu (2020), Lemma 9.7 adapted)**.** *Under Assumptions 1,3, 4 and 5, with* $\mathcal{C} := \mathcal{S} \cup \mathcal{X}$, *there exists a constant* $0 < C_F \leq 1$ *such that*

$$\|\nabla F(x)\| \geq C_F d(x, \mathcal{C}),$$

where

$$C_F := \min\left(1, \frac{\lambda^\dagger}{2}, \inf_{x:d(x,\mathcal{C}) > \frac{\lambda^\dagger}{4L'}} \frac{\|\nabla F(x)\|}{d(x,\mathcal{C})}\right).$$

Proof. First, observe that when F is (M, b)-dissipative, we have

$$\frac{1}{M}\|\nabla F(x)\|^2 + \frac{M}{2}\|x\|^2 \geq M\|x\|^2 - b,$$

which leads to

$$\frac{\|\nabla F(x)\|}{\|x\|} \geq \sqrt{\frac{M^2}{2} - Mb/\|x\|^2}.$$

We obtain thus

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla F(x)\|}{\|x\|} \geq \sqrt{\frac{M^2}{2}}. \tag{15}$$

Therefore,

$$\inf_{x:d(x,\mathcal{C})>\frac{\lambda^\dagger}{4L'}} \frac{\|\nabla F(x)\|}{d(x,\mathcal{C})} > 0,$$

since by equation (15) $\|\nabla F(x)\|/d(x,\mathcal{C}) > 0$ holds outside a compact set of x around the origin and since in this compact set and away from stationary points there exist an x that minimizes $\|\nabla F(x)\|/d(x,\mathcal{C})$ and that cannot be 0 as we are outside $\mathcal{C}$. As a result, we just have to consider the case when $d(x,\mathcal{C}) \le \frac{\lambda^\dagger}{4L'}$.

Let y be a stationary point such that $\|x-y\| < 2d(x,\mathcal{C}) \le \frac{\lambda^\dagger}{2L'}$. Since $\nabla^2 F(x)$ is L'-Lipschitz,

$$\|\nabla F(x) - \nabla F(y) - \nabla^2 F(y)^\top (x-y)\| \le \frac{L'}{2}\|x-y\|^2,$$

and we get

$$\begin{aligned}\|\nabla F(x)\| \ge& \|\nabla^2 F(y)(x-y)\| - \|\nabla F(x) - \nabla^2 F(y)(x-y)\|\\ \ge& \lambda^\dagger \|x-y\| - \frac{L'}{2}\|x-y\|^2\\ \ge& \frac{\lambda^\dagger}{2}\|x-y\|.\end{aligned}$$

We obtain thus the desired result with $C_F > 0$ by taking the minimum of the two constants.

Q.E.D

Lemma D.6 (Li and Erdogdu (2020), Lemma 9.8 adapted)**.** *Under Assumptions 1, 3, 4 and 5, for $a > 0$ and $\gamma > 0$ such that*

$$a^2 \ge \frac{24dL}{C_F^2},$$

the following holds:

$$\frac{\Delta F(x)}{2} - \frac{\gamma}{4}\|\nabla F(x)\|^2 \le -dL,\ \forall x \in \mathbb{R}^d : d(x,\mathcal{C}) \ge \frac{a^2}{4\gamma}.$$

Proof. From Lemma D.5 and smoothness of F, we have

$$\frac{\Delta F(x)}{2} - \frac{\gamma}{4}\|\nabla F(x)\|^2 \le \frac{dL}{2} - \frac{\gamma}{4}C_F^2 d(x,\mathcal{C}).$$

Since $d(x,\mathcal{C}) \ge \frac{a^2}{\gamma}$ and $a^2 \ge \frac{24dL}{C_F^2}$, we obtain

$$\begin{aligned}\frac{\Delta F(x)}{2} - \frac{\gamma}{4}\|\nabla F(x)\|^2 &\le \frac{dL}{2} - \frac{\gamma}{4}C_F^2 d(x,\mathcal{C})\\ &\le \frac{dL}{2} - \frac{\gamma}{4}C_F^2\frac{a^2}{\gamma}\\ &\le -dL.\end{aligned}$$

Q.E.D

D.5 Poincaré Inequality

In this section, we establish the Poincaré inequality for ν using Theorem D.1 and D.2. It is easy to find a Poincaré constant for $\nu|_{\mathcal{U}}$ which is our starting point.

Lemma D.7 (Li and Erdogdu (2020), Lemma 9.9 adapted)**.** *Under Assumptions 1, 3 and 4 to 6, with $a > 0$ and $\gamma > 0$ such that*

$$\gamma \ge a^2 \frac{4dL'^2}{\lambda^{\dagger 2}},$$

$\nu|_{\mathcal{U}}$ satisfies the Poincaré inequality with constant $\kappa_{\mathcal{U}} = \frac{\lambda^\dagger}{2}$.

Proof. Let x^* be the global minimum. Then $\mathcal{U} = \{x \in \mathbb{R}^d \mid \|x - x^*\|^2 < \frac{a^2}{\gamma}\}$. Using the same idea as Lemma D.5, we have

$$\min_{i \in \{1,\ldots,n\}} \lambda_i \left(\nabla^2 F(x)\right) \geq \lambda^\dagger - dL' \|x - x^*\| \geq \frac{\lambda^\dagger}{2},$$

where we used $\gamma \geq a^2 \frac{4dL'^2}{\lambda^{\dagger 2}}$ and $\|x - x^*\| \leq \sqrt{a^2 \gamma}$.

This implies for all $x \in \mathcal{U}$,

$$\nabla^2 F(x) \geq \frac{\lambda^\dagger}{2} I_{d \times d},$$

where $I_{d\times d}$ is the $d \times d$ unit matrix. Therefore, $\nu|_{\mathcal{U}}$ satisfies the Poincaré inequality with constant $\kappa_{\mathcal{U}} = \frac{\lambda^\dagger}{2}$.

Q.E.D

Next, we show that $\nu|_{\mathcal{U}\cup\mathcal{A}}$ satisfies the Poincaré inequality.

Lemma D.8 (Li and Erdogdu (2020), Lemma 9.11 adapted)**.** *Under Assumptions 1, 3 and 4 to 6, with $a > 0$ and $\gamma > 0$ such that*

$$a^2 \geq \frac{24dL}{C_F^2},$$

and

$$\gamma \geq a^2 \frac{4dL'^2}{\lambda^{\dagger 2}},$$

$\nu|_{\mathcal{U}\cup\mathcal{A}}$ satisfies the Poincaré inequality with constant

$$\kappa_{\mathcal{U}\cup\mathcal{A}} = \frac{1}{1 + 3/(2\kappa_{\mathcal{U}})}.$$

Proof. Let us choose the candidate Lyapunov function $V_1(x) = \mathrm{e}^{\frac{\gamma}{2}F(x)}$. Then,

$$\frac{\mathcal{L}V_1}{V_1} = \frac{1}{2}\Delta F - \frac{\gamma}{4}\|\nabla F\|^2.$$

From Lemma D.6, for all $x \in \mathcal{A}$ we have

$$\frac{\mathcal{L}V_1}{V_1} \leq -dL.$$

On the other hand, for all $x \in \mathcal{U}$ we obtain

$$\frac{\mathcal{L}V_1}{V_1} = \frac{1}{2}\Delta F \leq \frac{1}{2}dL.$$

This leads to

$$\frac{\mathcal{L}V_1}{V_1} \leq -dL + \frac{3dL}{2}\mathbb{1}_{\mathcal{U}}$$

for all $x \in \mathcal{U} \cup \mathcal{A}$. Since the assumptions of Theorem D.1 are satisfied, we conclude that $\nu|_{\mathcal{U}\cup\mathcal{A}}$ satisfies the Poincaré inequality with a constant

$$\kappa_{\mathcal{U}\cup\mathcal{A}} = \frac{dL}{1 + 3dL/(2\kappa_{\mathcal{U}})}.$$

Since $L \geq 1$ and $d \geq 1$, we can replace this value by its lower bound

$$\kappa_{\mathcal{U}\cup\mathcal{A}} = \frac{1}{1 + 3/(2\kappa_{\mathcal{U}})}.$$

Q.E.D

Finally, we can establish the Poincaré inequality for ν.

Lemma D.9 (Li and Erdogdu (2020), Proposition 9.12 adapted)**.** *Under Assumptions 1, 3 and 4 to 6, with $a > 0$ and $\gamma > 0$ such that*

$$a^2 \geq \frac{24dL}{C_F^2},$$

and

$$\gamma \geq \max\left(a^2 \frac{4dL'^2}{\lambda^{\dagger 2}}, 4L'^2 a^6\right),$$

ν satisfies the Poincaré inequality with constant

$$\kappa = \frac{\lambda^\dagger}{35}.$$

Proof. Let us select the candidate Lyapunov function

$$V_2(x) = \mathbb{E}[\mathrm{e}^{\lambda^\dagger \tau_{\mathcal{B}^c}/4} | X_0 = x],$$

where $\tau_{\mathcal{B}^c} = \inf\{t \geq 0 | X_t \notin \mathcal{B}\}$.

From Theorem D.3 and D.4, we have that V_2 satisfies the Lyapunov condition with

$$\frac{\mathcal{L}V_2}{V_2} \leq -\frac{\lambda^\dagger}{4}.$$

Now, using Theorem D.2 with $\theta = \frac{\lambda^\dagger}{4}$, $\tilde{r} = \frac{1}{2} r = \frac{a}{2\sqrt{\gamma}}$, we conclude that ν satisfies the Poincaré inequality with a constant κ such that

$$\frac{1}{\kappa} = \frac{16}{\lambda^\dagger} + \left(\frac{16}{\lambda^\dagger \gamma}\frac{4\gamma}{a^2} + 2\right)\frac{1}{\kappa_{\mathcal{U}\cup\mathcal{A}}}.$$

Since $C_F \leq 1$, $d \geq 1$ and $L \geq 1$, we have $\frac{1}{a^2} \leq \frac{C_F^2}{24dL} \leq \frac{1}{24}$, which leads to

$$\frac{1}{\kappa} \leq \frac{16}{\lambda^\dagger} + \left(\frac{8}{3\lambda^\dagger} + 2\right)\frac{1}{\kappa_{\mathcal{U}\cup\mathcal{A}}}.$$

Plugging $1/\kappa_{\mathcal{U}\cup\mathcal{A}} = 1 + 3/(2\kappa_{\mathcal{U}}) = 1 + 3/\lambda^\dagger$ from Lemma D.7 and D.8, we obtain

$$\begin{aligned}\frac{1}{\kappa} &\leq \frac{16}{\lambda^\dagger} + \left(\frac{8}{3\lambda^\dagger} + 2\right)\left(1 + \frac{3}{\lambda^\dagger}\right)\\ &\leq \frac{35}{\lambda^\dagger},\end{aligned}$$

where we used $\frac{1}{\lambda^\dagger} \geq 1$ in the last inequality. *Q.E.D*

D.6 Log-Sobolev Inequality

Finally, we can establish the Log-Sobolev inequality thanks to the following theorem.

Theorem D.5 (Cattiaux et al. (2010))**.** *Suppose the following conditions hold for the generator defined in Definition D.4.*

1. *There exist constants $\theta > 0$ and $b > 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that for all $x \in \mathbb{R}^d$*

$$\frac{\gamma \mathcal{L}V(x)}{V(x)} \leq -\theta \|x\|^2 + b.$$

2. *ν satisfies the Poincaré inequality with a constant κ.*

3. *There exists some constant $K > 0$, such that $\nabla^2 F \succeq -LI_{d\times d}$.*

Then, ν satisfies the Log-Sobolev inequality with a constant α such that

$$\frac{1}{\alpha} = C_1 + (C_2 + 2)\frac{1}{\kappa},$$

where

$$C_1 = \frac{2\gamma L}{\theta} + \frac{2}{\gamma K},$$

and

$$C_2 = \frac{2\gamma L}{\theta}\left(b + \theta\int_{\mathbb{R}^d}\|x\|^2 \mathrm{d}\nu\right).$$

Theorem D.6. *Under Assumptions 1, 3 and 4 to 6, with $a > 0$ and $\gamma \geq 1$ such that*

$$a^2 \geq \frac{24dL}{C_F^2},$$

and

$$\gamma \geq \max\left(1, a^2\frac{4dL'^2}{\lambda^{\dagger 2}}, 4L'^2 a^6\right),$$

ν satisfies the Log-Sobolev inequality with constant α such that

$$\frac{1}{\alpha} = \left(\frac{2M^2 + 8L^2}{M^2 L} + \left(\frac{6L(d+1))}{M} + 2\right)\frac{35}{\lambda^\dagger}\right)\gamma.$$

Proof. Let us consider the candidate Lyapunov function $V(x) = \mathrm{e}^{M\gamma\|x\|^2/4}$. Then from $V \geq 1$ and Assumption 3, we obtain

$$\begin{aligned}\gamma\mathcal{L}V(x) &= \left(\frac{M\gamma d}{2} + \frac{M^2\gamma^2}{4}\|x\|^2 - \frac{M\gamma^2}{2}\langle x, \nabla F(x)\rangle\right)V(x)\\ &\leq \left(\frac{M\gamma(d + b\gamma)}{2} - \frac{M^2\gamma^2}{4}\|x\|^2\right)V(x).\end{aligned}$$

Under

$$a^2 \geq \frac{24dL}{C_F^2},$$

and

$$\gamma \geq \max\left(a^2\frac{4dL'^2}{\lambda^{\dagger 2}}, 4L'^2 a^6\right),$$

ν satisfies the Poincaré inequality with a constant $\frac{\lambda*}{35}$. Moreover, since F is L-smooth, $\nabla^2 F \succeq -LI_{d\times d}$. Therefore, all the conditions of Theorem D.5 are satisfied.

We conclude that from Theorem D.5, ν satisfies the Log-Sobolev inequality with a constant α such that

$$\frac{1}{\alpha} \leq C_1 + (C_2 + 2)\frac{35}{\lambda^\dagger},$$

where constants C_1 and C_2 can be calculated as

$$C_1 = \frac{2M^2 + 8L^2}{M^2 L\gamma},$$

and

$$C_2 \leq \frac{6L(d+\gamma)}{M}$$

from Raginsky et al. (2017). We can replace this value by a simple upper bound which gives us

$$\frac{1}{\alpha} \leq \left(\frac{2M^2 + 8L^2}{M^2 L} + \left(\frac{6L(d+1))}{M} + 2\right)\frac{35}{\lambda^\dagger}\right)\gamma$$

since $\gamma \geq 1$.

$Q.E.D$

Remark D.1. *Once we obtain the Poincaré constant, they are several ways to construct the Log-Sobolev constant. Another approach is possible, maybe simpler, by proceeding as Li and Erdogdu (2020) did in their analysis. Even though their method is interesting, this should not seriously change our main point since we just wanted to show that a polynomial dependence of the Log-Solev constant on the inverse temperature was achievable under certain additional conditions.*

E Analysis of an annealing scheme

In this Appendix, we prove the global convergence of SVRG-LD and SARAH-LD combined with an annealed scheme.

E.1 Algorithm

In the context of optimization, we can use Algorithm 1 by setting a γ huge enough so that the stationary distribution concentrates on the global minimizer of F. On the other hand, we can also introduce to SVRG-LD and SARAH-LD an increasing inverse temperature and a decreasing step size as follows.

Algorithm 2: SVRG-LD / SARAH-LD with annealing

```
input: batch size $B$, epoch length $m$
annealing schedule: step size $\eta_s > 0$ and inverse temperature $\gamma_s \geq 1$
initialization: $X_0 = 0$, $X^{(0)} = X_0$
foreach $s = 0, 1, \dots, (K/m)$ do
    $v_{sm} = \nabla F(X^{(s)})$
    randomly draw $\epsilon_{sm} \sim N(0, I_{d\times d})$
    $X_{sm+1} = X_{sm} - \eta_s v_{sm} + \sqrt{2\eta_s/\gamma_s}\epsilon_{sm}$
    foreach $l = 1, \dots, m-1$ do
        $k = sm + l$
        randomly pick a subset $I_k$ from $\{1, \dots, n\}$ of size $|I_k| = B$
        randomly draw $\epsilon_k \sim N(0, I_{d\times d})$
        if SVRG-LD then
            $v_k = \frac{1}{B}\sum_{i_k \in I_k}(\nabla f_{i_k}(X_k) - \nabla f_{i_k}(X^{(s)})) + v_{sm}$
        else if SARAH-LD then
            $v_k = \frac{1}{B}\sum_{i_k \in I_k}(\nabla f_{i_k}(X_k) - \nabla f_{i_k}(X_{k-1})) + v_{k-1}$
        end
        $X_{k+1} = X_k - \eta_s v_k + \sqrt{2\eta_s/\gamma_s}\epsilon_k$
    end
    $X^{(s+1)} = X_{(s+1)m}$
end
```

Definition 2. *We define ψ_k as the distribution of X_k generated at the kth step of Algorithm 2.*

E.2 Preparation for the Proof

Let us first establish some special notations to keep the proof clear and simple.

Notation E.1. *We define ν_{γ_k} as the stationary Gibbs distribution of SDE (1) when the inverse temperature parameter is set at γ_k, namely,*

$$\nu_{\gamma_k} := \mathrm{e}^{-\gamma_k F}/Z_{\gamma_k},$$

where Z_{γ_k} is the normalizing constant, and α_k as the Log-Sobolev constant of ν_{γ_k} under Assumptions 1 and 3. We also abbreviate the KL divergence between the distribution ψ_{sm+r} of the random variable X_{sm+r} generated by Algorithm 3 and the Gibbs distribution ν_{γ_s} as follows, where $s \in \mathbb{N} \cup \{0\}$ and $r = 1, \dots, m$:

$$H_{sm+r} := H_{\nu_{\gamma_s}}(\psi_{sm+r}).$$

Moreover, $H_0 := H_{\nu_{\gamma_0}}(\psi_0)$.

We will also need the following technical lemma.

Lemma E.1. *For all $s \in \mathbb{N} \cup \{0\}$, $\sigma \geq 3$ and $\mu > 2$,*

$$\left(\frac{2}{3}\right)^{\frac{2}{\mu}} (s+1)^{1-\frac{2}{\mu}} \sigma^{-\frac{2}{\mu}} \leq \sum_{i=0}^{s} (i+\sigma)^{-\frac{2}{\mu}},$$

where C_μ is a constant independent of s and σ.

Proof. By a simple argument of area under the curve $y = x^{-\frac{2}{\mu}}$,

$$\sum_{i=0}^{s}(i+\sigma)^{-\frac{2}{\mu}} \geq \int_{\sigma}^{s+\sigma+1} x^{-\frac{2}{\mu}}\mathrm{d}x.$$

According to the mean value theorem for integrals, there exist a constant $c_s \in [\sigma, s+\sigma+1]$ such that

$$\sum_{i=0}^{s}(i+\sigma)^{-\frac{2}{\mu}} \geq \int_{\sigma}^{s+\sigma+1} x^{-\frac{2}{\mu}}\mathrm{d}x = c_s^{-\frac{2}{\mu}}(s+1).$$

We have also

$$\begin{aligned} c_s^{-\frac{2}{\mu}} &\geq (s+1+\sigma)^{-\frac{2}{\mu}} \\ &= (s+1)^{-\frac{2}{\mu}}(1+\frac{\sigma}{s+1})^{-\frac{2}{\mu}} \\ &\geq (s+1)^{-\frac{2}{\mu}}(1+\sigma)^{-\frac{2}{\mu}} \\ &= (s+1)^{-\frac{2}{\mu}}\sigma^{-\frac{2}{\mu}}\left(\frac{1+\sigma}{\sigma}\right)^{-\frac{2}{\mu}} \\ &\geq (s+1)^{-\frac{2}{\mu}}\sigma^{-\frac{2}{\mu}}\left(\frac{2}{3}\right)^{\frac{2}{\mu}}. \end{aligned}$$

In the last inequality, we used $\sigma \geq 2$. This implies the inequality of the statement.

$Q.E.D$

Now, considering that we only change the step size and the inverse temperature parameter at the beginning of every inner loop, all statements proved in Appendix A (Lemmas A.3 and A.4) and in Appendix B (Lemmas B.1, B.2 and B.3) that consider only the inner loop hold for Algorithm 2 as well.

Moreover, let us consider the annealing schedule

$$\eta_s = \bar{\eta}(s+\sigma)^{-\frac{1}{\mu}}, \tag{16}$$

$$\gamma_s = \bar{\gamma}\log\left\{g(s+\sigma)^{\frac{1}{\mu}}\right\}, \tag{17}$$

where we suppose $\bar{\eta} > 0$, $\bar{\gamma} > 0$, $\sigma \geq 3$, $g \geq \mathrm{e}$ and $\mu > 2$. This annealing schedule is chosen on the one hand so that $\sum_{i=0}^{s}\frac{\alpha_i}{\gamma_i}\eta_i$ is explicitly computable, and on the other hand because Chiang et al. (1987) showed that the annealed continuous time GLD

$$\mathrm{d}X_t^{\mathrm{Ann}} = \nabla F(X_t^{\mathrm{Ann}})\mathrm{d}t + \sqrt{T(t)}\mathrm{d}B_t$$

could find the global minimum with the annealing schedule $T(t) \propto \frac{1}{\log t}$, which corresponds to equation (17).

Then, the following theorem holds under this annealing schedule.

Theorem E.1. *With the annealing schedule* (16) *and* (17)*, under Assumptions 1 and 3,* $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, $\bar{\gamma} = \frac{1}{C_2}$, $\mu > 2$, $g \geq \mathrm{e}$, *and* $B \geq m$, *for all* $k = sm + r$ *where* $s \in \mathbb{N}\cup\{0\}$ *and* $r = 0, \dots, m-1$, *the following holds in the update of Algorithm 2:*

$$\begin{aligned} H_{\nu_{\gamma_s}}(\psi_{sm+r+1}) \leq{}& \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1+\frac{32\gamma_s L^4\eta_s^3}{\alpha_s}\right)H_{\nu_{\gamma_s}}(\psi_{sm+r}) \\ &+ \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\sum_{i=0}^{r-1}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s}H_{\nu_{\gamma_s}}(\psi_{sm+i}) \\ &+ 56\eta_s^2 dL^2. \end{aligned}$$

Here, $C = \frac{(n-B)}{B(n-1)}$.

Proof. From Property C.3, ν_{γ_s} satisfies Log-Sobolev inequality with a constant α_s such that $\frac{\alpha_s}{\gamma_s} = C_1 \mathrm{e}^{-C_2\gamma_s}$. It thus suffices to notice that under $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$ and $\bar{\gamma} = \frac{1}{C_2}$, we have for all $s \in \mathbb{N} \cup \{0\}$,

$$\begin{aligned}\eta_s &= \bar{\eta}(s+\sigma)^{-\frac{1}{\mu}} \\ &\leq \frac{C_1}{16\sqrt{6}g(s+\sigma)^{\frac{1}{\mu}}L^2m} \\ &= \frac{\alpha_s}{16\sqrt{6}\gamma_s L^2 m}.\end{aligned}$$

In the inequality, we used the fact that with $\bar{\gamma} = \frac{1}{C_2}$,

$$\begin{aligned}\frac{\alpha_s}{\gamma_s} &= C_1\mathrm{e}^{-C_2\gamma_s} \\ &= C_1\mathrm{e}^{-C_2\bar{\gamma}\log\left\{g(s+\sigma)^{\frac{1}{\mu}}\right\}} \\ &= \frac{C_1}{g(s+\sigma)^{\frac{1}{\mu}}}.\end{aligned}$$

Therefore, all the assumptions of Theorem A.1 and B.1 are satisfied. From the proof of each theorem, we immediately obtain the inequality of the statement from equations (9) and (14).

Q.E.D

The problem with changing the inverse temperature parameter of each inner loop is that we cannot immediately give an upper bound for each H_k as Theorem A.2 and B.2. The main challenge resides in linking $H_{\nu_{\gamma_s}}(\psi_{sm})$ and $H_{\nu_{\gamma_{s-1}}}(\psi_{sm})$, which corresponds to the shift of optimization trajectory in the space of measures generated by the change of inverse temperature parameter at the beginning of each inner loop. The following lemma suggests that a small enough difference between two consecutive inverse temperatures will solve this issue.

Lemma E.2. *Under Assumptions 1, 3 and $F \geq 0$, for all $s \in \mathbb{N}$ and $\gamma_0 \geq \frac{2}{M}$,*

$$H_{\nu_{\gamma_s}}(\psi_{sm}) \leq \left(1 + \Delta\gamma_s \frac{2L}{\alpha_{s-1}}\right) H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \Delta\gamma_s\left(\chi + F(X^*)\right).$$

Here, $\Delta\gamma_s := \gamma_s - \gamma_{s-1}$, $\chi := \max_{\gamma\geq 1}\left\{\frac{d}{\gamma}\log\left(\frac{\mathrm{e}L}{M}\left(\frac{b\gamma}{d}+1\right)\right)\right\}$ and X^* is the global minimum of F.

Proof.

$$\begin{aligned}H_{\nu_{\gamma_s}}(\psi_{sm}) &= H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + H_{\nu_{\gamma_s}}(\psi_{sm}) - H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\ &= H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \int \psi_{sm}\log\frac{\psi_{sm}}{\nu_{\gamma_s}}\mathrm{d}z - \int \psi_{sm}\log\frac{\psi_{sm}}{\nu_{\gamma_{s-1}}}\mathrm{d}z \\ &= H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \int \psi_{sm}\log\frac{\nu_{\gamma_{s-1}}}{\nu_{\gamma_s}}\mathrm{d}z \\ &= H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \int \psi_{sm}\log\frac{\mathrm{e}^{-\gamma_{s-1}F}/Z_{\gamma_{s-1}}}{\mathrm{e}^{-\gamma_s F}/Z_{\gamma_s}}\mathrm{d}z \\ &= H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \int \psi_{sm}(\gamma_s - \gamma_{s-1})F\mathrm{d}z + \log\frac{Z_{\gamma_s}}{Z_{\gamma_{s-1}}}.\end{aligned}$$

Here, as $\gamma_s \geq \gamma_{s-1}$and $F \geq 0$, we have that

$$-\gamma_s F \leq -\gamma_{s-1}F,$$

which means

$$Z_{\gamma_s} \leq Z_{\gamma_{s-1}}.$$

Thus,

$$H_{\nu_{\gamma_s}}(\psi_{sm}) \le H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \Delta\gamma_s \mathbb{E}_{X\sim\psi_{sm}}[F(X)].$$

Now, from Corollary C.1.1 and Theorem C.2, we know that

$$\begin{aligned}
\mathbb{E}_{X\sim\psi_{sm}}[F(X)] &= \mathbb{E}_{X\sim\psi_{sm}}[F(X)] - F(X^*) + F(X^*) \\
&\le LW_2^2(\psi_k, \nu_{\gamma_{s-1}}) + 2\left(\mathbb{E}_{X\sim\nu_{\gamma_{s-1}}}[F(X)] - F(X^*)\right) + F(X^*) \\
&\le LW_2^2(\psi_k, \nu_{\gamma_{s-1}}) + \frac{d}{\gamma_{s-1}}\log\left(\frac{\mathrm{e}L}{M}\left(\frac{b\gamma_{s-1}}{d}+1\right)\right) + F(X^*) \\
&\le \frac{2L}{\alpha_{s-1}} H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) + \chi + F(X^*).
\end{aligned}$$

We used Corollary C.1.1 at the first inequality, Theorem C.2 at the second inequality and Talagrand's inequality at the last inequality. This gives the desired result.

$Q.E.D$

Since the logarithmic function $\log x$ is strictly increasing while its derivative decreases according to x, we can find an adequate bound of σ to assure that $\Delta\gamma_s$ is small enough. As a reminder, we set $\gamma_s = \bar\gamma \log\left\{g(s+\sigma)^{\frac{1}{\mu}}\right\}$ and $\eta_s = \bar\eta(s+\sigma)^{\frac{1}{\mu}}$.

Lemma E.3. *With the annealing* (16) *and* (17)*, when* $\alpha_s = \gamma_s C_1 \mathrm{e}^{-C_2\gamma_s}$,

$$\sigma \ge 3 \vee \left(\frac{8Lg^2}{C_1^2\bar\eta}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar\eta^2}\right)^{\frac{\mu}{\mu-2}},$$

$\bar\gamma = \frac{1}{C_2}$, $\mu > 3$ *and* $g \ge \mathrm{e}$*, we have*

$$\Delta\gamma_s \frac{2L}{\alpha_{s-1}} \le \frac{\alpha_s\eta_s}{2\gamma_s}, \tag{18}$$

and

$$\Delta\gamma_s \le \eta_s^2 L^2 \le \frac{1}{4} \tag{19}$$

for all $s \in \mathbb{N}\cup\{0\}$.

Proof. First of all, by the mean value theorem, there exists a $c \in [s-1, s]$ such that,

$$\Delta\gamma_s = \frac{\bar\gamma/\mu}{c+\sigma}.$$

Thus,

$$\Delta\gamma_s = \frac{\bar\gamma/\mu}{c+\sigma} \le \frac{\bar\gamma/\mu}{s-1+\sigma} = \frac{1}{\mu C_2}\frac{1}{(s-1+\sigma)}.$$

Therefore, in order to satisfy inequality (18), it suffices to have

$$\begin{aligned}
\frac{1}{\mu C_2}\frac{1}{(s-1+\sigma)} &\le \frac{\alpha_{s-1}}{2L}\frac{\alpha_s\eta_s}{2\gamma_s} \\
&= \frac{\gamma_{s-1}C_1\mathrm{e}^{-C_2\gamma_{s-1}}}{2L}\frac{1}{2}C_1\mathrm{e}^{-C_2\gamma_s}\eta_s.
\end{aligned}$$

A sufficient condition to this is

$$\frac{1}{\mu C_2}\frac{1}{(s-1+\sigma)} \le \frac{C_1 C_2^{-1}\log\left\{g(s-1+\sigma)^{\frac{1}{\mu}}\right\}}{2Lg(s+\sigma)^{\frac{1}{\mu}}}\frac{\bar\eta C_1}{2g(s+\sigma)^{\frac{2}{\mu}}},$$

which gives

$$\begin{aligned}\frac{4Lg^2}{C_1^2\bar{\eta}} &\le \frac{s+\sigma-1}{(s+\sigma)^{\frac{3}{\mu}}}\log\{g(s-1+\sigma)\} \\ &= (s+\sigma)^{1-\frac{3}{\mu}}\frac{s+\sigma-1}{s+\sigma}\log\{g(s-1+\sigma)\} \\ &= (s+\sigma)^{1-\frac{3}{\mu}}\left(1-\frac{1}{s+\sigma}\right)\log\{g(s-1+\sigma)\}.\end{aligned}$$

As $\log\{g(s-1+\sigma)\} \ge 1$ and $1-\frac{1}{s+\sigma} \ge \frac{1}{2}$ when $g \ge \mathrm{e}$, $s \ge 0$ and $\sigma \ge 2$, it suffices to have the following inequality satisfied when $s = 0$:

$$\frac{8L}{C_1^2\bar{\eta}} \le (s+\sigma)^{1-\frac{3}{\mu}}.$$

From this, we obtain $\sigma \ge \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}}$.

Likewise, in order to satisfy inequality (19), it suffices to have

$$\frac{1}{\mu C_2}\frac{1}{(s-1+\sigma)} \le \frac{\bar{\eta}^2L^2}{(s+\sigma)^{\frac{2}{\mu}}},$$

which gives

$$\frac{1}{\mu C_2L^2\bar{\eta}^2} \le \frac{s+\sigma-1}{(s+\sigma)^{\frac{2}{\mu}}}.$$

It thus suffices to have the following inequality satisfied when $s = 0$:

$$\frac{2}{\mu C_2L^2\bar{\eta}^2} \le (s+\sigma)^{1-\frac{2}{\mu}}.$$

This gives $\sigma \ge \left(\frac{2}{\mu C_2L^2\bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}}$.

The last inequality $\eta_s^2L^2 \le \frac{1}{2}$ is immediately satisfied with $\eta_s \le \bar{\eta} \le \frac{1}{4L}$.

$Q.E.D$

E.3 Main Proof

We are now ready to prove the main results. We first evaluate how H_k decreases compared with the previous step.

Theorem E.2. *With the annealing schedule* (16) *and* (17), *under Assumptions 1, 3 and* $F \ge 0$, $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, $\bar{\gamma} = \frac{1}{C_2}$, $B \ge m$, $\mu > 3$, $g \ge \mathrm{e}$ *and*

$$\sigma \ge 3 \vee \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2L^2\bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}},$$

for all $k = sm + r$ *where* $s \in \mathbb{N}\cup\{0\}$ *and* $r = 0, \dots, m-1$, *the following holds in the update of Algorithm 2:*

$$\begin{aligned}H_{sm+r+1} \le{}& \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1+\frac{\alpha_s}{4\gamma_s}\eta_s\right)H_{sm+r} \\ &+ \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\sum_{i=0}^{r-1}\frac{\alpha_s}{4m\gamma_s}\eta_s\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s}H_{sm+i} \\ &+ \eta_s^2dL^2E.\end{aligned}$$

Here, $E = 56 + 2\chi + 2F(X^*)$

Proof. When $r = 0$, from Theorem E.1, we have

$$H_{\nu_{\gamma_s}}(\psi_{sm+1}) \le \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_s}}(\psi_{sm}) \\ + 56\eta_s^2 d L^2.$$

Under

$$\sigma \ge 3 \vee \left(\frac{8Lg^2}{C_1^2 \bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}},$$

we can derive the following bound:

$$\begin{aligned}
H_{\nu_{\gamma_s}}(\psi_{sm+1}) &\le \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right)\left(1 + \Delta\gamma_s \frac{2L}{\alpha_{s-1}}\right) H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\
&\quad + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right)\Delta\gamma_s\left(\chi + F(X^*)\right) \\
&\quad + 56\eta_s^2 d L^2 \\
&\le \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right)\left(1 + \frac{\alpha_s}{2\gamma_s}\eta_s\right) H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\
&\quad + \left(1 + 2L^2\eta_s^2\right)\eta_s^2 L^2\left(\chi + F(X^*)\right) \\
&\quad + 56\eta_s^2 d L^2 \\
&\le \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right)\mathrm{e}^{\frac{\alpha_s}{2\gamma_s}\eta_s} H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\
&\quad + (1 + 1)\,\eta_s^2 L^2\left(\chi + F(X^*)\right) \\
&\quad + 56\eta_s^2 d L^2 \\
&\le \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\
&\quad + \eta_s^2 d L^2\left(56 + 2\chi + 2F(X^*)\right).
\end{aligned}$$

We used Lemma E.2 at the first inequality, Lemma E.3 and $\eta_s \le \frac{\alpha_s}{16\gamma_s L^2 m}$ at the second inequality and $\eta_s^2 L^2 \le \frac{1}{2}$ at the last inequality.

When $r \ge 1$, from Theorem E.1, we have

$$\begin{aligned}
H_{\nu_{\gamma_s}}(\psi_{sm+r+1}) \le \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}&\left(1 + \frac{32\gamma_s L^4 \eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_s}}(\psi_{sm+r}) \\
&+ \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\sum_{i=0}^{r-1}\frac{128\gamma_s L^4 \eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{\nu_{\gamma_s}}(\psi_{sm+i}) \\
&+ 56\eta_s^2 d L^2.
\end{aligned}$$

Under

$$\sigma \ge 3 \vee \left(\frac{8Lg^2}{C_1^2 \bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}},$$

$$\begin{aligned}
H_{\nu_{\gamma_s}}(\psi_{sm+r+1}) \leq{} & \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1+\frac{32\gamma_s L^4\eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_s}}(\psi_{sm+r}) \\
& + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\sum_{i=1}^{r-1}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{\nu_{\gamma_s}}(\psi_{sm+i}) \\
& + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{\nu_{\gamma_s}}(\psi_{sm}) \\
& + 56\eta_s^2 dL^2 \\
\leq{} & \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1+\frac{32\gamma_s L^4\eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_s}}(\psi_{sm+r}) \\
& + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\sum_{i=1}^{r-1}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{\nu_{\gamma_s}}(\psi_{sm+i}) \\
& + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s}\left(1+\Delta\gamma_s\frac{2L}{\alpha_{s-1}}\right) H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\
& + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s}\Delta\gamma_s\left(\chi+F(X^*)\right) \\
& + 56\eta_s^2 dL^2 \\
\leq{} & \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\left(1+\frac{32\gamma_s L^4\eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_s}}(\psi_{sm+r}) \\
& + \mathrm{e}^{-\frac{3\alpha_s}{2\gamma_s}\eta_s}\sum_{i=1}^{r-1}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{\nu_{\gamma_s}}(\psi_{sm+i}) \\
& + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{\nu_{\gamma_{s-1}}}(\psi_{sm}) \\
& + 2\eta_s^2 L^2\Delta\gamma_s\left(\chi+F(X^*)\right)+56\eta_s^2 dL^2 \\
\leq{} & \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1+\frac{32\gamma_s L^4\eta_s^3}{\alpha_s}\right) H_{\nu_{\gamma_s}}(\psi_{sm+r}) \\
& + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\sum_{i=0}^{r-1}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{sm+i} \\
& + \eta_s^2 dL^2\left(56+2\chi+2F(X^*)\right).
\end{aligned}$$

Therefore, for all $r = 0, \dots, m-1$,

$$\begin{aligned}
H_{sm+r+1} \leq{} & \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1+\frac{32\gamma_s L^4\eta_s^3}{\alpha_s}\right) H_{sm+r} \\
& + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\sum_{i=0}^{r-1}\frac{128\gamma_s L^4\eta_s^3}{\alpha_s}\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{sm+i} \\
& + \eta_s^2 dL^2\left(56+2\chi+2F(X^*)\right) \\
\leq{} & \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1+\frac{\alpha_s}{4\gamma_s}\eta_s\right) H_{sm+r} \\
& + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\sum_{i=0}^{r-1}\frac{\alpha_s}{4m\gamma_s}\eta_s\mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{sm+i} \\
& + \eta_s^2 dL^2 E.
\end{aligned}$$

In the last inequality, we used $\eta_s \leq \frac{\alpha_s}{16\sqrt{2}L^2 m\gamma_s}$.

$Q.E.D$

Theorem E.3. *With the annealing schedule* (16) *and* (17)*, under Assumptions 1, 3 and* $F \geq 0$*,* $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$*,* $B \geq m$*,* $\mu > 3$*,* $\bar{\gamma} = \frac{1}{C_2}$*,* $g \geq \mathrm{e}$ *and*

$$\sigma \geq 3 \vee \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}},$$

for all $k = sm$ *where* $s \in \mathbb{N}$*, the following holds in the update of Algorithm 3:*

$$H_k \leq \mathrm{e}^{-\frac{C_1\bar{\eta}}{2g}\left(\frac{2}{3}\right)^{\frac{2}{\mu}} k^{1-\frac{2}{\mu}} m^{\frac{2}{\mu}} \sigma^{-\frac{2}{\mu}}} H_0 + \frac{8}{3}\bar{\eta} dL^2 E g C_1^{-1} k^{\frac{2}{\mu}} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}},$$

where $E = 56 + 2\chi + 2F(X^*)$*.*

Proof. First of all, we will prove by mathematical induction that in each inner loop the following inequality holds for all $r = 0, \ldots, m-1$:

$$H_{sm+r+1} \leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s(r+1)} H_{sm} + \eta_s^2 dL^2 E \left(\sum_{i=0}^{r} \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i}\right) \qquad \ldots (**)$$

When $r = 0$, from Theorem E.2, we have

$$\begin{aligned} H_{sm+1} &\leq \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{4\gamma_s}\eta_s\right) H_{sm} + \eta_s^2 dL^2 E \\ &\leq \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{2\gamma_s}\eta_s\right) H_{sm} + \eta_s^2 dL^2 E \\ &\leq \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s} \mathrm{e}^{\frac{\alpha_s}{2\gamma_s}\eta_s} H_{sm} + \eta_s^2 dL^2 E \\ &\leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s} H_{sm} + \eta_s^2 dL^2 E. \end{aligned}$$

Thus, $(**)$ holds for $r = 0$.

Now, let us suppose that $(**)$ is true for all $r \leq l$. Then, when $r = l+1$ from Theorem E.2, we have

$$\begin{aligned} H_{sm+l+2} \leq\ & \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{4\gamma_s}\eta_s\right) H_{sm+l+1} + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s} \sum_{i=0}^{l} \frac{\alpha_s}{4m\gamma_s}\eta_s \mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s} H_{sm+i} \\ & + \eta_s^2 dL^2 E \\ \leq\ & \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{4\gamma_s}\eta_s\right)\left(\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s(l+1)} H_{sm} + \eta_s^2 dL^2 E \left(\sum_{j=0}^{l} \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right)\right) \\ & + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s} \sum_{i=0}^{l} \frac{\alpha_s}{4m\gamma_s}\eta_s \mathrm{e}^{-\frac{\alpha_s m}{\gamma_s}\eta_s}\left(\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i} H_{sm} + \eta_s^2 dL^2 E \left(\sum_{j=0}^{i-1} \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right)\right) \\ & + \eta_s^2 dL^2 E \\ \leq\ & \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{4\gamma_s}\eta_s\right)\left(\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s(l+1)} H_{sm} + \eta_s^2 dL^2 E \left(\sum_{j=0}^{l} \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right)\right) \\ & + \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s} \sum_{i=0}^{l} \frac{\alpha_s}{4m\gamma_s}\eta_s \left(\mathrm{e}^{-\frac{\alpha_s\eta_s}{2\gamma_s}(l+1)} H_{sm} + \eta_s^2 dL^2 E \left(\sum_{j=0}^{l} \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right)\right) \\ & + \eta_s^2 dL^2 E \\ \leq\ & \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s(l+1)} \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{4\gamma_s}\eta_s + \sum_{i=0}^{l} \frac{\alpha_s}{4m\gamma_s}\eta_s\right) H_{sm} \\ & + \eta_s^2 dL^2 E \mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\left(1 + \frac{\alpha_s}{4\gamma_s}\eta_s + \sum_{i=0}^{l} \frac{\alpha_s}{4m\gamma_s}\eta_s\right)\left(\sum_{j=0}^{l} \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right) \\ & + \eta_s^2 dL^2 E. \end{aligned}$$

This further implies,

$$\begin{aligned}H_{sm+l+2} &\leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s(l+1)}\mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\mathrm{e}^{\frac{\alpha_s}{2\gamma_s}\eta_s}H_{sm} \\ &\qquad + \eta_s^2 dL^2 E\mathrm{e}^{-\frac{\alpha_s}{\gamma_s}\eta_s}\mathrm{e}^{\frac{\alpha_s}{2\gamma_s}\eta_s}\left(\sum_{j=0}^{l}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right) + \eta_s^2 dL^2 E \\ &\leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s(l+2)}H_{sm} + \eta_s^2 dL^2 E\left(\sum_{i=0}^{l+1}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i}\right).\end{aligned}$$

Therefore, $(**)$ holds for all inner loop and $r = 0, \ldots, m-1$.

Especially, when $r = m-1$, we obtain

$$H_{(s+1)m} \leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s m}H_{sm} + \eta_s^2 dL^2 E\left(\sum_{i=0}^{m-1}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i}\right).$$

Consecutively using this inequality, we obtain

$$\begin{aligned}H_{(s+1)m} &\leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s m}H_{sm} + \eta_s^2 dL^2 E\left(\sum_{i=0}^{m-1}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i}\right) \\ &\leq \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s m}\left(\mathrm{e}^{-\frac{\alpha_{s-1}}{2\gamma_{s-1}}\eta_{s-1}m}H_{(s-1)m} + \eta_{s-1}^2 dL^2 E\left(\sum_{i=0}^{m-1}\mathrm{e}^{-\frac{\alpha_{s-1}}{2\gamma_{s-1}}\eta_{s-1}i}\right)\right) \\ &\quad + \eta_s^2 dL^2 E\left(\sum_{i=0}^{m-1}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i}\right) \\ &= \mathrm{e}^{-\frac{m}{2}\left(\frac{\alpha_s}{\gamma_s}\eta_s + \frac{\alpha_{s-1}}{\gamma_{s-1}}\eta_{s-1}\right)}H_{(s-1)m} \\ &\quad + dL^2 E\left(\eta_s^2\sum_{i=0}^{m-1}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i} + \eta_{s-1}^2\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s m}\sum_{i=0}^{m-1}\mathrm{e}^{-\frac{\alpha_{s-1}}{2\gamma_{s-1}}\eta_{s-1}i}\right) \\ &\cdots \\ &\leq \mathrm{e}^{-\frac{m}{2}\sum_{i=0}^{s}\frac{\alpha_i}{\gamma_i}\eta_i}H_0 \\ &\quad + dL^2 E\sum_{i=0}^{s}\left(\eta_i^2\mathrm{e}^{-\frac{m}{2}\sum_{j=i+1}^{s}\frac{\alpha_j}{\gamma_j}\eta_j}\sum_{j=0}^{m-1}\mathrm{e}^{-\frac{\alpha_i}{2\gamma_i}\eta_i j}\right),\end{aligned}$$

which implies

$$\begin{aligned}H_{(s+1)m} &\leq \mathrm{e}^{-\frac{m}{2}\sum_{i=0}^{s}\frac{\alpha_i}{\gamma_i}\eta_i}H_0 + dL^2 E\sum_{i=0}^{s}\left(\eta_i^2\mathrm{e}^{-\frac{m}{2}\sum_{j=i+1}^{s}\frac{\alpha_s}{\gamma_s}\eta_s}\sum_{j=0}^{m-1}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s j}\right) \\ &\leq \mathrm{e}^{-\frac{m}{2}\sum_{i=0}^{s}\frac{\alpha_i}{\gamma_i}\eta_i}H_0 + \bar{\eta}^2 dL^2 E\sum_{i=0}^{\infty}\mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s i} \\ &= \mathrm{e}^{-\frac{m}{2}\sum_{i=0}^{s}\frac{\alpha_i}{\gamma_i}\eta_i}H_0 + \bar{\eta}^2 dL^2 E\left(1 - \mathrm{e}^{-\frac{\alpha_s}{2\gamma_s}\eta_s}\right)^{-1} \\ &\leq \mathrm{e}^{-\frac{m}{2}\sum_{i=0}^{s}\frac{\alpha_i}{\gamma_i}\eta_i}H_0 + \bar{\eta}^2 dL^2 E\left(\frac{3}{4}\frac{\alpha_s}{2\gamma_s}\eta_s\right)^{-1} \\ &= \mathrm{e}^{-\frac{mC_1\bar{\eta}}{2g}\sum_{i=0}^{s}(i+\sigma)^{-\frac{2}{\mu}}}H_0 + \frac{8}{3}\bar{\eta}dL^2 EgC_1^{-1}(s+\sigma)^{\frac{2}{\mu}} \\ &\leq \mathrm{e}^{-\frac{C1\bar{\eta}}{2g}\left(\frac{2}{3}\right)^{\frac{2}{\mu}}m(s+1)^{1-\frac{2}{\mu}}\sigma^{-\frac{2}{\mu}}}H_0 + \frac{8}{3}\bar{\eta}dL^2 EgC_1^{-1}(s+1)^{\frac{2}{\mu}}\sigma^{\frac{2}{\mu}}.\end{aligned}$$

For the first inequality, we used $\frac{\alpha_i}{\gamma_i}\eta_i \geq \frac{\alpha_s}{\gamma_s}\eta_s$ for all $i \leq s$, for the third inequality, we used $1 - \mathrm{e}^{-c} \geq \frac{3}{4}c$ holds for all $0 < c = \frac{\alpha_s}{2\gamma_s}\eta_s \leq \frac{1}{4}$, and for the last inequality, we used Lemma E.1.

Setting $k = (s+1)m$, we obtain

$$\begin{aligned} H_k \leq \mathrm{e}^{-\frac{C_1\bar{\eta}}{2g}\left(\frac{2}{3}\right)^{\frac{2}{\mu}} k^{1-\frac{2}{\mu}} m^{\frac{2}{\mu}} \sigma^{-\frac{2}{\mu}}} H_0 \\ + \frac{8}{3}\bar{\eta} d L^2 E g C_1^{-1} k^{\frac{2}{\mu}} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}}. \end{aligned}$$

Q.E.D

Finally, we obtain the following global convergence guarantee for Algorithm 2.

Theorem E.4. *Using Algorithm 2 with the annealing schedule $\eta_s = \bar{\eta}(s+\sigma)^{-\frac{1}{\mu}}$ and $\gamma_s = \bar{\gamma}\log\left\{g(s+\sigma)^{\frac{1}{\mu}}\right\}$, under Assumptions 1, 3 and $F \geq 0$, $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, $B \geq m$, $\epsilon = O\left(\frac{LH_0}{C_1C_2^{-1}}\right)$, $\mu \geq 13$, $g = \mathrm{e}^{\frac{h(\epsilon)\vee 2M \vee \bar{\gamma}}{\bar{\gamma}}}$, $\bar{\gamma} = \frac{1}{C_2}$ and*

$$\sigma = 3 \vee \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}},$$

where

$$h(\epsilon) := \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8bd}{\epsilon^2} \vee 1,$$

if we take $B = m = \sqrt{n}$, the largest permissible step size according to the value of σ, the gradient complexity to reach a precision of

$$\mathbb{E}_{X_k \sim \rho_k}[F(X_k)] - F(X^*) \leq \epsilon$$

is

$$\tilde{O}\left(GC_1 + GC_2 + GC_3\right).$$

where

$$\begin{aligned} GC_1 &= ng^{\frac{2\mu}{\mu-2}} C_1^{\frac{-2\mu}{\mu-2}} L^{\frac{2\mu}{\mu-2}} + n^{\frac{1}{2}-\frac{5}{2(\mu-5)}} \epsilon^{-\frac{\mu}{\mu-5}} g^{\frac{3\mu}{\mu-5}} C_1^{-\frac{3\mu}{\mu-5}} C_2^{\frac{\mu}{\mu-5}} (dE)^{\frac{\mu}{\mu-5}} L^{\frac{3\mu}{\mu-5}}, \\ GC_2 &= n^{\frac{1}{2}+\frac{\mu^2-3\mu+6}{2(\mu-2)(\mu-3)}} \left(\frac{gL}{C_1}\right)^{\frac{2\mu^2}{(\mu-2)(\mu-3)}} \\ &\quad + n^{\frac{1}{2}-\frac{5(\mu-3)}{2(\mu^2-11\mu+15)}} \epsilon^{-\frac{\mu(\mu-1)}{\mu^2-11\mu+15}} \left(\frac{gL}{C_1}\right)^{\frac{(3\mu^2-7\mu+6)\mu}{(\mu-3)(\mu^2-11\mu+15)}} (dE)^{\frac{\mu(\mu-1)}{\mu^2-11\mu+15}}, \\ GC_3 &= n^{\frac{1}{2}+\frac{\mu^2+4}{2(\mu-2)^2}} \left(\frac{gL}{C_1}\right)^{\frac{2\mu^2}{(\mu-2)^2}} C_2^{-\frac{2\mu}{(\mu-2)^2}} \\ &\quad + n^{\frac{1}{2}-\frac{5(\mu-2)}{2(\mu^2-13\mu+10)}} \epsilon^{-\frac{\mu(\mu+2)}{\mu^2-13\mu+10}} \left(\frac{gL}{C_1}\right)^{\frac{(3\mu-4)\mu}{\mu^2-13\mu+10}} (dE)^{\frac{\mu(\mu+2)}{\mu^2-13\mu+10}} C_2^{\frac{(\mu^2-12\mu-6)\mu}{(\mu^2-13\mu+10)(\mu-2)}}, \\ E &= 56 + 2\max_{\gamma\geq 1}\left(\frac{d}{\gamma}\log\left(\frac{\mathrm{e}L}{M}\left(\frac{b\gamma}{d}+1\right)\right)\right) + 2F^*, \end{aligned}$$

and C_1 and C_2 are defined in Property C.3.

Proof. Let us take $k = (s+1)m$, where $s \in \mathbb{N} \cup \{0\}$. From Corollary C.1.1, the sufficient condition for

$$\mathbb{E}_{X_k \sim \psi_k}[F(X_k)] - F(X^*) \leq \epsilon$$

is $LW_2^2(\psi_k, \nu_{\gamma_s}) \leq \epsilon/2$ and $\mathbb{E}_{X\sim\nu_{\gamma_s}}[F(X)] - F(X^*) \leq \epsilon/4$. From $g \geq \mathrm{e}^{\frac{2M}{\bar{\gamma}}}$, which implies $\gamma_s \geq \frac{2}{M}$, and from Corollary C.2.1, the latter condition is satisfied when $\gamma_s \geq \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8bd}{\epsilon^2} \vee 1$. Let us define

$$h(\epsilon) := \frac{4d}{\epsilon}\log\left(\frac{\mathrm{e}L}{M}\right) \vee \frac{8bd}{\epsilon^2} \vee 1.$$

Then, as $\gamma_s = \bar{\gamma}\log\left\{g(s+\sigma)^{\frac{1}{\mu}}\right\}$ and $s + \sigma \geq \mathrm{e}$, a sufficient condition is

$$\bar{\gamma}\log g \geq h(\epsilon).$$

This is satisfied with

$$g = \mathrm{e}^{\frac{h(\epsilon)\vee 2M\vee\bar{\gamma}}{\bar{\gamma}}} \geq \mathrm{e}^{\frac{h(\epsilon)}{\bar{\gamma}}}.$$

Moreover, concerning the former condition, from Talagrand's inequality

$$W_2^2(\rho_k, \nu_{\gamma_s}) \leq \frac{2}{\alpha_s} H_\nu(\rho_k),$$

it suffices to have

$$H_{\nu_{\gamma_s}}(\rho_k) \leq \frac{\alpha_s \epsilon}{4L} = \frac{\epsilon}{4L} \frac{C_1 C_2^{-1} \log g(s+\sigma)^{\frac{1}{\mu}}}{g(s+\sigma)^{\frac{1}{\mu}}}.$$

As $g \geq \mathrm{e}$, $s+\sigma \geq 1$ and $(s+\sigma) \leq (s+1)\sigma$, we obtain a simpler sufficient condition which is

$$H_{\nu_{\gamma_s}}(\rho_k) \leq \frac{\epsilon C_1 C_2^{-1}}{4Lk^{\frac{1}{\mu}} m^{\frac{-1}{\mu}} \sigma^{\frac{1}{\mu}} g}.$$

Therefore, from Theorem E.3, it is enough to take $\bar{\eta}$ and k such that

$$\frac{8}{3}\bar{\eta} dL^2 E g C_1^{-1} k^{\frac{2}{\mu}} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}} \leq \frac{\epsilon C_1 C_2^{-1}}{8Lk^{\frac{1}{\mu}} m^{\frac{-1}{\mu}} \sigma^{\frac{1}{\mu}} g}, \tag{20}$$

and

$$k^{1-\frac{2}{\mu}} \geq 2gC_1^{-1} \left(\frac{3}{2}\right)^{\frac{2}{\mu}} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}} \bar{\eta}^{-1} \log\left(\frac{8Lk^{\frac{1}{\mu}} m^{\frac{-1}{\mu}} \sigma^{\frac{1}{\mu}} H_0}{\epsilon C_1 C_2^{-1}}\right). \tag{21}$$

Concerning the first inequality (20), we obtain

$$\bar{\eta}\sigma^{\frac{3}{\mu}} \leq \frac{3C_1^2 C_2^{-1}}{64dL^3 E g^2} \epsilon k^{-\frac{3}{\mu}} m^{\frac{3}{\mu}}. \tag{22}$$

On the other hand, since $(s+1)^{\frac{1}{\mu}} \geq 1$ and $\sigma^{\frac{1}{\mu}} \geq 1$, as long as $\epsilon = O\left(\frac{LH_0}{C_1 C_2^{-1}}\right)$, we can consider the following condition for the second inequality (21):

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(2gC_1^{-1}\left(\frac{3}{2}\right)^{\frac{2}{\mu}} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}} \bar{\eta}^{-1}\right) = \tilde{\Theta}\left(gC_1^{-1} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}} \bar{\eta}^{-1}\right). \tag{23}$$

(I) When

$$\sigma = 3 \vee \left(\frac{8Lg^2}{C_1^2 \bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}} = 3,$$

equation (22) becomes

$$\bar{\eta} \leq \left(\frac{3C_1^2 C_2^{-1} 3^{\frac{-3}{\mu}}}{64dL^3 E g^2}\right) \epsilon k^{-\frac{3}{\mu}} m^{\frac{3}{\mu}}.$$

On the other hand, by plugging $\sigma = 3$ to (23), we obtain

$$\begin{aligned} k^{1-\frac{2}{\mu}} &\geq \tilde{\Theta}\left(gC_1^{-1} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}} \bar{\eta}^{-1}\right) \\ &\geq \tilde{\Theta}\left(gC_1^{-1} m^{-\frac{2}{\mu}} \bar{\eta}^{-1}\right). \end{aligned}$$

If inequality (22) is stronger than $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}L^2 m}$, then

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(\frac{g^3 dEL^3}{C_1^3 C_2^{-1}} k^{\frac{3}{\mu}} m^{-\frac{5}{\mu}} \epsilon^{-1}\right).$$

This leads to

$$k \geq \tilde{\Theta}\left(\left(\frac{g^3 dEL^3}{C_1^3 C_2^{-1}}\right)^{\frac{\mu}{\mu-5}} m^{-\frac{5}{\mu-5}} \epsilon^{-\frac{\mu}{\mu-5}}\right).$$

From this, if we take the largest permissible step size and the smallest permissible σ, the gradient complexity can be calculated with an optimal order when $B = m = n^{\frac{1}{2}}$ as

$$\begin{aligned}\tilde{\Theta}\left(kB + \frac{k}{m}n\right) &= \tilde{\Theta}(k\sqrt{n}) \\ &=\tilde{\Theta}\left(n^{\frac{1}{2}-\frac{5}{2(\mu-5)}}\epsilon^{-\frac{\mu}{\mu-5}}g^{\frac{3\mu}{\mu-5}}C_1^{-\frac{3\mu}{\mu-5}}C_2^{\frac{\mu}{\mu-5}}(dE)^{\frac{\mu}{\mu-5}}L^{\frac{3\mu}{\mu-5}}\right).\end{aligned}$$

If inequality (22) is weaker than $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, then

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(g^2C_1^{-2}L^2m^{1-\frac{2}{\mu}}\right).$$

This leads to

$$k \geq \tilde{\Theta}\left(\left(g^2C_1^{-2}L^2\right)^{\frac{\mu}{\mu-2}}m\right).$$

From this, if we take the largest permissible step size and the smallest permissible σ, the gradient complexity can be calculated with an optimal order when $B = m = n^{\frac{1}{2}}$ as

$$\begin{aligned}\tilde{\Theta}\left(kB + \frac{k}{m}n\right) &= \tilde{\Theta}(k\sqrt{n}) \\ &= \tilde{\Theta}\left(ng^{\frac{2\mu}{\mu-2}}C_1^{\frac{-2\mu}{\mu-2}}L^{\frac{2\mu}{\mu-2}}\right).\end{aligned}$$

Therefore, we obtain the following gradient complexity for this case:

$$\tilde{\Theta}\left(ng^{\frac{2\mu}{\mu-2}}C_1^{\frac{-2\mu}{\mu-2}}L^{\frac{2\mu}{\mu-2}} + n^{\frac{1}{2}-\frac{5}{2(\mu-5)}}\epsilon^{-\frac{\mu}{\mu-5}}g^{\frac{3\mu}{\mu-5}}C_1^{-\frac{3\mu}{\mu-5}}C_2^{\frac{\mu}{\mu-5}}(dE)^{\frac{\mu}{\mu-5}}L^{\frac{3\mu}{\mu-5}}\right). \tag{24}$$

(II) When

$$\sigma = 3 \vee \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2L^2\bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}} = \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}},$$

equation (22) becomes

$$\bar{\eta} \leq \left(\frac{3C_1^2C_2^{-1}}{64dL^3Eg^2}\right)^{\frac{\mu-3}{\mu-6}}\left(\frac{8Lg^2}{C_1^2}\right)^{\frac{-3}{\mu-6}}\epsilon^{\frac{\mu-3}{\mu-6}}k^{-\frac{3(\mu-3)}{\mu(\mu-6)}}m^{\frac{3(\mu-3)}{\mu(\mu-6)}}.$$

On the other hand, by plugging $\sigma = \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}}$ to (23), we obtain

$$\begin{aligned}k^{1-\frac{2}{\mu}} &\geq \tilde{\Theta}\left(gC_1^{-1}m^{-\frac{2}{\mu}}\sigma^{\frac{2}{\mu}}\bar{\eta}^{-1}\right) \\ &\geq \tilde{\Theta}\left(g^{\frac{\mu+1}{\mu-3}}C_1^{-\frac{\mu+1}{\mu-3}}L^{\frac{2}{\mu-3}}m^{-\frac{2}{\mu}}\bar{\eta}^{-\frac{\mu-1}{\mu-3}}\right).\end{aligned}$$

If inequality (22) is stronger than $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, then

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(\left(\frac{C_1^2C_2^{-1}}{dEL^3g^2}\right)^{-\frac{\mu-1}{\mu-6}}\left(\frac{Lg^2}{C_1^2}\right)^{\frac{3(\mu-1)}{(\mu-6)(\mu-3)}}\frac{g^{\frac{\mu+1}{\mu-3}}L^{\frac{2}{\mu-3}}}{C_1^{\frac{\mu+1}{\mu-3}}}\epsilon^{-\frac{\mu-1}{\mu-6}}k^{\frac{3(\mu-1)}{\mu(\mu-6)}}m^{-\frac{5(\mu-3)}{\mu(\mu-6)}}\right).$$

This leads to

$$k \geq \tilde{\Theta}\left(\frac{\left(\left(\frac{C_1^2C_2^{-1}}{dEL^3g^2}\right)^{-\frac{\mu-1}{\mu-6}}\left(\frac{Lg^2}{C_1^2}\right)^{\frac{3(\mu-1)}{(\mu-6)(\mu-3)}}g^{\frac{\mu+1}{\mu-3}}C_1^{-\frac{\mu+1}{\mu-3}}L^{\frac{2}{\mu-3}}\right)^{\frac{\mu(\mu-6)}{\mu^2-11\mu+15}}}{m^{\frac{5(\mu-3)}{\mu^2-11\mu+15}}\epsilon^{\frac{\mu(\mu-1)}{\mu^2-11\mu+15}}}\right).$$

From this, if we take the largest permissible step size and the smallest permissible σ, the gradient complexity can be calculated with an optimal order when $B = m = n^{\frac{1}{2}}$ as

$$\begin{aligned}\tilde{\Theta}\left(kB + \frac{k}{m}n\right) &= \tilde{\Theta}(k\sqrt{n}) \\ &= \tilde{\Theta}\left(n^{\frac{1}{2} - \frac{5(\mu-3)}{2(\mu^2-11\mu+15)}} \epsilon^{-\frac{\mu(\mu-1)}{\mu^2-11\mu+15}} \left(\frac{gL}{C_1}\right)^{\frac{(3\mu^2-7\mu+6)\mu}{(\mu-3)(\mu^2-11\mu+15)}} (dE)^{\frac{\mu(\mu-1)}{\mu^2-11\mu+15}}\right).\end{aligned}$$

If inequality (22) is weaker than $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, then

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(\left(\frac{gL}{C_1}\right)^{\frac{2\mu}{\mu-3}} m^{\frac{\mu^2-3\mu+6}{\mu(\mu-3)}}\right).$$

This leads to

$$k \geq \tilde{\Theta}\left(\left(\frac{gL}{C_1}\right)^{\frac{2\mu^2}{(\mu-2)(\mu-3)}} m^{\frac{\mu^2-3\mu+6}{(\mu-2)(\mu-3)}}\right).$$

From this, if we take the largest permissible step size and the smallest permissible σ, the gradient complexity can be calculated with $B = m = n^{\frac{1}{2}}$ as

$$\begin{aligned}\tilde{\Theta}\left(kB + \frac{k}{m}n\right) &= \tilde{\Theta}(k\sqrt{n}) \\ &= \tilde{\Theta}\left(n^{\frac{1}{2} + \frac{\mu^2-3\mu+6}{2(\mu-2)(\mu-3)}} \left(\frac{gL}{C_1}\right)^{\frac{2\mu^2}{(\mu-2)(\mu-3)}}\right).\end{aligned}$$

Therefore, we obtain the following gradient complexity for this case:

$$\tilde{\Theta}\left(n^{\frac{2\mu^2-8\mu+12}{2(\mu-2)(\mu-3)}} \left(\frac{gL}{C_1}\right)^{\frac{2\mu^2}{(\mu-2)(\mu-3)}} + \frac{n^{\frac{1}{2} - \frac{5(\mu-3)}{2(\mu^2-11\mu+15)}}}{\epsilon^{\frac{\mu(\mu-1)}{\mu^2-11\mu+15}}} \left(\frac{gL}{C_1}\right)^{\frac{(3\mu^2-7\mu+6)\mu}{(\mu-3)(\mu^2-11\mu+15)}} (dE)^{\frac{\mu(\mu-1)}{\mu^2-11\mu+15}}\right). \tag{25}$$

(III) When

$$\sigma = 3 \vee \left(\frac{8Lg^2}{C_1^2\bar{\eta}}\right)^{\frac{\mu}{\mu-3}} \vee \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}} = \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}},$$

equation (22) becomes

$$\bar{\eta} \leq \left(\frac{3C_1^2C_2^{-1}}{64dL^3Eg^2}\right)^{\frac{\mu-2}{\mu-8}} \left(\frac{2}{\mu C_2 L^2}\right)^{\frac{-3}{\mu-8}} \epsilon^{\frac{\mu-2}{\mu-8}} k^{-\frac{3(\mu-2)}{\mu(\mu-8)}} m^{\frac{3(\mu-2)}{\mu(\mu-8)}}.$$

On the other hand, by plugging $\sigma = \left(\frac{2}{\mu C_2 L^2 \bar{\eta}^2}\right)^{\frac{\mu}{\mu-2}}$ to (23), we obtain

$$\begin{aligned}k^{1-\frac{2}{\mu}} &\geq \tilde{\Theta}\left(gC_1^{-1} m^{-\frac{2}{\mu}} \sigma^{\frac{2}{\mu}} \bar{\eta}^{-1}\right) \\ &\geq \tilde{\Theta}\left(gC_1^{-1} (C_2L^2)^{-\frac{2}{\mu-2}} m^{-\frac{2}{\mu}} \bar{\eta}^{-\frac{\mu+2}{\mu-2}}\right).\end{aligned}$$

If inequality (22) is stronger than $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, then

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(gC_1^{-1}(C_2L^2)^{\frac{-2}{\mu-2}} \left(\frac{C_1^2C_2^{-1}}{dEL^3g^2}\right)^{-\frac{\mu+2}{\mu-8}} (C_2L^2)^{\frac{-3(\mu+2)}{(\mu-8)(\mu-2)}} k^{\frac{3(\mu+2)}{\mu(\mu-8)}} m^{-\frac{5(\mu-2)}{\mu(\mu-8)}} \epsilon^{\frac{\mu+2}{\mu-8}}\right).$$

This leads to

$$k \geq \tilde{\Theta}\left(\frac{\left(gC_1^{-1}(C_2L^2)^{\frac{-2}{\mu-2}} \left(\frac{C_1^2C_2^{-1}}{dEL^3g^2}\right)^{-\frac{\mu+2}{\mu-8}} (C_2L^2)^{\frac{-3(\mu+2)}{(\mu-8)(\mu-2)}}\right)^{\frac{\mu(\mu-8)}{\mu^2-13\mu+10}}}{m^{\frac{5(\mu-2)}{\mu^2-13\mu+10}} \epsilon^{\frac{\mu(\mu+2)}{\mu^2-13\mu+10}}}\right).$$

From this, if we take the largest permissible step size and the smallest permissible σ, the gradient complexity can be calculated with an optimal order when $B = m = n^{\frac{1}{2}}$ as

$$\begin{aligned}\tilde{\Theta}\left(kB + \frac{k}{m}n\right) &= \tilde{\Theta}(k\sqrt{n}) \\ &= \tilde{\Theta}\left(\frac{n^{\frac{1}{2}-\frac{5(\mu-2)}{2(\mu^2-13\mu+10)}}}{\epsilon^{\frac{\mu(\mu+2)}{\mu^2-13\mu+10}}}\left(\frac{gL}{C_1}\right)^{\frac{(3\mu-4)\mu}{\mu^2-13\mu+10}}(dE)^{\frac{\mu(\mu+2)}{\mu^2-13\mu+10}}C_2^{\frac{(\mu^2-12\mu-6)\mu}{(\mu^2-13\mu+10)(\mu-2)}}\right).\end{aligned}$$

If inequality (22) is weaker than $0 < \bar{\eta} < \frac{C_1}{16\sqrt{6}gL^2m}$, then

$$k^{1-\frac{2}{\mu}} \geq \tilde{\Theta}\left(gC_1^{-1}(C_2L^2)^{-\frac{2}{\mu-2}}\left(\frac{gL^2}{C_1}\right)^{\frac{\mu+2}{\mu-2}} m^{\frac{\mu^2+4}{\mu(\mu-2)}}\right).$$

This leads to

$$k \geq \tilde{\Theta}\left(\left(gC_1^{-1}(C_2L^2)^{-\frac{2}{\mu-2}}\left(\frac{gL^2}{C_1}\right)^{\frac{\mu+2}{\mu-2}}\right)^{\frac{\mu}{\mu-2}} m^{\frac{\mu^2+4}{(\mu-2)^2}}\right).$$

From this, if we take the largest permissible step size and the smallest permissible σ, the gradient complexity can be calculated with $B = m = n^{\frac{1}{2}}$ as

$$\begin{aligned}\tilde{\Theta}\left(kB + \frac{k}{m}n\right) &= \tilde{\Theta}(k\sqrt{n}) \\ &= \tilde{\Theta}\left(n^{\frac{1}{2}+\frac{\mu^2+4}{2(\mu-2)^2}}\left(\frac{gL}{C_1}\right)^{\frac{2\mu^2}{(\mu-2)^2}} C_2^{-\frac{2\mu}{(\mu-2)^2}}\right).\end{aligned}$$

Therefore, we obtain the following gradient complexity for this case:

$$\tilde{\Theta}\left(n^{\frac{\mu^2-2\mu+4}{(\mu-2)^2}}\left(\frac{gL}{C_1C_2^{\frac{1}{\mu}}}\right)^{\frac{2\mu^2}{(\mu-2)^2}} + \frac{n^{\frac{\mu^2-18\mu+20}{2(\mu^2-13\mu+10)}}}{\epsilon^{\frac{\mu(\mu+2)}{\mu^2-13\mu+10}}}\left(\frac{gL}{C_1}\right)^{\frac{(3\mu-4)\mu}{\mu^2-13\mu+10}}(dE)^{\frac{\mu(\mu+2)}{\mu^2-13\mu+10}}C_2^{\frac{(\mu^2-12\mu-6)\mu}{(\mu^2-13\mu+10)(\mu-2)}}\right). \quad (26)$$

The statement of the theorem is obtained by grouping (24), (25) and (26).

$Q.E.D$

Now, looking at the gradient complexity of Theorem E.4, we remark that the dependence on n is slightly improved for finite values of μ compared with $\mu \to \infty$. However, taking into account that $g = \mathrm{e}^{C_2h(\epsilon)}$ most of the time, the dependence on ϵ becomes worse as the exponent of g of the first term is greater than 2, and that of the second term is greater than 3 for all GC_i $(i = 1, 2, 3)$. Since this influence cannot be ignored in this case, we conclude that the best value of μ is $\mu = \infty$ in our analysis, i.e., the method that keeps η and γ constants.