OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Langevin Dynamics: A Unified Perspective on Optimization via Lyapunov Potentials

August Y. Chen AYC74@CORNELL.EDU
Ayush Sekhari SEKHARI@MIT.EDU
Karthik Sridharan KS999@CORNELL.EDU

Abstract

We study the problem of non-convex optimization using Stochastic Gradient Langevin Dynamics (SGLD). SGLD is a natural and popular variation of stochastic gradient descent where at each step, appropriately scaled Gaussian noise is added. To our knowledge, the only strategy for showing global convergence of SGLD on the loss function is to show that SGLD can sample from a stationary distribution which assigns larger mass when the function is small (the Gibbs measure), and then to convert these guarantees to optimization results.

We employ a new strategy to analyze the convergence of SGLD to global minima, based on Lyapunov potentials and optimization. This adapts well to the case with a stochastic gradient oracle, which is natural for machine learning applications where one wants to minimize population loss but only has access to stochastic gradients via minibatch training samples. Here we provide 1) improved rates in the setting of previous works studying SGLD for optimization under mild regularity assumptions, and 2) the first finite gradient complexity guarantee for SGLD where the function is Lipschitz and the Gibbs measure defined by the function satisfies a Poincaré Inequality.

1. Introduction

We consider the minimization problem

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}).$$

More specifically we are interested in returning a vector $\mathbf{w}$ such that $F(\mathbf{w}) - \min_{\mathbf{w}} F(\mathbf{w}) \leq \varepsilon$ for some desired sub-optimality $\varepsilon > 0$. In Machine Learning (ML) settings, F can be thought of as population loss and $\mathbf{w}$ as the parameters of a model we are using for the learning problem. Additionally, in ML one does not have direct access to F but only via samples $\mathbf{z}_1, \ldots, \mathbf{z}_n$ drawn iid from some unknown but fixed distribution D and we assume that $\mathbb{E}_{\mathbf{z}\sim D}[f(\mathbf{w};\mathbf{z})] = F(\mathbf{w})$. Here the $\mathbf{z}_i$ can be thought of as input-output pairs and $f(\mathbf{w};\mathbf{z})$ can be thought of as the loss of the model parametrized by weights $\mathbf{w}$ on instance $\mathbf{z}$. When the objective function/loss function is differentiable (or sub-differentiable), then a common method of choice in practice is to use gradient descent (GD), stochastic gradient descent (SGD) and its variants to perform the optimization. To understand their properties theoretically, we aim to understand how many gradient computations are necessary to find an ε-suboptimal $\mathbf{w}$, and for which functions F this is possible. Under geometric conditions such as convexity, the properties of GD and SGD are well-understood. For convex functions, methods from acceleration to variance reduction have been developed to speed up runtime in a variety of settings.

Matching lower and upper bounds exist for both exact and stochastic gradients for convex functions and smaller classes such as strongly convex functions [6].

In recent years, machine learning has seen an explosion of success employing non-convex models. However, despite intensive study, the empirical success of optimizing non-convex functions to global optima is not at all well-understood theoretically. Beyond convexity, GD/SGD converges to global minima under general conditions such as Polyak-Łojasiewicz (PŁ) [28] [22] and Kurdyka-Łojasiewicz (KŁ) [20] functions. Much more general geometric properties where GD/SGD can converge to global minima were found in [15], by considering what properties hold if and only if gradient flow succeeds. Additionally, researchers have proved GD/SGD with appropriate initialization can find global minima of particular non-convex problems such as matrix square root [17] [15], matrix completion [18], phase retrieval [7] [11] [31] [15], and dictionary learning [2].

While gradient descent/stochastic gradient descent has been shown to be successful in the aforementioned cases, there are well-known cases where GD/SGD does not work. A natural variant of gradient descent that is used for optimization is *perturbed* gradient descent, where Gaussian noise is added to the iterates of stochastic gradient descent – known as Langevin Dynamics – is frequently analyzed. Formally, the iterates of *Gradient Langevin Dynamics* (GLD) are given as follows:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla F(\mathbf{w}_t) + \sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t. \tag{1}$$

Here $\eta > 0$ is the step size, $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \mathbb{I}_d)$ is a d-dimensional standard Gaussian, and $\beta > 0$ is the *inverse temperature parameter* (when larger, noise is weighted less). When we use a stochastic gradient oracle $\nabla f(\mathbf{w}_t; \mathbf{z}_t)$ in place of $\nabla F(\mathbf{w}_t)$, these iterates become those of *Stochastic Gradient Langevin Dynamics* (SGLD). Langevin Dynamics has been shown to work in several highly non-convex settings where even gradient descent fails [29].

The continuous time version of (1) is the following Stochastic Differential Equation (SDE):

$$\mathrm{d}\mathbf{w}(t) = -\nabla F(\mathbf{w}(t))\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathbf{B}(t). \tag{2}$$

Here $\mathbf{B}(t)$ denotes a standard Brownian motion in $\mathbb{R}^d$. This is known as the *Langevin Diffusion*. Broadly, all of these recursions are known as *Langevin Dynamics*. Note as $\beta \to \infty$, these iterates become exactly those of GD/SGD (for (1)) or Gradient Flow (for (2)).

The only strategy in literature we know for proving *global optimization guarantees* for GLD is by first showing sampling guarantees, and then connecting it back to optimization. Consider the Gibbs measure $\mu_\beta = e^{-\beta F}/Z$, where Z denotes the partition function. It is well known that the continuous-time Langevin Diffusion with inverse temperature β (2) converges to μ_β [14] (although this is in fact false in discrete time). When β is sufficiently large, one can use this convergence to get optimization guarantees. This was exactly the strategy of the works [19, 29, 35, 39]. These works prove that under their conditions, this measure μ_β can be sampled from, and therefore non-convex optimization can succeed. Sampling from μ_β is generally known as *Langevin Monte Carlo* (LMC).

For optimization, we need the inverse temperature $\beta = \widetilde{\Omega}(\frac{d}{\varepsilon})$; consider the natural example when $F(\mathbf{w}) = \|\mathbf{w}\|^2$, thus μ_β is a Gaussian with covariance $\frac{1}{\beta}\mathbb{I}_d$. By standard results on Gaussian concentration [34], we see we need $\beta = \widetilde{\Omega}(\frac{d}{\varepsilon})$ for even exact oracle access to μ_β to succeed for efficient optimization. If $\varepsilon = o(\frac{d}{\beta})$, then $\mu_\beta(\{\mathbf{w} : F(\mathbf{w}) < \varepsilon\})$ is exponentially small in d.

The most general condition under which LMC has been proven to be successful is when μ_β satisfies a *Poincaré Inequality* [13]. A Poincaré Inequality is defined as follows:

. Emails: {ayc74@cornell.edu, sekhari@mit.edu, ks999@cornell.edu}

Definition 1 *A measure μ on $\mathbb{R}^d$ satisfies a Poincaré Inequality with Poincaré constant $\mathsf{C}_{\mathrm{PI}}(\mu)$ if for all infinitely differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$\int_{\mathbb{R}^d} f^2 \mathrm{d}\mu - \left(\int_{\mathbb{R}^d} f \mathrm{d}\mu\right)^2 \le \mathsf{C}_{\mathrm{PI}}(\mu) \int_{\mathbb{R}^d} \|\nabla f\|^2 \mathrm{d}\mu.$$

If the above is not satisfied, following the convention, we set $\mathsf{C}_{\mathrm{PI}}(\mu) = \infty$.

There is evidence that in several cases, LMC does not succeed efficiently under looser conditions on μ_β such as a weak Poincaré Inequality [25]. Ultimately, a Poincaré Inequality being satisfied by μ_β is a geometric condition on F. It is quite natural: when F is convex (μ_β is log-concave), μ_β satisfies a Poincaré Inequality [5]. But a Poincaré Inequality is in fact much more general. It is stable under bounded perturbations (at the expense of worsening the Poincar'e constant), hence covering a wide range of cases that log-concave measures (when F is convex) does not (see Proposition 4.2.7, Bakry et al. [3]). Poincaré Inequalities are also stable under convolutions and mixtures, in the sense that for distributions which all satisfy a Poincaré Inequality, their mixture or convolutions between any two of them will also satisfy a Poincaré Inequality (again, at the expense of worsening the Poincaré constant; see Propositions 2.3.7 and 2.3.8, Chewi [12]).

However, the approach of studying optimization guarantees for GLD/SGLD via sampling is not necessarily optimal. It does not handle stochastic gradients well (the more relevant setting for optimization), only works well when F is approximately smooth, and converting sampling results back to optimization guarantees often incurs extra runtime. Moreover, it is not clear whether sampling, i.e. proving mixing, is necessary to study optimization. In this paper, we take a different route. We highlight the benefit our approach brings next in Subsection B.2.

Notation. Unless otherwise specified the domain is $\mathbb{R}^d$, with origin $\vec{\mathbf{0}}$. We denote the Laplacian (sum of second derivatives) of a twice-differentiable function f by Δf. Here $\mathbb{B}(p, R)$ denotes the Euclidean l_2 ball centered at $p \in \mathbb{R}^d$ with radius $R \ge 0$. $\mathcal{S}^{d-1}$ denotes the surface of the d-dimensional unit sphere. $\widetilde{\Omega}$, $\widetilde{\Theta}$, $\widetilde{O}$ hide universal constants, log factors in β, d, ε, as well as $\mathbf{w}_0$-dependence. Sometimes we will write exponentials as exp for readability. When we write vectors $\mathbf{w}_t$ this denotes time t in discrete time, and when we write $\mathbf{w}(t)$ this denotes time t in continuous time. Unless indicated otherwise, $\mathbb{E}$ refers to expectation over the Brownian motion/random variables $\boldsymbol{\varepsilon}_t$ (as well as the data samples $\mathbf{z}_t$ in the SGLD case), and $\mathbb{E}_{\mathbf{w}}$ denotes the same expectation when the stochastic processes is initialized at $\mathbf{w}$. For any set $\mathcal{U} \subset \mathbb{R}^d$, let the hitting time of the Langevin Diffusion (2) initialized at $\mathbf{w}$ to $\mathcal{U}$ be $\tau_{\mathcal{U}}(\mathbf{w})$. We assume that first order tensors, i.e. vectors, are equipped with l_2 Euclidean norm and that all second order tensors (i.e. matrices) and above are equipped with operator norm. When we write $\|\cdot\|$ without specifying the norm, we implicitly mean the l_2 Euclidean norm of a vector. For some f differentiable to k orders, we will let $\nabla^k f$ denote the tensor of all the k-th order derivatives of f, and $\|\cdot\|_{\mathrm{op}}$ denotes the corresponding tensor's operator norm.

2. Lyapunov Potentials and Optimization

In the rest of this paper, suppose F has a global minimum $\mathbf{w}^\star$, which need not be unique (thus $\mathbf{w}^\star$ can refer to any of these). Furthermore, without loss of generality, assume that $F(\mathbf{w}^\star) = 0$.

2.1. Our Strategy

Optimization under Langevin Dynamics can ultimately be posed as a question of *hitting time*: how long does it take to reach a point $\mathbf{w}$ such that $F(\mathbf{w}) \le \varepsilon$? In the probability theory and stochastic partial differential equations (PDEs) literature, an extensive program has been devoted to studying the connection between isoperimetric inequalities such as a Poincaré Inequality, hitting times of the Langevin Diffusion to sets $A \subset \mathbb{R}^d$, and *Lyapunov potentials*. As mentioned in Section 1, Poincaré Inequalities are the loosest conditions under which global optimization guarantees for Langevin Dynamics have been well-studied. This literature connects these inequalities to the *geometry of* F.

Definition 2 *Say a non-negative function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a Lyapunov potential (for Langevin Dynamics at inverse temperature β given in (2)) if $\Phi \ge 1$ and on the set $\{\mathbf{w} : F(\mathbf{w}) > \varepsilon\}$ we have*

$$\langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \ge \lambda \Phi(\mathbf{w}) + \frac{1}{\beta} \Delta \Phi(\mathbf{w}), \tag{3}$$

where β refers to the inverse temperature of (2).

Our main method to study optimization is to track the progress of GLD/SGLD using the Lyapunov potential $\Phi(\mathbf{w})$, which we outline in Section B. The geometric condition (3) turns out to be closely linked to a Poincaré Inequality: as a corollary of Theorem 2.1 of Cattiaux and Guillin [8], we obtain the following:

Theorem 3 *Assume that μ_β satisfies a Poincaré inequality with constant $\mathsf{C}_{\mathrm{PI}}(\mu_\beta)$ and has finite second second moment for some $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. Then on $\mathcal{A}_\varepsilon^c = \{\mathbf{w} : F(\mathbf{w}) > \varepsilon\}$,*

$$\langle \nabla F(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle \ge \lambda \Phi(\mathbf{w}) + \frac{1}{\beta} \Delta \Phi(\mathbf{w}) \quad \text{for} \quad \lambda \in \left[\frac{1}{8\beta} \min\left(\frac{1}{\mathsf{C}_{\mathrm{PI}}(\mu_\beta)}, \frac{1}{2} \right), \frac{1}{4\beta} \min\left(\frac{1}{\mathsf{C}_{\mathrm{PI}}(\mu_\beta)}, \frac{1}{2} \right) \right], \tag{4}$$

for some non-negative Φ that is differentiable to all orders such that on $\mathcal{A}_\varepsilon^c$, Φ takes the explicit form

$$\Phi(\mathbf{w}') = \mathbb{E}_{\mathbf{w}'}[\exp(\lambda \tau_{\mathcal{A}_\varepsilon})].$$

Remark 4 *Note that on $\mathcal{A}_\varepsilon^c$, $\Phi \ge 1$. Also note Φ generally behaves in a 'dimension free' manner, depending on how $\tau_{\mathcal{A}_\varepsilon}(\mathbf{w}')$ behaves, as $\lambda \le \frac{1}{4\beta} \min\left(\frac{1}{\mathsf{C}_{\mathrm{PI}}(\mu_\beta)}, \frac{1}{2} \right)$ is very small.*

2.2. Results

Now, we state our results in full detail. Complete statements and proofs, including all explicit dependencies, are in Section D. For all of our results, recall from the above that the desired tolerance $\varepsilon = \widetilde{\Omega}\left(\frac{d}{\beta}\right)$; no results so far in literature yield meaningful optimization guarantees for smaller tolerance levels.

Before we state our results more explicitly, we state our assumptions, which are in fact necessary. Our first assumption, generalized to higher order derivatives from [15], is that the Lyapunov potential Φ satisfies 'self-bounding regularity' in the following sense:

Definition 5 *A k times differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies k-th order self-bounding regularity if*

$$\left\| \nabla^k f(\mathbf{w}) \right\|_{\mathrm{op}} \le \rho_{f,k}(|f(\mathbf{w})|)$$

for some increasing function $\rho_{f,k} : \mathbb{R} \to \mathbb{R}_{\geq 0}$.

We say f *satisfies polynomial-like self-bounding regularity at order* k *if we can express* $\rho_{f,k}(z) = \sum_{j=0}^{n} c_j z^{d_j}$ *where all* $d_j \geq 0$. *Note without loss of generality we can assume all* $c_j, d_j \geq 0$ *and* $\rho_{f,k}(z) = A(z+1)^p$ *or* $\rho_{f,k}(z) = A + Az^p$ *by the AM-GM Inequality.*

Assumption 1 *Suppose* Φ *satisfies first, second, and third order polynomial-like self-bounding regularity where the monomials in the self-bounding regularity functions have degree at most 1.*

Such an assumption on the relevant Lyapunov potential is necessary to go from continuous to discrete-time optimization: Theorem 3 from De Sa et al. [15] shows even for Gradient Flow/Gradient Descent, there are examples where discrete-time optimization fails when continuous-time optimization succeeds, but the Lyapunov potential did not satisfy self-bounding regularity. Analysis of the same or similar examples hold for the Langevin Diffusion/GLD, where the Langevin Diffusion succeeds as an optimization strategy but discrete-time GLD/SGLD does not. Note Assumption 1 is satisfied by many Lyapunov functions, e.g. when the Lyapunov function Φ has tail growth polynomial in $\|\mathbf{w}\|$ or of the form $e^{r\|\mathbf{w}\|^s}$ for $s \leq 1$, going well beyond smoothness.

Now we state our assumptions on F. We consider the most general setting of previous works [4, 13] for analyzing LMC where we assume F is Hölder continuous with parameter $0 \leq s \leq 1$:

Assumption 2 (Hölder continuity) *Suppose* ∇F *satisfies* L*-Hölder continuity for some* $0 \leq s \leq 1$*:*

$$\|\nabla F(\mathbf{u}) - \nabla F(\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|^s.$$

When $s > 0$, that is F is not Lipschitz, we also require an assumption on the growth of F. This significantly generalizes the dissipation assumption (when $s = 1$ and $\gamma = 2$) made in several previous works studying non-convex optimization [24, 29, 35, 39].

Assumption 3 *There exists* $\gamma \geq 2s$ *such that for some* $m, b > 0$ *and all* $\mathbf{w} \in \mathbb{R}^d$,

$$\langle \mathbf{w}, \nabla F(\mathbf{w}) \rangle \geq m\|\mathbf{w}\|^\gamma - b.$$

Analyzing growth rates, we can see $\gamma \leq s + 1$, which leads to no issues for $0 \leq s \leq 1$. Note this assumption is quite reasonable: in some sense it states that the gradient will push us towards the origin when we are sufficiently far away. Moreover, all critical points of F are in $\mathbb{B}(\vec{\mathbf{0}}, (b/m)^{1/\gamma})$. However, we allow for arbitrary non-convexity inside this ball. In fact, by adding a suitable regularizer penalizing solutions lying outside $\mathbb{B}(\vec{\mathbf{0}}, (b/m)^{1/\gamma})$, we can ensure F satisfies the above, which is discussed on page 15 of Raginsky et al. [29].

Theorem 6 *Suppose that* F *satisfies Assumption 2 and Assumption 3,* μ_β *satisfies a Poincaré Inequality with constant* $\mathsf{C}_{\text{PI}}(\mu_\beta)$ *for* $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$*, and* μ_β *has finite second moment* $S < \infty$. *(In our results dependence on* S *will be logarithmic.) Suppose* Φ *(from Theorem 3) satisfies Assumption 1. Then running* GLD*, with probability at least* $1 - \delta$*, across all the runs we will reach a* $\mathbf{w}$ *with* $F(\mathbf{w}) \leq \varepsilon$ *in at most*

$$\widetilde{O}\left(\max\left\{d^3 \max(\mathsf{C}_{\text{PI}}(\mu_\beta), 1)^3, \frac{d^{2+\frac{s}{2}} \max(\mathsf{C}_{\text{PI}}(\mu_\beta), 1)^{2+\frac{s}{2}}}{\varepsilon^{2+\frac{s}{2}}}\right\} \log\left(\frac{1}{\delta}\right)\right) \tag{5}$$

gradient evaluations.

We note considering Assumption 2 for any $s \geq 0$ and a Poincaré Inequality not only compatible but natural to study in tandem, as discussed in [13].

We now move on to the stochastic gradient oracle case. Some control over the stochastic gradient estimates is necessary: if they are very inaccurate, following them will be meaningless.

Assumption 4 (Bound of variance of gradient estimates) *The unbiased gradient estimate $\nabla f(\mathbf{w}; \mathbf{z})$ of $\nabla F(\mathbf{w})$ satisfies the sub-Gaussian property that for all $\mathbf{w} \in \mathbb{R}^d$ and $t \geq 0$,*

$$\mathbb{P}_{\mathbf{z}}(\|\nabla f(\mathbf{w}; \mathbf{z}) - \nabla F(\mathbf{w})\|_2 \geq t) \leq e^{-t^2/\sigma_F^2}. \tag{6}$$

Assumption 4 covers the classic setting of stochastic optimization where $\nabla f(\mathbf{w}; \mathbf{z}) = \nabla F(\mathbf{w}) + \boldsymbol{\varepsilon}_t$ where $\boldsymbol{\varepsilon}_t$ is sub-Gaussian with mean 0 and variance σ_F^2 [26]. We expect our techniques to hold when gradient noise scales in function value, a more general setting discussed in De Sa et al. [15], but for simplicity we work with Assumption 4.

We also need the following assumption made in Raginsky et al. [29] studying stochastic optimization in this setting. This is also reasonable, saying stochastic gradients contain reasonable signal and will push us towards the origin when sufficiently far away.

Assumption 5 *For every $\mathbf{z}$, $\nabla f(\mathbf{w}; \mathbf{z})$ satisfy Assumption 2 and Assumption 3. (Note they may be satisfied with larger L and b and smaller m.)*

Then, we have the following:

Theorem 7 *Suppose μ_β, F, Φ satisfy the same assumptions as in Theorem 6. Then running* SGLD *with a stochastic gradient oracle satisfying Assumption 4 and Assumption 5, we obtain the same guarantee (5) of the query complexity of our stochastic gradient oracle as in Theorem 6.*

To our knowledge, our result Theorem 7 is the first finite iteration guarantee for the setting of F Hölder-continuous and μ_β satisfying a Poincaré Inequality with a stochastic gradient oracle. The stronger assumption of smoothness is not satisfied by many canonical non-convex optimization problems [15], so analyzing optimization with a stochastic gradient oracle in this more general setting is highly relevant to study.

Recall from our conditions Assumption 2 and Assumption 3 that by analyzing the implied growth rates of F, we have $2s \leq \gamma \leq s + 1$. Thus when $s = 1$, $\gamma = 1$ is forced, so this recovers as a special case of our assumption the smooth and dissipative setting from [19, 29, 35, 39]. In turn, $s = 1, \gamma = 1$ actually implies μ_β satisfies a Poincaré Inequality for all $\beta \geq \frac{2}{m}$ [29]. In this setting we have the following result which is stronger than directly applying Theorem 6:

Theorem 8 *Suppose F is L-smooth and (m, b)-dissipative (that is, there exist $m, b > 0$ such that $\langle \mathbf{w}, \nabla F(\mathbf{w}) \rangle \geq m\|\mathbf{w}\|^2 - b$). Moreover suppose the resulting Φ satisfies Assumption 1. Running either* GLD *or* SGLD *with a stochastic gradient oracle satisfying Assumption 4 and Assumption 5, with probability at least $1 - \delta$, across all the runs we will reach a $\mathbf{w}$ with $F(\mathbf{w}) \leq \varepsilon$ in at most*

$$\widetilde{O}\left(\max\left\{d^3 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^3, \frac{d^2 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^2}{\varepsilon^2}\right\} \log\left(\frac{1}{\delta}\right)\right)$$

gradient/stochastic gradient evaluations.

For all these results, we highlight the improvement on literature in Subsection B.2.

References

[1] Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2169–2176. IEEE, 2023.

[2] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, pages 113–149. PMLR, 2015.

[3] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

[4] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

[5] Sergey G Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999.

[6] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[7] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[8] Patrick Cattiaux and Arnaud Guillin. Hitting times, functional inequalities, lyapunov conditions and uniform ergodicity. *Journal of Functional Analysis*, 272(6):2361–2391, 2017.

[9] Patrick Cattiaux, Arnaud Guillin, and Pierre André Zitt. Poincaré inequalities and hitting times. *Annales de l'IHP Probabilités et Statistiques*, 49(1):95–118, 2013.

[10] Xi Chen, Simon S Du, and Xin T Tong. On stationary-point hitting time and ergodicity of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 21(68):1–41, 2020.

[11] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

[12] Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github.io*, 2024.

[13] Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/chewi22a.html.

[14] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in r^n. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

[15] Christopher M De Sa, Satyen Kale, Jason D Lee, Ayush Sekhari, and Karthik Sridharan. From gradient flow on population loss to learning with stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:30963–30976, 2022.

[16] Xunpeng Huang, Difan Zou, Yi-An Ma, Hanze Dong, and Tong Zhang. Faster sampling via stochastic gradient proximal sampler. *arXiv preprint arXiv:2405.16734*, 2024.

[17] Prateek Jain, Chi Jin, Sham Kakade, and Praneeth Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488. PMLR, 2017.

[18] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29, 2016.

[19] Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization. *Advances in Neural Information Processing Systems*, 35:19022–19034, 2022.

[20] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[21] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

[22] Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

[23] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[24] Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity. *Bernoulli*, 28(3):1577–1601, 2022.

[25] Alireza Mousavi-Hosseini, Tyler K Farghly, Ye He, Krishna Balasubramanian, and Murat A Erdogdu. Towards a complete analysis of langevin monte carlo: Beyond poincaré inequality. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1–35. PMLR, 2023.

[26] A Nemirovski, A Juditsky, G Lan, and A Shapiro. Stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[27] Goran Peskir and Albert Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006.

[28] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[29] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

[30] Brian Street. What else about... hypoellipticity? *Notices of the AMS*, 65(4), 2018.

[31] Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *Journal of Machine Learning Research*, 24(58):1–47, 2023.

[32] Michalis K Titsias and Omiros Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):749–767, 2018.

[33] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[35] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[36] Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact langevin algorithm and score-based generative models in kl divergence. *arXiv preprint arXiv:2211.01512*, 2022.

[37] Zhipeng Yang. What is hypoellipticity? *https://www.yzpmath.com/post/2020-2/2020-2.pdf*, 2020.

[38] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022. PMLR, 2017.

[39] Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pages 1152–1162. PMLR, 2021.

Contents of Appendix

A Setup for Rest of Paper **11**
A.1 Additional Notation 11

B Proof Sketch **11**
B.1 Proof Sketch 11
B.2 Comparison of our Results to Literature 12

C Proofs for Continuous Time **16**
C.1 Proof of Theorem 3 and Related Results 16

D Proofs for Section 2 **23**
D.1 Proofs of Theorem 6, 7, and 8 23
D.2 Details for Comparison to Literature 42

E Additional Proofs **44**
E.1 Additional Helper Results 44

Appendix A. Setup for Rest of Paper

The appendix is organized as follows. We first give a proof sketch and resulting comparison to literature in Section B. We derive our 'continuous time' result Theorem 3 in Section C. Then in Section D we prove Theorem 6, 7, and 8.

A.1. Additional Notation

In the following, log always denotes natural logarithm. The notation $U([a,b])$ refers to the uniform distribution on $[a,b]$. The notation $\delta_{\mathcal{A}}$ denotes the Dirac Delta on some event $\mathcal{A}$. The notation Γ refers to the Gamma function.

The notation $d(p,\mathcal{A})$ refers to the minimum distance from a point $p \in \mathbb{R}^d$ to a set $\mathcal{A} \subset \mathbb{R}^d$. For a set $\mathcal{U} \subset \mathbb{R}^d$, $\partial\mathcal{U}$ denotes its boundary. For a vector $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w}_i$ refers to its i-th coordinate. For a k-th order tensor operator T and $\mathbf{v}_1,\ldots,\mathbf{v}_k \in \mathbb{R}^d$, $T[\mathbf{v}_1,\ldots,\mathbf{v}_k]$ refers to applying T to the k-th order tensor $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_k$, that is, $\langle T, \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_k \rangle$.

Again, we will refer to the measure on $\mathbb{R}^d$ proportional to $e^{-\beta F(\mathbf{w})}$ by μ_β (the subscript shows the dependence on the temperature, which is crucial for optimization). When we write Z, it refers to the normalizing constant $\int_{\mathbb{R}^d} e^{-\beta F(\mathbf{w})} d\mathbf{w}$ of the measure, unless specified otherwise (so it may change line-to-line if we refer to different measures). For any set $\mathcal{U} \subset \mathbb{R}^d$, let the hitting time of the SDE (7) initialized at $\mathbf{w}$ to $\mathcal{U}$ be $\tau'_{\mathcal{U}}(\mathbf{w})$.

Before we apply results from probability regarding the continuous-time Langevin Diffusion, consider the SDE

$$d\mathbf{w}(t) = -\beta \nabla F(\mathbf{w}(t))dt + \sqrt{2} d\mathbf{B}(t). \tag{7}$$

We refer to this SDE when we directly use results from Cattiaux et al. [9] and Cattiaux and Guillin [8], so that our convention for Poincaré and Log-Sobolev constants will match theirs. Note (7) is equivalent to (2). For a given realization of a Brownian motion driving both SDEs, both SDEs will trace out the same path. However in (7) time passes 'β times faster' than in (2). Hence for any set $\mathcal{U} \subset \mathbb{R}^d$, the hitting time of the SDE (7) to $\mathcal{U}$ is $\frac{1}{\beta}$ (i.e. faster if $\beta \geq 1$) than that of the hitting time of (2) to $\mathcal{U}$, if both SDEs are driven by the same Brownian motion. That is, using our notation, we have $\tau'_{\mathcal{U}} = \frac{1}{\beta}\tau_{\mathcal{U}}$ for all $\mathcal{U} \subset \mathbb{R}^d$.

Appendix B. Proof Sketch

B.1. Proof Sketch

The fundamental idea of how we use Theorem 3 is as follows. Consider $\tau_{\mathcal{A}_\varepsilon}(\mathbf{w}_0)$, the hitting time of GLD/SGLD initialized at $\mathbf{w}_0$ to $\mathcal{A}_\varepsilon$. Denote this by τ for short in the following. Consider the random variable $X := \frac{1}{\tau}\sum_{t=0}^{\tau-1} \lambda\Phi(\mathbf{w}_t)$. Suppose that Φ is L-smooth and L-Hessian Lipschitz. The idea is that, by the following, we can make X relatively small if τ is relatively large, by Taylor expanding Φ to third order and using (8) (it turns out to be possible to control the higher order discretization terms). However, by definition none of $\mathbf{w}_0,\ldots,\mathbf{w}_{\tau-1}$ lie in $\mathcal{A}_\varepsilon$. Clearly X is lower-bounded by λ, since $\Phi \geq 1$. But we just showed X is small if τ is relatively large. This gives contradiction! Hence, we can upper bound τ. This idea, while currently informal, can be made rigorous (using discrete-time Dynkin's formula, Theorem 11.3.1, page 277 of Meyn and Tweedie [23]). See Section D for details. To show how X can be made small, using definition (1), we Taylor

expand Φ to third order (using that it is L-smooth and L-Hessian Lipschitz) to obtain

$$\begin{aligned}\Phi(\mathbf{w}_{t+1}) &= \Phi\Big(\mathbf{w}_t - \eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t\Big)\\ &\le \Phi(\mathbf{w}_t) + \langle -\eta\nabla F(\mathbf{w}_t), \nabla\Phi(\mathbf{w}_t)\rangle + \Big\langle\sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t, \nabla\Phi(\mathbf{w}_t)\Big\rangle\\ &\quad + \frac{1}{2}\Big\langle\nabla^2\Phi(\mathbf{w}_t)\Big(-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t\Big), -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t\Big\rangle\\ &\quad + \frac{L}{6}\Big\|-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t\Big\|^3.\end{aligned}$$

We first use (3), which gives

$$\langle -\eta\nabla F(\mathbf{w}_t), \nabla\Phi(\mathbf{w}_t)\rangle \le -\eta\lambda\Phi(\mathbf{w}_t) - \frac{\eta}{\beta}\Delta\Phi(\mathbf{w}_t).$$

Now, take expectations with respect to $\boldsymbol{\varepsilon}_t$. The term $\Big\langle\sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t, \nabla\Phi(\mathbf{w}_t)\Big\rangle$ disappears, in addition to the cross term $-2\eta\sqrt{2\eta\beta^{-1}}\Big\langle\nabla^2\Phi(\mathbf{w}_t)\boldsymbol{\varepsilon}_t, \nabla F(\mathbf{w}_t)\Big\rangle$ from the second-order term. Note now that

$$\mathbb{E}\Big[\frac{1}{2}\Big\langle\nabla^2\Phi(\mathbf{w}_t)\cdot\sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t, \sqrt{2\eta\beta^{-1}}\boldsymbol{\varepsilon}_t\Big\rangle\Big] = \frac{\eta}{\beta}\Delta\Phi(\mathbf{w}_t).$$

Therefore, the Laplacian terms $\frac{\eta}{\beta}\Delta\Phi(\mathbf{w}_t)$ cancel in the above after taking expectations, and what we obtain is (upon dividing by η)

$$\lambda\mathbb{E}[\Phi(\mathbf{w}_t)] \le \mathbb{E}[\Phi(\mathbf{w}_t)] - \mathbb{E}[\Phi(\mathbf{w}_{t+1})] + \{\text{higher order discretization error terms}\}.$$

Summing and telescoping this relation, and using that Φ is non-negative, we obtain

$$\mathbb{E}[X] = \frac{1}{\tau}\sum_{t=0}^{\tau-1}\lambda\mathbb{E}[\Phi(\mathbf{w}_t)] \le \frac{\Phi(\mathbf{w}_0)}{\tau} + \frac{1}{\tau}\cdot\{\text{higher order discretization error terms}\}.$$

If we can control higher order discretization error terms, which it turns out we can do as discussed in Section D, then if τ is large then $\mathbb{E}[X]$ will be small. But as discussed earlier $X \ge \lambda$ pointwise, hence $\mathbb{E}[X] \ge \lambda$. This lets us control τ, the hitting time of GLD/SGLD to the set $\mathcal{A}_\varepsilon$. One might note this idea of considering the hitting time of SGLD to $\mathcal{A}_\varepsilon$ bears resemblance to the style of proof from Chen et al. [10], Zhang et al. [38]. However, Chen et al. [10], Zhang et al. [38] considered the hitting time to second-order stationary points, and so our results (in addition to the techniques) are fairly different. To fully generalize this, using Lemma 16, this idea can be extended to cover essentially all Lyapunov functions of interest (far beyond when Φ is smooth and Hessian Lipschitz). Due to the stochasticity already present in GLD, our analysis for GLD vs SGLD is extremely similar.

B.2. Comparison of our Results to Literature

In our work, we prove optimization results for GLD/SGLD through *Lyapunov potentials* that are implied by Poincaré Inequalities. To our knowledge, this is the first time such a proof has been used to analyze global convergence of GLD/SGLD. Techniques to analyze sampling of GLD/SGLD generally go through a Girsanov change of measure style argument [4, 13, 29]. This is both fragile, and does not work as well for the more natural case of stochastic gradients (SGLD). In contrast, our

Lyapunov-potential based method is more direct, robust, and naturally handles stochastic gradients. Rather than in sampling or even expected suboptimality, our geometric properties allow us to study the *hitting time* of GLD/SGLD. This leads to better bounds for optimizing non-convex functions, both in general and especially via SGLD. We highlight these improvements as follows; earlier the full statements were given in Subsection 2.2:

1. Theorem 6 and Theorem 7: Consider the case where F is s-Hölder continuous for some $0 \le s \le 1$, there exists $\gamma \ge 2s$ such that for some $m, b > 0$ we have $\langle \mathbf{w}, \nabla F(\mathbf{w}) \rangle \ge m\|\mathbf{w}\|^\gamma - b$, and μ_β satisfies a Poincaré Inequality with constant $\mathsf{C}_{\mathrm{PI}}(\mu_\beta)$ for $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. This is the setting of Balasubramanian et al. [4] and Chewi et al. [13][1]. For both GLD and SGLD, with probability at least $1 - \delta$ we will reach a $\mathbf{w}$ with ε-suboptimality to the global minimum using at most

$$\widetilde{O}\left(\max\left\{d^3 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^3, \frac{d^{2+\frac{s}{2}} \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^{2+\frac{s}{2}}}{\varepsilon^{2+\frac{s}{2}}}\right\} \log(1/\delta)\right)$$

gradient/stochastic gradient evaluations. Here, the $\tilde{O}$ hides universal constants and polynomial log factors in β, d, ε.

2. Theorem 6 and Theorem 7, special case: Consider the case where F is Lipschitz and μ_β satisfies a Poincaré Inequality with constant $\mathsf{C}_{\mathrm{PI}}(\mu_\beta)$ for $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. Here, unlike the above, we do not need lower bounds on the tails of F. For both GLD and SGLD, with probability at least $1 - \delta$ we will reach a $\mathbf{w}$ with ε-suboptimality to the global minimum using at most

$$\widetilde{O}\left(\max\left\{d^3 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^3, \frac{d^2 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^2}{\varepsilon^2}\right\} \log(1/\delta)\right)$$

gradient/stochastic gradient evaluations.

3. Theorem 8: Consider the case when F is smooth (∇F is Lipschitz) and (m, b)-dissipative (that is, there exist $m, b > 0$ such that $\langle \mathbf{w}, \nabla F(\mathbf{w}) \rangle \ge m\|\mathbf{w}\|^2 - b$; see Raginsky et al. [29], Xu et al. [35], Zou et al. [39], and Mou et al. [24] for more details on dissipativeness). By F smooth and dissipative, one can show that μ_β satisfies a Poincaré Inequality for $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$; see Proposition 9 of Raginsky et al. [29]. For both GLD and SGLD, with probability at least $1 - \delta$ we will reach a $\mathbf{w}$ with ε-suboptimality using

$$\widetilde{O}\left(\max\left\{d^3 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^3, \frac{d^2 \max(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 1)^2}{\varepsilon^2}\right\} \log(1/\delta)\right)$$

gradient/stochastic gradient evaluations.

4. Theorem 3: We show a tight connection between μ_β satisfying a Poincaré Inequality and the hitting time of the Langevin Diffusion to the set of ε-suboptimal global minima of F. This is a corollary of literature in probability theory and partial differential equations (PDEs) [8, 9]; we believe we are the first to connect these results to optimization.

1. Although these works do not make our assumption on the tail growth of F, this assumption is mild and natural for non-convex optimization problems motivated by machine learning.

Additionally, sampling and optimization runtime guarantees are *not* the same. As mentioned above, as done in Raginsky et al. [29], Xu et al. [35], Zou et al. [39], and Kinoshita and Suzuki [19], one uses the sampling result to upper bound $\mathbb{E}_{\mathbf{w}\sim\mu_T}[F(\mathbf{w})] - \mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]$. However, techniques to do this can and often do pick up extra dependence in d, ε, and isoperimetric constants such as $\mathsf{C}_{\text{PI}}(\mu_\beta)$, depending on the information metric the sampling guarantee is for. Moreover, for papers such as Chewi et al. [13], Balasubramanian et al. [4], and Yang and Wibisono [36] which study sampling in the constant temperature regime, when converting their results to optimization, we must scale their smoothness parameter by β, which again changes the runtime. Therefore, the runtime for optimization for other papers may not reflect the runtime written in said paper for sampling, as we compute the rate implied by the literature for our task of optimization (which requires low temperature, that is, large $\beta = \Omega\left(\frac{d}{\varepsilon}\right)$): refer to Subsection D.2 for full derivation of the rates of literature.

We summarize the comparison to literature in Table 1 on page 17. Note in our comparisons, we assume other results in literature are done with an $O(1)$ warm-start, which is the most favorable for pre-existing literature (i.e. the least favorable comparisons for our results).[2]

Remark 9 *We additionally note that unlike the strategy for converting sampling to optimization guarantees outlined in Raginsky et al. [29] and followed in Xu et al. [35], Zou et al. [39], and Kinoshita and Suzuki [19], which is to upper bound $\mathbb{E}_{\mathbf{w}\sim\mu_T}[F(\mathbf{w})] - \mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]$ using sampling guarantees, there is a more elegant and faster approach. To our knowledge it has not been mentioned in literature. The approach is to simply sample until $TV\left(\mu_T, \mu_\beta\right) \le 0.1 = O(1)$. For any $\varepsilon > 0$, denote the set $\{\mathbf{w} : F(\mathbf{w}) \le \varepsilon\}$ by $\mathcal{A}_\varepsilon$. For $\beta = \Omega\left(\frac{d}{\varepsilon}\right)$, one can show (see Lemma 14) that $\mu_\beta(\mathcal{A}_\varepsilon) \ge 0.5$. Therefore $\mu_T(\mathcal{A}_\varepsilon) \ge 0.4$ by definition of TV distance – that is, the probability our iterate $\mathbf{w}_T \in \mathcal{A}_\varepsilon$ is at least 0.4. When $\beta = o\left(\frac{d}{\varepsilon}\right)$, $\mu_\beta(\mathcal{A}_\varepsilon)$ can be exponentially small in d as seen from the Gaussian example, so this strategy still requires large β. Table 1 on page 17 shows the results using the strategy from Raginsky et al. [29] known in the literature, but below we discuss the comparisons using both methods. While the rates of literature do improve, our rates are still more favorable.*

Here we expand on these comparisons:

1. Consider the case where F is s-Hölder continuous, there exists $\gamma \ge 2s$ such that $\langle \mathbf{w}, \nabla F(\mathbf{w})\rangle \ge m\|\mathbf{w}\|^\gamma - b$, and μ_β satisfies a Poincaré Inequality for $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. This case has been studied in Chewi et al. [13] and Balasubramanian et al. [4].

   In the GLD case, using the strategy of Raginsky et al. [29], Theorem 7 of Chewi et al. [13] obtains a rate of $\tilde{O}\left(\frac{d^{2+\frac{3}{s}}\mathsf{C}_{\text{PI}}(\mu_\beta)^{1+\frac{1}{s}}}{\varepsilon^{\frac{4}{s}}}\right)$. Following the method suggested by Remark 9, the rate becomes $\tilde{O}\left(\frac{d^{1+\frac{2}{s}}\mathsf{C}_{\text{PI}}(\mu_\beta)^{1+\frac{1}{s}}}{\varepsilon^{\frac{2}{s}}}\right)$. When $s \le \frac{1}{2}$, our result from Theorem 6 is always better or equal to both of these in all parameters. When $s \in (\frac{1}{2}, 1]$, our result from Theorem 6 is superior to the rate obtained following the strategy of Raginsky et al. [29] when $\varepsilon < \frac{d^{\frac{1}{4}(3-s)}}{\mathsf{C}_{\text{PI}}(\mu_\beta)^{\frac{1}{2}(s-\frac{1}{2})}}$. Theorem 6 is superior to the rate obtained following Remark 9 when $\varepsilon < \frac{d^{1-s}}{\mathsf{C}_{\text{PI}}(\mu_\beta)^{s-\frac{1}{2}}}$.

2. For simplicity, in our comparisons we assume $\mathsf{C}_{\text{PI}}(\mu_\beta) = \tilde{\Omega}(1)$, which is generally the case (for example this is true if μ_β is isotropic and F is convex, and perturbations to F will increase $\mathsf{C}_{\text{PI}}(\mu_\beta)$). All explicit expressions for our rates and those of the literature are given, so one can still perform these comparisons when $\mathsf{C}_{\text{PI}}(\mu_\beta) = o(1)$.

When $s \leq \frac{1}{2}$, Corollary 19 of Balasubramanian et al. [4] improves on Chewi et al. [13]. Using the strategy of [29], the rate is $\tilde{O}\left(\frac{d^{\frac{6}{1+s}+8-3s}\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^{\frac{16-2s}{1+s}}}\right)$, which is using $s \leq \frac{1}{2}$ at least $\tilde{O}\left(\frac{d^{10.5}\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^{10}}\right)$. Following Remark 9, the rate becomes $\tilde{O}\left(\frac{d^{3+\frac{6}{1+s}-2s}\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^{\frac{6}{1+s}}}\right)$, which using $s \leq \frac{1}{2}$ is at least $\tilde{O}\left(\frac{d^8\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^6}\right)$. Our result from Theorem 6 is superior or equal to both of these in all parameters, oftentimes by a significant amount.

In the SGLD case, our rate from Theorem 7 is the *first finite gradient complexity guarantee.*

2. Consider the case when F is Lipschitz and μ_β satisfies a Poincaré Inequality for $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. This has not been well-studied in the sampling or optimization literature, and the only work we know of with finite gradient complexity is Balasubramanian et al. [4], namely $s = 0$ in Corollary 19, in the GLD case. The rate here using the strategy of Raginsky et al. [29] is $\widetilde{O}\left(\frac{d^{14}\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^{16}}\right)$, or following Remark 9, is $\widetilde{O}\left(\frac{d^9\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^6}\right)$. Our rate from Theorem 6 is superior or equal to both of these in every parameter, oftentimes by a significant amount. Again, Theorem 7 is the first finite gradient complexity guarantee for the SGLD case.

3. Consider the case for SGLD and when F is smooth and dissipative, which has been well-studied in the works Raginsky et al. [29], Xu et al. [35], Zou et al. [39], Yang and Wibisono [36], and Kinoshita and Suzuki [19]. Theorem 1 of Raginsky et al. [29] requires gradient noise δ to be potentially exponentially small in d, which does not make sense (we only require gradient noise of constant order, which is more realistic). Similarly, the relevant result of Yang and Wibisono [36] which are Theorems 2, 3 also require gradient noise σ_F to be potentially exponentiall small in d. Their gradient noise ε_{mgf} is lower bounded by our σ_F from Assumption 4.

   In particular their results give sampling results in KL divergence that, with constant gradient noise/score estimation error, are of order at least $O\left(\sigma_F^2\mathsf{C}_{\text{LSI}}(\mu_\beta)\right)$ where $\mathsf{C}_{\text{LSI}}(\mu_\beta)$ denotes the Log-Sobolev constant. A Log-Sobolev Inequality is a stronger functional inequality which implies a Poincaré Inequality, with the corresponding Log-Sobolev constant $\mathsf{C}_{\text{LSI}}(\mu_\beta) \geq \mathsf{C}_{\text{PI}}(\mu_\beta)$. $\mathsf{C}_{\text{LSI}}(\mu_\beta)$ is worst-case exponentially large in d (in fact in β), while using either the strategy of Raginsky et al. [29] or Remark 9, we need at most $O(1)$ sampling error in KL divergence, necessitating that σ_F be tiny for their results to be meaningful for optimization. It should be noted, however, that Yang and Wibisono [36] does not assume dissipativity; it only assumes a Log-Sobolev Inequality. Recall the stronger Log-Sobolev Inequality is implied by dissipativeness [29], although dissipativeness (i.e. quadratic tail growth) can also be thought of as a canonical case of a Log-Sobolev Inequality.

   For using the results from Xu et al. [35] and Zou et al. [39], we must account for total gradient complexity for a stochastic gradient oracle with $O(1)$ noise. After doing so we obtain $\widetilde{O}\left(\frac{d^7}{\varepsilon^5\lambda_*^5}\right)$ for Xu et al. [35] and a rate of $\widetilde{O}\left(\frac{d^5}{\lambda_*^4\varepsilon^4}\right)$ for variance-reduced SGLD from Xu et al. [35]. Here, λ_* is a quantity similar to $\frac{1}{\mathsf{C}_{\text{PI}}(\mu_\beta)}$ (but not directly comparable)[3]. Our rate from Theorem 8 thus is generally superior to both of these in every parameter. (The results of

3. It is the spectral gap of discrete-time SGLD.

Xu et al. [35], being phrased directly in optimization, can't be directly improved using Remark 9.) The rate from Zou et al. [39] is, using Cheeger's Inequality, at least $\widetilde{O}\left(\frac{d^8 \mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^4}\right)$ using the strategy of Raginsky et al. [29]. Thus our rate is superior when $\varepsilon < \frac{d^{1.25}}{\mathsf{C}_{\text{PI}}(\mu_\beta)^{0.25}}$. Following Remark 9, Zou et al. [39] yields a rate of $\widetilde{O}\left(\frac{d^6 \mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^2}\right)$; our rate is superior when $\varepsilon < \frac{d^{1.5}}{\mathsf{C}_{\text{PI}}(\mu_\beta)^{0.5}}$.

For Kinoshita and Suzuki [19], we directly apply their Theorem 3 which applies their results on sampling to non-convex optimization to obtain a rate of at least $\tilde{O}\left(\frac{d^3 \mathsf{C}_{\text{LSI}}(\mu_\beta)^3}{\varepsilon^3}\right) \geq \tilde{O}\left(\frac{d^3 \mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^3}\right)$ (their γ is our β), which our rate is always superior to. Using their sampling result and then Remark 9 improves their result to $\tilde{O}\left(\frac{d^3 \mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^2}\right)$; our rate is superior when $\varepsilon < \frac{1}{\mathsf{C}_{\text{PI}}(\mu_\beta)^{0.5}}$. We note their result uses a variance reduction technique, analogous to variance reduction in convex optimization, to discretize the Langevin Diffusion in a slightly different way. We also note that Kinoshita and Suzuki [19] also assumes dissipativity for optimization, see the statement of their Theorem 3.

4. We additionally touch on other discretizations of the Langevin Diffusion. To our knowledge, the only other discretization of (2) successful beyond log-concavity is the Proximal Sampler first introduced in Lee et al. [21], Titsias and Papaspiliopoulos [32]. With exact gradients, Altschuler and Chewi [1] showed it succeeds under a Poincaré Inequality when F is smooth; the Proximal Sampler can only be implementable with smoothness for non-convex F. In the stochastic gradient setting, the only work we are aware of showing its success is Theorems 4.1 and 4.2 of Huang et al. [16], showing the Proximal Sampler succeeds under smoothness and a Log-Sobolev Inequality (which is satisfied in the smooth and dissipative setting as shown in Proposition 9 of Raginsky et al. [29]). The rate from there is, using the strategy of Raginsky et al. [29], $\tilde{O}\left(\frac{d^{5.5} \mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^5}\right)$. Or following Remark 9, the rate is $\tilde{O}\left(\frac{d^{3.5} \mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^3}\right)$. Our rate from Theorem 8 is superior or equal in every parameter, often by a significant amount.

Appendix C. Proofs for Continuous Time

C.1. Proof of Theorem 3 and Related Results

Now we restate Theorem 3 formally here. Note Theorem 10 requires us to control $\mu(\mathcal{A}_\varepsilon)$ in the λ from Theorem 3, for which we need Lemma 14. Lemma 14 is precisely where we need $\varepsilon = \widetilde{\Omega}\left(\frac{d}{\beta}\right)$. This leads to consistency between our results and our discussion from the Introduction. We defer Lemma 14 to later in this section and note Theorem 3 follows immediately from combining Theorem 10 and Lemma 14.

Theorem 10 *Assume that μ_β satisfies a Poincaré inequality with constant $\mathsf{C}_{\text{PI}}(\mu_\beta)$. Then there exists a non-negative Lyapunov function Φ differentiable to all others such that on $\mathcal{A}_\varepsilon^c$, we have $\Phi \geq 1$ and*

$$-\langle \nabla F(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle + \frac{1}{\beta} \Delta \Phi(\mathbf{w}) \leq -\lambda \Phi(\mathbf{w}), \tag{8}$$

Problem Setting	**Our Result**	**Best in Literature**
GLD Poincaré & Lipschitz	$\widetilde{O}\left(\max\left\{d^3\mathsf{C}_{\text{PI}}(\mu_\beta)^3, \frac{d^2\mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^2}\right\}\right)$	$\widetilde{O}\left(\frac{d^{14}\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^{16}}\right)$ [4]
SGLD Poincaré & Lipschitz	$\widetilde{O}\left(\max\left\{d^3\mathsf{C}_{\text{PI}}(\mu_\beta)^3, \frac{d^2\mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^2}\right\}\right)$	No finite guarantee
SGLD smooth & dissipative	$\widetilde{O}\left(\max\left\{d^3\mathsf{C}_{\text{PI}}(\mu_\beta)^3, \frac{d^2\mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^2}\right\}\right)$	$\widetilde{O}\left(\min\left\{\frac{d^8\mathsf{C}_{\text{PI}}(\mu_\beta)^2}{\varepsilon^4}, \frac{d^7}{\varepsilon^5\lambda_*^5}, \frac{d^3\mathsf{C}_{\text{PI}}(\mu_\beta)^3}{\varepsilon^3}\right\}\right)$ [19, 35, 39]

Table 1: Gradient complexity comparisons. We compare our optimization results to those obtained by proving sampling results and then converting back to optimization using the strategy known from pre-existing literature, from Raginsky et al. [29]. In the table, d refers to dimension and ε refers to tolerance. $\beta = \widetilde{\Theta}\left(\frac{d}{\varepsilon}\right)$, and $\mathsf{C}_{\text{PI}}(\mu_\beta)$ denotes the Poincaré constant of μ_β. λ_* is a spectral gap comparable to $\frac{1}{\mathsf{C}_{\text{PI}}(\mu_\beta)}$.

where

$$\lambda = \frac{1}{\beta}\mu_\beta(\mathcal{A}_\varepsilon)\min\left(\frac{1}{4\mathsf{C}_{\text{PI}}(\mu_\beta)}, \frac{1}{8}\right).$$

In fact, on $\mathcal{A}_\varepsilon^c$, Φ has the explicit form

$$\Phi(\mathbf{w}') = \mathbb{E}_{\mathbf{w}'}[\exp(\lambda\tau_{\mathcal{A}_\varepsilon})].$$

Proof We first need to introduce some concepts from Markov processes and Partial Differential Equations (PDEs). First, we introduce the concept of the (infinitesimal) generator of a Markov process, which will make this exposition much more natural. We give only what is needed for our proof and refer the reader to Chewi [12] for more details.

Definition 11 *The (infinitesimal) generator of a Markov process $\mathbf{w}(t)$ is the operator $\mathcal{L}$ defined on all (sufficiently differentiable) functions f by*

$$\mathcal{L}f(\mathbf{w}) = \lim_{t\to 0}\frac{\mathbb{E}[f(\mathbf{w}(t))] - f(\mathbf{w})}{t}.$$

It is well-known and can be easily checked that for the Langevin Diffusion given in the form (7), the generator

$$\mathcal{L}f(\mathbf{w}) = -\langle\beta\nabla F(\mathbf{w}), \nabla f(\mathbf{w})\rangle + \Delta f(\mathbf{w}). \tag{9}$$

For example, this calculation can be found in Example 1.2.4 of Chewi et al. [13].

Note the similarity of the above to (3). This is no coincidence; our discrete-time proofs, specifically Lemma 17 and Lemma 19, are essentially re-deriving the generator of the Langevin diffusion. In Lemma 17 and Lemma 19 we Taylor expand to third order (so we have the full second order quadratic form); intuitively that is all that is needed by Itô's Lemma.

We also need to introduce the idea of symmetry of the measure μ_β with respect to the stochastic process. In particular, we say μ_β is *symmetric* (with respect to the Langevin Diffusion (7)) if for all infinitely differentiable f, g,

$$\int f\mathcal{L}g\mathrm{d}\mu_\beta = \int \mathcal{L}fg\mathrm{d}\mu_\beta.$$

Here $\mathcal{L}$ refers to the generator (9) for the Langevin Diffusion (7). It is well-known and can be easily checked again that μ_β is symmetric, see Example 1.2.18 of Chewi et al. [13] or the discussion on page 3 of Cattiaux and Guillin [8].

Finally, we need to introduce some ideas from PDE theory. Consider a second-order differential operator

$$\mathcal{P} = \frac{1}{2}\sum_{1\le i<j\le d} a_{ij}\frac{\partial^2}{\partial \mathbf{w}_i \partial \mathbf{w}_j} + \sum_{1\le i\le d} b_i \frac{\partial}{\partial \mathbf{w}_i} + c.$$

The following definitions generalize far beyond second-order differential operators, but this is all we need for our work. We say that $\mathcal{P}$ is *elliptic* if, for every $\mathbf{w} \neq 0 \in \mathbb{R}^d$,

$$\sum_{1\le i,j\le d} a_{ij}\mathbf{w}_i\mathbf{w}_j \neq 0.$$

We say $\mathcal{P}$ is uniformly elliptic if we can write

$$\mathcal{P} = \frac{1}{2}\sum_{1\le i<j\le d} \left(\sigma\sigma^T\right)_{ij}\frac{\partial^2}{\partial \mathbf{w}_i \partial \mathbf{w}_j} + \sum_{1\le i\le d} b_i \frac{\partial}{\partial \mathbf{w}_i} + c,$$

for some $\sigma \in \mathbb{R}^d$ where uniformly on $\mathbb{R}^d$ we have

$$\sigma\sigma^T \succcurlyeq a > 0$$

in the PSD order [8, 30].

A canonical example of $\mathcal{P}$ that is uniformly elliptic is the Laplacian, where $a_{ij} = 2\delta_{i=j}$ [37]. Beyond this, note for the Langevin Diffusion (7), we have $a_{ij} = 2\delta_{i=j}$ as well, from (9). Thus, it is clear that $\mathcal{L}$ for the Langevin Diffusion (7) is uniformly elliptic.

Ellipticity is well-known to imply that solutions u to the Dirichlet problem $\mathcal{P}u = f$ in some open domain $\Omega \subset \mathbb{R}^d$ are smooth, which is all we need here [37].[4] Ellipticity implies maximal hypoellipticity, which in turn implies strong hypoellipticity/Hormander's condition from Cattiaux et al. [9], as discussed in Yang [37]. Thus uniform ellipticity implies strong uniform hypoellipticity as defined in Cattiaux et al. [9]. Using the results of Cattiaux et al. [9] requires strong uniform hypoellipticity and symmetry with respect to the stochastic process, and Cattiaux and Guillin [8] requires uniform ellipticity and symmetry. We have uniform ellipticity and symmetry, and so can use all those results.

Now we move to the main proof. Our main tool is Theorem 2.1 of Cattiaux and Guillin [8], which connects Poincaré Inequalities to more explicit geometric conditions that we can use in an 'optimization-styled' proof analysis later.[5] Specialized to the Langevin Diffusion (7) on the domain $\mathcal{D} = \mathbb{R}^d$, it states the following:

4. For this, ellipticity is sufficient but not necessary. The loosest such condition for this is hypoellipticity [30, 37], which is not relevant for this work.
5. We presume here F is sufficiently differentiable to use the results of Cattiaux et al. [9] and Cattiaux and Guillin [8], for example this holds if F is infinitely differentiable. The careful reader will notice that F can be approximated by

Theorem 12 (Theorem 2.1 of Cattiaux and Guillin [8]) *Suppose that μ_β satisfies a Poincaré Inequality with constant $\mathsf{C}_{\text{PI}}(\mu_\beta)$. Then for all open subsets $\mathcal{U}$ of $\mathbb{R}^d$, there exists a function Φ differentiable to all orders such that on $\mathcal{U}^c$ we have $\Phi \geq \delta' > 0$ for some δ', as well as*

$$\mathcal{L}\Phi(\mathbf{w}) = -\langle \beta \nabla F(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle + \Delta \Phi(\mathbf{w}) \leq -\lambda' \Phi(\mathbf{w}), \tag{10}$$

where $\lambda' = \mu_\beta(\mathcal{U}) \min\Big(\frac{1}{4\mathsf{C}_{\text{PI}}(\mu_\beta)}, \frac{1}{8}\Big)$.

Note to prove this result in $\mathcal{D} = \mathbb{R}^d$ all that is needed is ellipticity, which is clearly satisfied here in the case of the Langevin diffusion (following the discussion on page 9 of Cattiaux and Guillin [8]). Hence, applying Theorem 12 with $\mathcal{U} = \{\mathbf{w} : F(\mathbf{w}) < \varepsilon\}$ which is clearly open, this gives the existence of such a Φ.

Suppose $\{\mathbf{w} : \Phi(\mathbf{w}) \leq \frac{\delta'}{2}\} \neq \varnothing$. In this case, consider $\{\mathbf{w} : \Phi(\mathbf{w}) \leq \frac{\delta'}{2}\} \subset \{\mathbf{w} : \Phi(\mathbf{w}) < \frac{3\delta'}{4}\} \subset \{\mathbf{w} : F(\mathbf{w}) < \varepsilon\}$. Apply the standard construction of bump functions to the compact set $\{\mathbf{w} : \Phi(\mathbf{w}) \leq \frac{\delta'}{2}\}$ contained in the open set $\{\mathbf{w} : \Phi(\mathbf{w}) < \frac{3\delta'}{4}\}$ to obtain a function χ differentiable to all orders supported on $\{\mathbf{w} : \Phi(\mathbf{w}) < \frac{3\delta'}{4}\}$ and identically 1 on $\{\mathbf{w} : \Phi(\mathbf{w}) \leq \frac{\delta'}{2}\}$. Let $B = \inf \Phi \leq \frac{\delta'}{2}$. It is easy to check that $\Phi + \Big(\frac{\delta'}{2} + \max(0, -B)\Big)\chi \geq \frac{\delta'}{2}$, and differentiable to all orders as Φ and χ are, and is identical to Φ on $\{\mathbf{w} : F(\mathbf{w}) \geq \varepsilon\}$. Taking $\Phi \leftarrow \Phi + \Big(\frac{\delta'}{2} + \max(0, -B)\Big)\chi \geq \frac{\delta'}{2}$, this gives us the existence of $\Phi \geq \frac{\delta'}{2}$ differentiable to all orders where we know on $\{\mathbf{w} : F(\mathbf{w}) \geq \varepsilon\}$, it satisfies (10).

Notice $\mu_\beta(\{\mathbf{w} : F(\mathbf{w}) < \varepsilon\}) = \mu_\beta(\mathcal{A}_\varepsilon)$, since $\mu_\beta(\partial \mathcal{A}_\varepsilon) = \mu_\beta(\{\mathbf{w} : F(\mathbf{w}) = \varepsilon\})$ is simply a positive constant times the Lebesgue measure of $\partial \mathcal{A}_\varepsilon$, and hence is 0. Therefore we know for this Φ,

$$\mathcal{L}\Phi(\mathbf{w}) = -\langle \beta \nabla F(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle + \Delta \Phi(\mathbf{w}) \leq -\lambda' \Phi(\mathbf{w}) = -\beta\lambda \Phi(\mathbf{w}). \tag{11}$$

We claim with such a Φ, the moment generating function $\mathbb{E}_{\mathbf{w}'}\big[\exp\big(\beta\lambda \tau'_{\mathcal{A}_\varepsilon}\big)\big]$ exists (i.e. is finite). The argument is done explicitly on page 8 of Cattiaux et al. [9] (connectivity of $\mathcal{A}$ is not necessary, as one will see below). We write it here explicitly here for the reader. Clearly this MGF is finite for $\mathbf{w}' \in \mathcal{A}_\varepsilon$, so consider any $\mathbf{w}' \in \mathcal{A}_\varepsilon^c$. Consider any $t < \infty$, any $R < \infty$ and consider the hitting time $\tau'_{\mathcal{A}_\varepsilon \cup \mathbb{B}(\vec{\mathbf{0}}, R)^c}$. Denote $\tau'_{t,\varepsilon,R} := t \wedge \tau'_{\mathcal{A}_\varepsilon \cup \mathbb{B}(\vec{\mathbf{0}}, R)^c}$ for short, which is clearly a stopping time. Apply Dynkin's Formula to the map $(s, \mathbf{w}) \to e^{\beta\lambda s}\Phi(\mathbf{w})$ with the stopping time $\tau'_{t,\varepsilon,R}$; thus for all $s < \tau'_{t,\varepsilon,R}$, we know $\Phi(\mathbf{w}(s))$ satisfies (11). We obtain:

$$\begin{aligned}
\frac{\delta'}{2}\mathbb{E}_{\mathbf{w}'}\big[\exp\big(\beta\lambda\tau'_{t,\varepsilon,R}\big)\big] &\leq \mathbb{E}_{\mathbf{w}'}\big[\exp\big(\beta\lambda\tau'_{t,\varepsilon,R}\big)\Phi\big(\mathbf{w}(\tau'_{t,\varepsilon,R})\big)\big] \\
&= \Phi(\mathbf{w}') + \mathbb{E}_{\mathbf{w}'}\left[\int_0^{\tau'_{t,\varepsilon,R}} \exp(\beta\lambda s)(\beta\lambda\Phi(\mathbf{w}(s)) + \mathcal{L}\Phi(\mathbf{w}(s)))\mathrm{d}s\right] \\
&\leq \Phi(\mathbf{w}') + \mathbb{E}_{\mathbf{w}'}\left[\int_0^{\tau'_{t,\varepsilon,R}} \exp(\beta\lambda s)(\beta\lambda\Phi(\mathbf{w}(s)) - \beta\lambda\Phi(\mathbf{w}(s)))\mathrm{d}s\right]
\end{aligned}$$

an infinitely differentiable function to arbitrary precision. We also assume the boundary $\partial\mathcal{A}_\varepsilon = \{\mathbf{w} : F(\mathbf{w}) = \varepsilon\}$ is differentiable to all orders, non-characteristic for (7) in the sense described in Cattiaux et al. [9] and Cattiaux and Guillin [8], and has Lebesgue measure 0. In the F Lipschitz case we assume this set is bounded and hence compact; boundedness and hence compactness follows from Assumption 3 in all other cases. Since we can approximate F by an infinitely differentiable function to arbitrary precision, this boundary in turn will be infinitely differentiable.

$$= \Phi(\mathbf{w}').$$

For justification, the first line above follows as $\Phi(\mathbf{w}) \geq \frac{\delta'}{2}$. Dynkin's Formula and then Chain Rule and Itô's Lemma are used in the second line (an analogous calculation is done formally on page 121, Peskir and Shiryaev [27]). The third line uses the geometric condition (11) that we know $\Phi(\mathbf{w}(s))$ satisfies for $s < \tau'_{t,\varepsilon,R}$.

Thus, we have for all $t < \infty$, $R < \infty$ that

$$\mathbb{E}_{\mathbf{w}'}\left[\exp\left(\beta\lambda\tau'_{t,\varepsilon,R}\right)\right] \leq \frac{2\Phi(\mathbf{w}')}{\delta'} < \infty.$$

Recalling $\delta' > 0$ is independent of R, t, letting first $R \to \infty$ and then $t \to \infty$, Dominated Convergence Theorem gives the result $\mathbb{E}_{\mathbf{w}'}\left[\exp\left(\beta\lambda\tau'_{\mathcal{A}_\varepsilon}\right)\right] \leq \frac{2\Phi(\mathbf{w}')}{\delta'} < \infty$ (since the right hand side above is a finite upper bound independent of R, t).

We now claim the moment generating function $\mathbb{E}_{\mathbf{w}'}\left[\exp\left(\beta\lambda\tau'_{\mathcal{A}_\varepsilon}\right)\right]$, which we now know exists, satisfies (8). In fact this holds as an *equality* on $\mathcal{A}_\varepsilon^c$ (although we don't need this). This is shown on page 8 of Cattiaux et al. [9] and discussed on page 12 of Cattiaux and Guillin [8]. Thus, here we just give a sketch; it follows by literature on PDEs, specifically Dirichlet problems. The result used to prove this is result 1 of Section 7.2 of Peskir and Shiryaev [27]:

Theorem 13 (Result 1 of Section 7.2 of Peskir and Shiryaev [27]) *Let $\mathcal{U}$ be a bounded, open subset of $\mathbb{R}^d$. Given a continuous function $L : \mathcal{U} \to \mathbb{R}$ define*

$$F(\mathbf{w}) = \mathbb{E}_{\mathbf{w}}\left[\int_0^{\tau'_{\mathcal{U}^c}} L(\mathbf{w}(t))\mathrm{d}t\right],$$

where $\mathbf{w}(t)$ here denotes the iterates of any diffusion process and $\tau'_{\mathcal{U}^c}$ denotes the hitting time of $\mathbf{w}(t)$ to $\mathcal{U}^c$. Then F solves the Dirichlet problem

$$\mathcal{L}F = -L \text{ in } \mathcal{U}, F|_{\partial\mathcal{U}} = 0.$$

Here, $\mathcal{L}$ is the generator of this diffusion.

Consider any $R < \infty$. Consider $\mathcal{U}_{\varepsilon,R} := \mathcal{A}_\varepsilon^c \cap \{\mathbf{w} : \|\mathbf{w}\| < R\}$, which is clearly open. Now, we apply the same reasoning as Result 4 of Section 7.2 of Peskir and Shiryaev [27] (the *killed* version of the Dirichlet problem), except now we want to study the *created* version of the Dirichlet problem[6]. There is not much difference, thus we just give a sketch and refer the reader to Result 4 of Section 7.2 of Peskir and Shiryaev [27] and again page 8 of Cattiaux et al. [9]. Let $L \equiv \beta\lambda$ be a constant function and now let $\mathbf{w}(t)$ denotes the iterates of the Langevin diffusion (7). Consider

$$F(\mathbf{w}) = \mathbb{E}_{\mathbf{w}}\left[\int_0^{\tau'_{\mathcal{U}^c_{\varepsilon,R}}} e^{\beta\lambda t}\beta\lambda\mathrm{d}t\right] = \mathbb{E}_{\mathbf{w}}\left[\int_0^{\tau'_{\mathcal{U}^c_{\varepsilon,R}}} e^{\beta\lambda t}L(\mathbf{w}(t))\mathrm{d}t\right].$$

where $\tau'_{\mathcal{U}^c}$ now is consistent with our definition from Section A, being for the Langevin Diffusion (7). Observe that

$$F(\mathbf{w}) + 1 = \mathbb{E}_{\mathbf{w}}\left[1 + \int_0^{\tau'_{\mathcal{U}^c_{\varepsilon,R}}} e^{\beta\lambda t}\beta\lambda\mathrm{d}t\right] = \mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{U}^c_{\varepsilon,R}}}\right] \leq \mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{A}_\varepsilon}}\right] < \infty,$$

6. See Section 5.4, [27].

since $\frac{\partial}{\partial t}e^{\beta\lambda t} = \beta\lambda e^{\beta\lambda t}$, $\tau'_{\mathcal{U}^c_{\varepsilon,R}} \leq \tau'_{\mathcal{A}_\varepsilon}$. Hence, $F(\mathbf{w}) < \infty$ and so we may continue to analyze it.

Now consider $\tilde{\mathbf{w}}(t) := e^{\beta\lambda t}\mathbf{w}(t)$ (the created process). By the same reasoning as in Result 4 of Section 7.2 of Peskir and Shiryaev [27] but for the created rather than killed process, we have $F(\mathbf{w}) = \mathbb{E}_{\mathbf{w}}\left[\int_0^{\tilde{\tau}'_{\mathcal{U}^c_{\varepsilon,R}}} L(\mathbf{w}(t))\mathrm{d}t\right]$ where $\tilde{\tau}'_{\mathcal{U}^c_{\varepsilon,R}}$ denotes the hitting time of $\tilde{\mathbf{w}}(t)$ to $\mathcal{U}^c_{\varepsilon,R}$. Let the generator of $\tilde{\mathbf{w}}(t)$ be $\tilde{\mathcal{L}}$. Now, Theorem 13 implies that $F(\mathbf{w})$ solves the Dirichlet problem

$$\tilde{\mathcal{L}}F = -L = -\beta\lambda \text{ in } \mathcal{U}_{\varepsilon,R}, F|_{\partial\mathcal{U}_{\varepsilon,R}} = 0.$$

It can be readily seen that by Chain Rule that $\tilde{\mathcal{L}} = \mathcal{L} + \beta\lambda$; this calculation is done formally on page 121, Peskir and Shiryaev [27]. Therefore, we have

$$-\beta\lambda = \tilde{\mathcal{L}}F = \mathcal{L}F + \beta\lambda F \text{ in } \mathcal{U}_{\varepsilon,R}, F|_{\partial\mathcal{U}_{\varepsilon,R}} = 0.$$

Therefore, $\Phi_R = F + 1$ satisfies (note $\mathcal{L}\Phi_R = \mathcal{L}F$)

$$\mathcal{L}\Phi_R = \mathcal{L}F = -\beta\lambda(F+1) = -\beta\lambda\Phi_R \text{ in } \mathcal{U}_{\varepsilon,R}, \Phi_R|_{\partial\mathcal{U}_{\varepsilon,R}} = 1.$$

Note we showed earlier

$$\Phi_R(\mathbf{w}) = F(\mathbf{w}) + 1 = \mathbb{E}_{\mathbf{w}}\left[1 + \int_0^{\tau'_{\mathcal{U}^c_{\varepsilon,R}}} e^{\beta\lambda t}\beta\lambda\mathrm{d}t\right] = \mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{U}^c_{\varepsilon,R}}}\right].$$

Finally, since we've already shown $\mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{A}_\varepsilon}}\right] < \infty$, the same argument of page 8 of Cattiaux et al. [9] shows that the pointwise limit

$$\Phi(\mathbf{w}) := \mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{A}_\varepsilon}}\right] = \lim_{R\to\infty}\mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{U}^c_{\varepsilon,R}}}\right]$$

exists and solves the Dirichlet Problem

$$\mathcal{L}\Phi = -\beta\lambda\Phi \text{ in } \lim_{R\to\infty}\mathcal{U}_{\varepsilon,R} \cap \{\mathbf{w} : \|\mathbf{w}\| < R\} = \mathcal{A}^c_\varepsilon.$$

Thus, it satisfies (8). Moreover, since $\mathcal{L}$ is elliptic (and therefore hypoelliptic), the resulting solution

$$\Phi(\mathbf{w}) = \mathbb{E}_{\mathbf{w}}\left[e^{\beta\lambda\tau'_{\mathcal{A}_\varepsilon}}\right]$$

is differentiable to all orders in $\lim_{R\to\infty}\mathcal{A}^c_\varepsilon \cap \{\mathbf{w} : \|\mathbf{w}\| < R\} = \mathcal{A}^c_\varepsilon$. Note since the quantity in the exponential is always non-negative pointwise, $\Phi(\mathbf{w}) \geq 1$ on $\mathcal{A}^c_\varepsilon$.

Since the boundary $\partial\mathcal{A}_\varepsilon = \{\mathbf{w} : F(\mathbf{w}) = \varepsilon\}$ is compact and differentiable to all orders, through a standard compactness and $\delta - \varepsilon$ argument we can show by defining

$$\Phi(\mathbf{w}) = \lim_{\mathbf{w}'\to\mathbf{w}, \mathbf{w}'\in\mathcal{A}^c_\varepsilon}\Phi(\mathbf{w}') \text{ for all } \mathbf{w} \in \partial\mathcal{A}_\varepsilon,$$

the resulting Φ is differentiable to all orders on $\mathcal{A}^c_\varepsilon \cup \partial\mathcal{A}_\varepsilon$ (when we define derivatives as the limits coming from outside $\mathcal{A}^c_\varepsilon$). (Compactness here is important.) As $\mathcal{A}^c_\varepsilon \cup \partial\mathcal{A}_\varepsilon$ is closed, applying Whitney's Extension Theorem as mentioned in [9], Φ above can be extended to a function differentiable to all orders on all of $\mathbb{R}^d$ so that (8) holds on $\{\mathbf{w} : F(\mathbf{w}) \geq \varepsilon\}$. Note $\Phi \geq 1$ on $\{\mathbf{w} : F(\mathbf{w}) \geq \varepsilon\}$.

Suppose the resulting Φ from the extension was not non-negative. Let $B := \inf \Phi < 0$. Observe $\{\mathbf{w} : \Phi(\mathbf{w}) \le 0\} \subset \{\mathbf{w} : \Phi(\mathbf{w}) < \frac{1}{2}\} \subset \mathcal{A}_\varepsilon$. Apply the standard construction of bump functions to the compact set $\{\mathbf{w} : \Phi(\mathbf{w}) \le 0\}$ contained in the open set $\{\mathbf{w} : \Phi(\mathbf{w}) < \frac{1}{2}\}$ to obtain a function χ differentiable to all orders supported on $\{\mathbf{w} : \Phi(\mathbf{w}) < \frac{1}{2}\}$ and identically 1 on $\{\mathbf{w} : \Phi(\mathbf{w}) \le 0\}$. Then $\Phi - B\chi$ is non-negative (recall $B < 0$) and differentiable to all orders, and is identical to Φ on $\mathcal{A}_\varepsilon^c$. Taking $\Phi \leftarrow \Phi - B\chi$, this gives us the existence of $\Phi \ge 0$ differentiable to all orders where we have its explicit form and know it satisfies (11) and therefore (8) (upon dividing both sides by $\beta > 0$) on $\mathcal{A}_\varepsilon^c$.

To conclude, note from our remarks from Section A that

$$\tau'_{\mathcal{A}_\varepsilon}(\mathbf{w}') = \frac{1}{\beta}\tau_{\mathcal{A}_\varepsilon}(\mathbf{w}').$$

Therefore on $\mathcal{A}_\varepsilon^c$ we can also write

$$\Phi(\mathbf{w}') = \mathbb{E}_{\mathbf{w}'}[\exp(\lambda\tau_{\mathcal{A}_\varepsilon})] \ge 1.$$

This completes the proof. ■

Now we prove Lemma 14.

Lemma 14 *Suppose F satisfies Assumption 2 and μ_β has finite second moment $S < \infty$. Then for $\varepsilon \ge \frac{2d}{\beta}\log(4\pi e\beta LdS)$, we have $\mu_\beta(\mathcal{A}_\varepsilon) \ge \frac{1}{2}$.*

Proof As $F(\mathbf{w})$ is non-negative, by Markov's Inequality, we have

$$\mu_\beta(\mathcal{A}_\varepsilon^c) = \mu_\beta(\{\mathbf{w} : F(\mathbf{w}) > \varepsilon\}) \le \frac{\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]}{\varepsilon}.$$

Now we compute $\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]$ with the same strategy as in the proof of Proposition 11 of Raginsky et al. [29]. Write

$$\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})] = \int_{\mathbb{R}^d} F(\mathbf{w})\mu_\beta(\mathbf{w})\mathrm{d}\mathbf{w} = \frac{1}{\beta}\big(h(\mu_\beta) - \log Z\big).$$

Here Z is the partition function of μ_β and

$$h(\mu_\beta) = -\int_{\mathbb{R}^d} \mu_\beta(\mathbf{w})\log\mu_\beta(\mathbf{w})\mathrm{d}\mathbf{w}$$

is the differential entropy of μ_β.

To upper bound the differential entropy of μ_β, we use the same derivation as the proof of Proposition 11 of Raginsky et al. [29]. The assumption that $\int_{\mathbb{R}^d}\|\mathbf{w}\|^2\mathrm{d}\mu_\beta(\mathbf{w}) \le S$, as well as the fact that the differential entropy of a measure with finite second moment is upper bounded by the differential entropy of a Gaussian with the same second moment, yields

$$h(\mu_\beta) \le \frac{d}{2}\log\left(\frac{2\pi eS}{d}\right).$$

Now we aim to lower bound the partition function Z. Using Lemma 25 and Lemma 26, we obtain

$$\log Z = \log\int_{\mathbb{R}^d} e^{-\beta F(\mathbf{w})}\mathrm{d}\mathbf{w}$$

$$\begin{aligned}
&\geq \log \int_{\mathbb{R}^d} e^{-\beta L\|\mathbf{w}-\mathbf{w}^\star\|^{s+1}} \mathrm{d}\mathbf{w} \\
&= \log \int_{\mathbb{R}^d} e^{-\beta L\|\mathbf{w}\|^{s+1}} \mathrm{d}\mathbf{w} \\
&= \log\left(\frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{s+1} \cdot (\beta L)^{-\frac{d}{s+1}} \cdot \Gamma\left(\frac{d}{s+1}\right)\right).
\end{aligned}$$

It is well known that on $\mathbb{R}_{>0}$, $\Gamma(\cdot)$ attains a constant lower bound of at least $\frac{1}{2}$ (the real value is around 0.8856, but this is all we need for our purposes). Moreover, by well-known properties of $\Gamma(\cdot)$, we have $\Gamma(d/2) = \frac{d}{2} \cdot \frac{d-2}{2} \cdot \ldots \cdot \frac{d-2\lfloor d/2\rfloor+r'+2}{2} \cdot \Gamma\left(\frac{d-2\lfloor d/2\rfloor+r'}{2}\right)$, where $r' = 2(1-d \pmod 2)$. Since $\frac{d-2\lfloor d/2\rfloor+r'}{2} \in \{1/2, 1\}$ and $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) \leq 1$, this gives $\Gamma\left(\frac{d-2\lfloor d/2\rfloor}{2}\right) \leq d^{d/2}\sqrt{\pi}$. This implies (since $\beta L \geq 1$) the following very loose bound:

$$\begin{aligned}
\log Z &\geq \log\left(\frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{s+1} \cdot (\beta L)^{-\frac{d}{s+1}} \cdot \Gamma\left(\frac{d}{s+1}\right)\right) \\
&\geq \log\left(\frac{\pi^{d/2}}{2\sqrt{\pi}(\beta L)^d d^{d/2}}\right) \\
&\geq -d\log(2\beta L d).
\end{aligned}$$

Hence, we see

$$\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})] = \frac{1}{\beta}\big(h(\mu_\beta) - \log Z\big) \leq \frac{d}{\beta}\left(\frac{1}{2}\log\left(\frac{2\pi e S}{d}\right) + \log(2\beta L d)\right) \leq \frac{d}{\beta}\log(4\pi e \beta L d S).$$

The conclusion follows from our condition on β and the original application of Markov's Inequality.

Note it suffices to just take $\varepsilon \geq 2\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]$ to make this proof work; most of our work was to find a suitable upper bound for $\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]$. Also, $\varepsilon = \Omega\big(\mathbb{E}_{\mathbf{w}\sim\mu_\beta}[F(\mathbf{w})]\big)$ is necessary, as demonstrated by the Gaussian example in Subsection B.2. ■

Appendix D. Proofs for Section 2

In this section, we state all guarantees with constant probability. To obtain those results with probability $1-\delta$, one can simply use the standard log-boosting trick.

D.1. Proofs of Theorem 6, 7, and 8

Here we formally state and prove Theorem 6, 7, and 8, which are all subsumed by the following result.

Theorem 15 *Suppose that F satisfies Assumption 2 and Assumption 3. Suppose μ_β has second moment $S < \infty$ and satisfies a Poincaré Inequality with constant $\mathsf{C}_{\mathrm{PI}}(\mu_\beta)$ with $\beta = \widetilde{\Theta}\big(\frac{d}{\varepsilon}\big)$, namely*

$$\varepsilon \geq \frac{2d}{\beta}\log(4\pi e\beta L d S).$$

Suppose Φ (from Theorem 3) satisfies Assumption 1 (hence $0 \leq p \leq 1$). Define $\rho_\Phi = \max(\rho_{\Phi,1}, \rho_{\Phi,2}, \rho_{\Phi,3})$. We can assume without loss of generality that $\rho_\Phi(z) = A(z+1)^p$ for some constant $A > 0$.

Then consider running either GLD*, or* SGLD *using a stochastic gradient oracle ∇f satisfying Assumption 4 and Assumption 5, with constant step size for T iterations. We will reach a $\mathbf{w}$ in $\{\mathbf{w} : F(\mathbf{w}) \le \varepsilon\}$ with probability at least 0.8 in at most T gradient (for* GLD*) or stochastic gradient (for* SGLD*) evaluations respectively, where we set*

$$T \le 8^3 C_0 \max\left(1, \frac{4L^2}{m}, \frac{4\max(L,B)}{m}, \frac{4B^2}{m}, 6B, 120^2 A^2 B^2 M^2, 120 A C_0 M\right) \cdot \max\left(\beta \max\left(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right), d^3 \max\left\{\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right\}^3, \beta^{2+s/2} \max\left(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right)^{2+s/2}\right).$$

(An explicit expression can be found in our proof.)

Here we define the above constants as follows:

$$L_2 := \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4\left(m + b + \frac{4d+2}{\beta}\right)}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right)^{s/2}, L_3 := \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4\left(m + b + \frac{4d+2}{\beta}\right)}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right)^{3s/4},$$

$$B = \max(L \max(1, \|\mathbf{w}^\star\|), \sigma_F), C_0 = 50 A \theta(\Phi(\mathbf{w}_0)) \vee 1, C = \frac{4A^2 p(p+1) + 2Ap + 1}{3},$$

$$M = \max\left(\frac{1}{2}, 2C\right) \cdot \left(8\sigma_F^3 + 16 \max(L,B)^3 (\max(L_2, L_3) + 1)\right).$$

Here $\theta = \frac{1}{\rho_\Phi}$, as defined in Lemma 16. (Take $L \leftarrow \max(1, L)$, $\sigma_F \leftarrow \max(\sigma_F, 1)$ if necessary.)

Moreover, this generalizes to $s = 0, 1$ as follows:

1. *In the case when $s = 0$, this result holds with no dependence on L_2, L_3 and instead we have*

$$M = \max\left(\frac{1}{2}, 2C\right) \cdot \left(8\sigma_F^3 + 16 \max(L,B)^3\right).$$

2. *In the case when $s = 1$, we no longer need to make assumptions on μ_β: as $2s \le \gamma \le s + 1$, $s = 1$ forces $\gamma = 1$, the setting of F being L-smooth and (m, b) dissipative from Raginsky et al. [29], Xu et al. [35], and Zou et al. [39]. As shown in Raginsky et al. [29], these conditions imply μ_β satisfies a Poincaré Inequality for $\beta \ge \frac{2}{m}$, and also that μ_β has finite second moment $S \le \frac{b + d/\beta}{m}$.*

   Moreover, our guarantees improve as follows. Instead letting

$$L_2 := \|\mathbf{w}_0\|^2 + \frac{2}{m}\left(b + 2B^2 + \frac{d}{\beta}\right), L_3 := \left(\|\mathbf{w}_0\|^4 + C'' \vee \frac{2C''}{m}\right)^{3/4},$$

   where

$$C'' = \frac{4}{m}\left(2C'^2\left(4 + \frac{1}{m}\right) \vee \frac{1}{m}\left(3mB + C'\right)^2\right), C' = m + b + \frac{4d+2}{\beta},$$

   we have a runtime guarantee of

$$T \le 8^3 C_0 \max\left(1, \frac{4L^2}{m}, \frac{4\max(L,B)}{m}, \frac{4B^2}{m}, 6B, 120^2 A^2 B^2 M^2, 120 A C_0 M\right) \cdot \max\left(\beta \max\left(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right), d^3 \max\left\{\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right\}^3, \beta^2 \max\left(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right)^2\right).$$

First, note from our assumption that $\varepsilon \geq \frac{2d}{\beta}\log(4\pi e\beta LdS)$, we may apply Theorem 3 (in particular, by combining Theorem 10, Lemma 14) to obtain Φ satisfying the properties described in Theorem 3.

Now, we need to show that with self-bounding regularity, by composing with the appropriate function, we can obtain some analogue of third-order smoothness in order to perform optimization-style discretization. We detail this as follows via the following Lemmas.

Lemma 16 *Let Φ be any non-negative function that satisfies polynomial self-bounding regularity to first, second, and third orders*[7]*, that is we have $\left\|\nabla^i\Phi(\mathbf{w})\right\|_{\mathrm{op}} \leq \rho_{\Phi,i}(\Phi(\mathbf{w}))$ for $1 \leq i \leq 3$, where $\rho_{\Phi,i}(z) = \sum_{j=1}^{n_i} c_{i,j} z^{d_{i,j}}$ for all $z \geq 0$ (where all the $d_{i,j} \geq 0$). Then there exists some $\theta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\theta'(z) > 0$, $\theta''(z) < 0$, $\theta'''(z) \geq 0$ for all $z \geq 0$, and*

$$\theta(\Phi(\mathbf{w}+\mathbf{u})) \leq \theta(\Phi(\mathbf{w})) + \theta'(\Phi(\mathbf{w}))\langle\nabla\Phi(\mathbf{w}),\mathbf{u}\rangle + \frac{1}{2}\theta'(\Phi(\mathbf{w}))\left\langle\nabla^2\Phi(\mathbf{w})\mathbf{u},\mathbf{u}\right\rangle + \frac{C}{6}\|\mathbf{u}\|^3,$$

for some constant C that depends only on the form of the functions ρ_1, ρ_2, and ρ_3.

Moreover, we also have

$$\theta(\Phi(\mathbf{w}+\mathbf{u})) \leq \theta(\Phi(\mathbf{w})) + \theta'(\Phi(\mathbf{w}))\langle\nabla\Phi(\mathbf{w}),\mathbf{u}\rangle + \frac{1}{2}\|\mathbf{u}\|^2,$$

and

$$\|\nabla\Phi(\mathbf{w})\| \leq \rho_\Phi(\Phi(\mathbf{w}))\sqrt{2\theta(\Phi(\mathbf{w}))}.$$

Proof Note we can assume without loss of generality that all the $c_{i,j} \geq 0$, and thus again we can assume without loss of generality that for all $z \geq 0$ we have

$$\max\left(\rho_{\Phi,1}(z), \rho_{\Phi,1}(z)^3, \rho_{\Phi,2}(z), \rho_{\Phi,3}(z), \rho_{\Phi,1}(z)\rho_{\Phi,2}(z)\right) \leq A + Az^p \leq 2A(z+1)^p$$

for some $A \geq 0, p \geq 0$. The last step follows from Lemma 27.

Next, define $\rho_\Phi(z) := 2A(z+1)^p$, which is clearly non-negative and increasing. Thus for all $z \geq 0$ we have

$$\rho_\Phi(z) \geq \max\left(\rho_{\Phi,1}(z), \rho_{\Phi,1}(z)^3, \rho_{\Phi,2}(z), \rho_{\Phi,3}(z), \rho_{\Phi,1}(z)\rho_{\Phi,2}(z)\right).$$

Now let $\theta(z)$ be defined by $\theta'(z) = \frac{1}{\rho_\Phi(z)}$ and $\theta(0) = 0$. The potential Φ we consider is non-negative and so we only consider $z \geq 0$; thus, θ is differentiable to all orders. Clearly $\theta'(z) > 0$. We can also check that $\theta''(z) = -\frac{p}{2A}(z+1)^{-p-1} < 0$, thus

$$\left|\theta''(z)\right|\rho_\Phi(z) = \frac{p}{2A}(z+1)^{-p-1}\cdot 2A(z+1)^p \leq p(z+1)^{-1} \leq p.$$

for all $z \geq 0$. Finally, we can compute $\theta'''(z) = \frac{p(p+1)}{2A}(z+1)^{-p-2}$, thus

$$\left|\theta'''(z)\right|\rho_\Phi(z) = \theta'''(z)\rho_\Phi(z) = \frac{p(p+1)}{2A}(z+1)^{-p-2}\cdot 2A(z+1)^p = \frac{p(p+1)}{(z+1)^2} \leq p(p+1)$$

for all $z \geq 0$.

7. This implicitly assumes Φ is differentiable through third order.

Now define for all $0 \le \alpha \le 1$,

$$l(\alpha) := \theta(\Phi(\mathbf{w} + \alpha\mathbf{u})).$$

Recall Φ is non-negative, so all the inputs here to θ are non-negative. $l(\alpha)$ is differentiable to third order, since Φ is and θ is for non-negative inputs.

By standard calculation using the Chain Rule (this is also done in the proof of Lemma 11 of [15]),

$$l'(\alpha) = \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle.$$

We also have, from similar calculation (also done in the proof of Lemma 11 of [15]) and using that $\theta''(z) \le 0$ for all $z \ge 0$ which was established earlier,

$$\begin{aligned} l''(\alpha) &= \theta''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle^2 + \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\left\langle \nabla^2\Phi(\mathbf{w} + \alpha\mathbf{u})\mathbf{u}, \mathbf{u}\right\rangle \\ &\le \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\left\langle \nabla^2\Phi(\mathbf{w} + \alpha\mathbf{u})\mathbf{u}, \mathbf{u}\right\rangle. \end{aligned}$$

Similar calculation, noting $\theta'(z) \ge 0$ and the bounds we established earlier on $|\theta''(z)|\rho_\Phi(z)$ and $|\theta'''(z)|\rho_\Phi(z)$, gives

$$\begin{aligned} l'''(\alpha) &= \theta'''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle \cdot \langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle^2 \\ &\qquad + \theta''(\Phi(\mathbf{w} + \alpha\mathbf{u})) \cdot 2\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle\left\langle \nabla^2\Phi(\mathbf{w} + \alpha\mathbf{u})\mathbf{u}, \mathbf{u}\right\rangle \\ &\qquad + \theta''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle \cdot \left\langle \nabla^2\Phi(\mathbf{w} + \alpha\mathbf{u})\mathbf{u}, \mathbf{u}\right\rangle \\ &\qquad + \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\nabla^3\Phi(\mathbf{w} + \alpha\mathbf{u})[\mathbf{u}, \mathbf{u}, \mathbf{u}] \\ &= \theta'''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle^3 \\ &\qquad + 3\theta''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\left\langle \nabla^2\Phi(\mathbf{w} + \alpha\mathbf{u})\mathbf{u}, \mathbf{u}\right\rangle\langle \nabla\Phi(\mathbf{w} + \alpha\mathbf{u}), \mathbf{u}\rangle \\ &\qquad + \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\nabla^3\Phi(\mathbf{w} + \alpha\mathbf{u})[\mathbf{u}, \mathbf{u}, \mathbf{u}] \\ &\le \left|\theta'''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\right|\rho_{\Phi,1}(\Phi(\mathbf{w} + \alpha\mathbf{u}))^3\|\mathbf{u}\|^3 \\ &\qquad + 3\left|\theta''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\right|\rho_{\Phi,1}(\Phi(\mathbf{w} + \alpha\mathbf{u}))\rho_{\Phi,2}(\Phi(\mathbf{w} + \alpha\mathbf{u}))\|\mathbf{u}\|^3 \\ &\qquad + \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\rho_{\Phi,3}(\Phi(\mathbf{w} + \alpha\mathbf{u}))\|\mathbf{u}\|^3 \\ &\le \rho_\Phi(\Phi(\mathbf{w} + \alpha\mathbf{u}))\left(\left|\theta'''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\right| + 3\left|\theta''(\Phi(\mathbf{w} + \alpha\mathbf{u}))\right| + \theta'(\Phi(\mathbf{w} + \alpha\mathbf{u}))\right)\|\mathbf{u}\|^3 \\ &\le (p^2 + p + 3p + 1)\|\mathbf{u}\|^3. \end{aligned}$$

From here, we consider Taylor expansion of $l(1)$ around 0. By Taylor's formula for the remainder, we know for some $\alpha \in [0, 1]$ that

$$l(1) = l(0) + l'(0) + \frac{1}{2}l''(0) + \frac{1}{6}l'''(\alpha).$$

Plugging in the above inequalities, we get

$$\theta(\Phi(\mathbf{w} + \mathbf{u})) \le \theta(\Phi(\mathbf{w})) + \theta'(\Phi(\mathbf{w}))\langle \nabla\Phi(\mathbf{w}), \mathbf{u}\rangle + \frac{1}{2}\theta'(\Phi(\mathbf{w}))\left\langle \nabla^2\Phi(\mathbf{w})\mathbf{u}, \mathbf{u}\right\rangle + \frac{p^2 + 4p + 1}{6}\|\mathbf{u}\|^3.$$

The result follows since $C = p^2 + 4p + 1$ only depends on the form of the functions $\rho_{\Phi,1}, \rho_{\Phi,2}$, and $\rho_{\Phi,3}$.

The second part follows from noticing that ρ_Φ as defined here is an upper bound on $\rho_{\Phi,1}$ and $\rho_{\Phi,2}$, so the same derivation as in the proof of Lemma 11 of De Sa et al. [15] suffices.

Finally, if $\max_j(d_{i,j}) \le 1$ for all $1 \le i \le 3$ (i.e. the max degree of the self-bounding regularity functions is at most 1), we can be a bit tighter in how we define θ. Instead we can just say

$$\max(\rho_{\Phi,1}(z), \rho_{\Phi,1}(z), \rho_{\Phi,1}(z)) \le A + Az^p \le 2A(z+1)^p$$

where $0 \le p \le 1$, and we define $\rho_\Phi(z) = A(z+1)^p$. Defining θ by $\theta'(z) = \frac{1}{\rho_\Phi(z)}$, $\theta(0) = 0$ analogously as before, note we have for any $z \ge 0$ that

$$\theta'(z) > 0, \theta''(z) < 0, \theta'''(z) > 0,$$

$$\left|\theta'''(z)\right|\rho_{\Phi,1}(z)^3 = \frac{p(p+1)}{A}(z+1)^{-p-2} \cdot 8A^3(z+1)^{3p} = 8A^2p(p+1)(z+1)^{2p-2} \le 8A^2p(p+1),$$

$$\left|\theta''(z)\right|\rho_{\Phi,1}(z)\rho_{\Phi,2}(z) = \frac{p}{A}(z+1)^{-p-1} \cdot 4A^2(z+1)^{2p} = 4Ap(z+1)^{p-1} \le 4Ap,$$

$$\left|\theta'(z)\right|\rho_{\Phi,3}(z) = \frac{1}{A(z+1)^p} \cdot 2A(z+1)^p = 2.$$

The above three lines all use $p \le 1$ in the last inequality of those lines. Therefore, an analogous derivation as above gives

$$\theta(\Phi(\mathbf{w}+\mathbf{u})) \le \theta(\Phi(\mathbf{w})) + \theta'(\Phi(\mathbf{w}))\langle\nabla\Phi(\mathbf{w}), \mathbf{u}\rangle + \frac{1}{2}\theta'(\Phi(\mathbf{w}))\left\langle\nabla^2\Phi(\mathbf{w})\mathbf{u}, \mathbf{u}\right\rangle + \frac{4A^2p(p+1) + 2Ap + 1}{3}\|\mathbf{u}\|^3.$$

■

Lemma 17 *For one iteration of* GLD *starting at arbitrary* $\mathbf{w}_t$*,*

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\lambda\Phi(\mathbf{w}_t) + \frac{1}{2}\eta^2\|\nabla F(\mathbf{w}_t)\|^2 + \frac{2C}{3}\eta^3\|\nabla F(\mathbf{w}_t)\|^3 + 2C(\eta d/\beta)^{3/2},$$

where p *and* C *are defined from Lemma 16.*

Proof First, apply Lemma 16 with $\mathbf{w} = \mathbf{w}_t$ and $\mathbf{u} = -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t$ to obtain

$$\begin{aligned}\theta(\Phi(\mathbf{w}_{t+1})) &= \theta\Big(\Phi(\mathbf{w}_t - \eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t)\Big)\\ &\le \theta(\Phi(\mathbf{w}_t)) + \theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla\Phi(\mathbf{w}_t), -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\\ &+ \frac{1}{2}\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\Big(-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big), -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\\ &+ \frac{C}{6}\Big\|-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\|^3\end{aligned}$$

where C is defined in the proof of Lemma 16.

We take expectations of this inequality with respect to $\boldsymbol{\varepsilon}_t$. Let's consider what each term of the upper bound becomes when we take expectations.

- First order term: Since $\boldsymbol{\varepsilon}_t$ has mean as the 0 vector,

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla\Phi(\mathbf{w}_t), -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\Big] = -\eta\theta'(\Phi(\mathbf{w}_t))\langle\nabla\Phi(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\rangle.$$

- Second order term: Note

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)(-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t), -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\Big] \\
&= \eta^2\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\Big\rangle \\
&\qquad - 2\eta(2\eta/\beta)^{1/2}\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\nabla F(\mathbf{w}_t), \mathbb{E}_{\boldsymbol{\varepsilon}_t}[\boldsymbol{\varepsilon}_t]\Big\rangle \\
&\qquad + (2\eta/\beta)\theta'(\Phi(w_t))\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t\Big\rangle\Big].
\end{aligned}$$

In the above, the cross terms cancel because $\boldsymbol{\varepsilon}_t$ has mean of the 0 vector.

Now, consider $\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t\Big\rangle\Big]$. We perform similar analysis as in the derivation and application of Ito's Lemma. This is where we see how the Laplacian term here actually helps. To make the parallels and motivation to Stochastic Calculus clear, here η corresponds to $\mathrm{d}t$, and $\sqrt{\eta}(\boldsymbol{\varepsilon}_t)_i$ corresponds to $(\mathrm{d}\mathbf{B}_t)_i$. Note

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t\Big\rangle\Big] &= \sum_{1\le i,j\le d}\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[(\boldsymbol{\varepsilon}_t)_i(\boldsymbol{\varepsilon}_t)_j\big(\nabla^2\Phi(\mathbf{w}_t)\big)_{ij}\Big] \\
&= \sum_{1\le i,j\le d}\nabla^2\Phi(\mathbf{w}_t)_{ij}\mathbb{E}_{\boldsymbol{\varepsilon}_t}[(\boldsymbol{\varepsilon}_t)_i(\boldsymbol{\varepsilon}_t)_j].
\end{aligned}$$

We break into cases:

1. When $i \neq j$: Note by symmetry of the unit sphere that

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}[(\boldsymbol{\varepsilon}_t)_i(\boldsymbol{\varepsilon}_t)_j] = 0.$$

   In particular this follows because for any $\mathbf{x} \in \mathcal{S}^{d-1}$, $(\boldsymbol{\varepsilon}_t)_j$ has equal probability of being $\mathbf{x}$ or $-\mathbf{x}$.

2. When $i = j$: This is where we pick up the Laplacian. Note by symmetry,

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}\big[(\boldsymbol{\varepsilon}_t)_i^2\big] = \mathbb{E}_{\boldsymbol{\varepsilon}_t}\big[(\boldsymbol{\varepsilon}_t)_j^2\big] \text{ for all } i, j, \text{ and } d = \mathbb{E}_{\boldsymbol{\varepsilon}_t}[\sum_{i=1}^{d}(\boldsymbol{\varepsilon}_t)_i^2] = \sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\varepsilon}_t}\big[(\boldsymbol{\varepsilon}_t)_i^2\big].$$

   Therefore,

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}\big[(\boldsymbol{\varepsilon}_t)_i^2\big] = 1 \text{ for all } 1 \le i \le d.$$

Hence, we obtain the Laplacian $\Delta\Phi(\mathbf{w}_t)$: we have plugging this into the above that

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t\Big\rangle\Big] = \sum_{i=1}^{d}(\nabla^2\Phi(\mathbf{w}_t))_{ii} = \Delta\Phi(\mathbf{w}_t).$$

Hence,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}_t}\Big[\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)(-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t), -\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\Big] \\
&= \eta^2\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\Big\rangle + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t).
\end{aligned}$$

- Third order term: By AM-GM, we can prove for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,

$$\|\mathbf{a}+\mathbf{b}\|^3 \le (\|\mathbf{a}\| + \|\mathbf{b}\|)^3 \le 4\|\mathbf{a}\|^3 + 4\|\mathbf{b}\|^3.$$

  Thus using this inequality pointwise we obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t}\left[\left\|-\eta\nabla F(\mathbf{w}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\right\|^3\right] \le 4\eta^3\|\nabla F(\mathbf{w}_t)\|^3 + 4(2\eta/\beta)^{3/2}d^{3/2}.$$

  The last step is because deterministically $\|\boldsymbol{\varepsilon}_t\| \le \sqrt{d}$ always.

Using the geometric property (3) and $\theta'(\Phi(\mathbf{w}_t)) \ge 0$ from Lemma 16,

$$-\eta\theta'(\Phi(\mathbf{w}_t))\langle\nabla\Phi(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\rangle \le -\eta\theta'(\Phi(\mathbf{w}_t))\left(\lambda\Phi(\mathbf{w}_t) + \frac{1}{\beta}\Delta\Phi(\mathbf{w}_t)\right).$$

Putting these together, this gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \\
&\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\langle\nabla\Phi(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\rangle \\
&\qquad + \frac{1}{2}\left(\eta^2\theta'(\Phi(\mathbf{w}_t))\left\langle\nabla^2\Phi(\mathbf{w}_t)\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\right\rangle + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t)\right) \\
&\qquad + \frac{C}{6}(4\eta^3\|\nabla F(\mathbf{w}_t)\|^3 + 4(2\eta/\beta)^{3/2}d^{3/2}) \\
&\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\left(\lambda\Phi(\mathbf{w}_t) + \frac{1}{\beta}\Delta\Phi(\mathbf{w}_t)\right) \\
&\qquad + \frac{1}{2}\left(\eta^2\theta'(\Phi(\mathbf{w}_t))\left\langle\nabla^2\Phi(\mathbf{w}_t)\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\right\rangle + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t)\right) \\
&\qquad + \frac{C}{6}\left(4\eta^3\|\nabla F(\mathbf{w}_t)\|^3 + 4(2\eta/\beta)^{3/2}d^{3/2}\right).
\end{aligned}$$

Note the terms $\eta\theta'(\Phi(\mathbf{w}_t)) \cdot \frac{1}{\beta}\Delta\Phi(\mathbf{w}_t)$ and $\frac{1}{2}(2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t)$ cancel out. Moreover, note by definition of operator norm and since we set $\theta'(z) = \frac{1}{\rho_\Phi(z)} \le \frac{1}{\rho_{\Phi,2}(z)}$, we obtain

$$\begin{aligned}
\frac{1}{2}\eta^2\theta'(\Phi(\mathbf{w}_t))\left\langle\nabla^2\Phi(\mathbf{w}_t)\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\right\rangle &\le \frac{1}{2}\eta^2\theta'(\Phi(\mathbf{w}_t))\|\nabla F(\mathbf{w}_t)\|^2\rho_2(\Phi(\mathbf{w}_t)) \\
&\le \frac{1}{2}\eta^2\|\nabla F(\mathbf{w}_t)\|^2.
\end{aligned}$$

Thus our above bound becomes

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \le \theta(\Phi(\mathbf{w}_t)) &- \eta\theta'(\Phi(\mathbf{w}_t)) \cdot \lambda\Phi(\mathbf{w}_t) \\
&+ \frac{1}{2}\eta^2\|\nabla F(\mathbf{w}_t)\|^2 + \frac{2C}{3}\eta^3\|\nabla F(\mathbf{w}_t)\|^3 + 2C(\eta d/\beta)^{3/2}.
\end{aligned}$$

This is the desired result. ■

In the SGLD setting, we also need the following to control the error of the gradient estimates to adapt to the stochastic gradient setting.

Lemma 18 *Suppose Assumption 4 holds. Letting $\{\mathbf{w}_t\}_{0\le t\le T-1}$ be the sequence of iterates generated by any of the variants of* SGLD *used in our algorithms on F, using stochastic gradient estimates based on $\{\mathbf{z}_t\}_{0\le t\le T-1}$. Then we have*

$$\mathbb{E}_{\mathbf{z}_t}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^2\big]\le\sigma_F^2,$$

and moreover

$$\mathbb{E}_{\mathbf{z}_t}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\big]\le 2\sigma_F^2+2\|\nabla F(\mathbf{w}_t)\|^2,\mathbb{E}_{\mathbf{z}_t}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\big]\le 8\sigma_F^3+4\|\nabla F(\mathbf{w}_t)\|^3.$$

Also with probability at least $1-\delta$ we have

$$\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|\le\|\nabla F(\mathbf{w}_t)\|+\sigma_F\sqrt{\log(T/\delta)}\ \textit{for all}\ 0\le t\le T-1.$$

Here, all probabilities and expectations are taken over the $\mathbf{z}_t$.

Proof Clearly $\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^2$ is non-negative, therefore

$$\begin{aligned}\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^2\big]&=\int_{t=0}^{\infty}\mathbb{P}\big(\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^2\ge t\big)\mathrm{d}t\\&=\int_{t=0}^{\infty}\mathbb{P}\big(\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|\ge\sqrt{t}\big)\mathrm{d}t\\&\le\int_{t=0}^{\infty}e^{-t/\sigma_F^2}\mathrm{d}t\\&=\sigma_F^2.\end{aligned}$$

Now by Young's Inequality we have pointwise

$$\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\le 2\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^2+2\|\nabla F(\mathbf{w}_t)\|^2,$$

and combining with the above gives

$$\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\big]\le 2\sigma_F^2+2\mathbb{E}\big[\|\nabla F(\mathbf{w}_t)\|^2\big].$$

Analogously, note

$$\begin{aligned}\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^3\big]&=\int_{t=0}^{\infty}\mathbb{P}\big(\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|^3\ge t\big)\mathrm{d}t\\&=\int_{t=0}^{\infty}\mathbb{P}\Big(\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|\ge t^{1/3}\Big)\mathrm{d}t\\&\le\int_{t=0}^{\infty}e^{-t^{2/3}/\sigma_F^2}\mathrm{d}t\\&\le 2\sigma_F^3.\end{aligned}$$

The inequality $\|\mathbf{a}+\mathbf{b}\|^3\le 4\|\mathbf{a}\|^3+4\|\mathbf{b}\|^3$ thus yields

$$\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\big]\le 8\sigma_F^3+4\mathbb{E}\big[\|\nabla F(\mathbf{w}_t)\|^3\big].$$

For a high probability statement, note for any $0\le t\le T-1$, we have $\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)-\nabla F(\mathbf{w}_t)\|\ge\sigma_F\sqrt{\log(T/\delta)}$ with probability at most δ/T. A Union Bound and Triangle Inequality implies that with probability at least $1-\delta$ we have

$$\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|\le\|\nabla F(\mathbf{w}_t)\|+\sigma_F\sqrt{\log(T/\delta)}\ \text{for all}\ 0\le t\le T-1.$$

■

Now analogously as before with Lemma 17, we prove a one-step discretization result. The main difference now is that we have to do the argument in a way that handles the stochasticity of the gradient estimates, but the same idea goes through.

Lemma 19 *For one iteration of* SGLD *starting at arbitrary* $\mathbf{w}_t$*,*

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\lambda\Phi(\mathbf{w}_t) + \frac{1}{2}\eta^2\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\right] + \frac{2C}{3}\eta^3\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\right] + 2C(\eta d/\beta)^{3/2}$$

where p *and* C *are defined from Lemma 16.*

Proof First, apply Lemma 16 with $\mathbf{w} = \mathbf{w}_t$ and $\mathbf{u} = -\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t$ to obtain

$$\begin{aligned}\theta(\Phi(\mathbf{w}_{t+1})) &= \theta\Big(\Phi(\mathbf{w}_t - \eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t)\Big)\\ &\le \theta(\Phi(\mathbf{w}_t)) + \theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla\Phi(\mathbf{w}_t), -\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\\ &+ \frac{1}{2}\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\Big(-\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big), -\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\\ &+ \frac{C}{6}\Big\|-\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\|^3\end{aligned}$$

where C is defined in the proof of Lemma 16.

We take expectations of this inequality with respect to $\boldsymbol{\varepsilon}_t$ and $\mathbf{z}_t$. Let's consider what each term of the upper bound becomes when we take expectations.

- First order term: Since $\nabla f(\mathbf{w}_t;\mathbf{z}_t)$ is unbiased, $\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\nabla f(\mathbf{w}_t;\mathbf{z}_t)] = \nabla F(\mathbf{w}_t)$. Thus as $\boldsymbol{\varepsilon}_t$ has mean of the 0 vector,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}\Big[\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla\Phi(\mathbf{w}_t), -\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\Big]\\ &= \theta'(\Phi(\mathbf{w}_t))\Big(\langle\nabla\Phi(\mathbf{w}_t), -\eta\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\nabla f(\mathbf{w}_t;\mathbf{z}_t)]\rangle + \Big\langle\nabla\Phi(\mathbf{w}_t), \sqrt{2\eta/\beta}\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\boldsymbol{\varepsilon}_t]\Big\rangle\Big)\\ &= -\eta\theta'(\Phi(\mathbf{w}_t))\langle\nabla\Phi(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\rangle.\end{aligned}$$

- Second order term: Note

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}\Big[\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\Big(-\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big), -\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\Big\rangle\Big]\\ &= \eta^2\theta'(\Phi(\mathbf{w}_t))\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\nabla f(\mathbf{w}_t;\mathbf{z}_t), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\Big\rangle\Big]\\ &\qquad - 2\eta(2\eta/\beta)^{1/2}\theta'(\Phi(\mathbf{w}_t))\Big\langle\nabla^2\Phi(\mathbf{w}_t)\mathbb{E}_{\boldsymbol{\varepsilon}_t}[\varepsilon_t], \mathbb{E}_{\mathbf{z}_t}[\nabla f(\mathbf{w}_t;\mathbf{z}_t)]\Big\rangle\\ &\qquad + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\varepsilon_t, \varepsilon_t\Big\rangle\Big]\\ &= \eta^2\theta'(\Phi(\mathbf{w}_t))\mathbb{E}_{\mathbf{z}_t}\Big[\Big\langle\nabla^2\Phi(\mathbf{w}_t)\nabla f(\mathbf{w}_t;\mathbf{z}_t), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\Big\rangle\Big] + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t).\end{aligned}$$

  Here we used that $\boldsymbol{\varepsilon}_t$ has zero mean as a vector and that $\boldsymbol{\varepsilon}_t, \mathbf{z}_t$ are clearly independent to compute the cross term. The calculation of

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}\Big[\Big\langle\nabla^2\Phi(w_t)\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t\Big\rangle\Big] = \Delta\Phi(\mathbf{w}_t)$$

  is the same as before. Note this expectation has no $\mathbf{z}_t$ dependence.

- Third order term: Again we use for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,

$$\|\mathbf{a}+\mathbf{b}\|^3 \le 4\|\mathbf{a}\|^3 + 4\|\mathbf{b}\|^3.$$

  Using this inequality pointwise we obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}\left[\left\|-\eta\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{2\eta/\beta}\boldsymbol{\varepsilon}_t\right\|^3\right] \le 4\eta^3\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\right] + 4(2\eta/\beta)^{3/2}d^{3/2}.$$

  The last step is because $\|\boldsymbol{\varepsilon}_t\| = \sqrt{d}$ always holds deterministically.

We put this together, noting $\theta'(\Phi(\mathbf{w}_t)) \ge 0$ from Lemma 16 which means we can use the admissability condition (3) which we use to upper bound the first order term. This gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \\
&\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\langle\nabla\Phi(\mathbf{w}_t), \nabla F(\mathbf{w}_t)\rangle \\
&+ \frac{1}{2}\Big(\eta^2\theta'(\Phi(\mathbf{w}_t))\mathbb{E}_{\mathbf{z}_t}\left[\left\langle\nabla^2\Phi(\mathbf{w}_t)\nabla f(\mathbf{w}_t;\mathbf{z}_t), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\right\rangle\right] + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t)\Big) \\
&\qquad + \frac{C}{6}\Big(4\eta^3\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\right] + 4(2\eta/\beta)^{3/2}d^{3/2}\Big) \\
&\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\Big(\lambda\Phi(\mathbf{w}_t) + \frac{1}{\beta}\Delta\Phi(\mathbf{w}_t)\Big) \\
&+ \frac{1}{2}\Big(\eta^2\theta'(\Phi(\mathbf{w}_t))\|\nabla^2\Phi(\mathbf{w}_t)\|_{\mathrm{op}}\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\right] + (2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t)\Big) \\
&\qquad + \frac{C}{6}\Big(4\eta^3\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\right] + 4(2\eta/\beta)^{3/2}d^{3/2}\Big).
\end{aligned}$$

The second inequality follows analogously as in the proof of Lemma 17; pointwise we have

$$\nabla f(\mathbf{w}_t;\mathbf{z}_t)^T\nabla^2\Phi(\mathbf{w}_t)\nabla f(\mathbf{w}_t;\mathbf{z}_t) \le \|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\|\nabla^2\Phi(\mathbf{w}_t)\|_{\mathrm{op}},$$

and the fact that

$$\theta'(z) \le \frac{1}{\rho_{\Phi,2}(z)}$$

always holds. Also note the terms $\eta\theta'(\Phi(\mathbf{w}_t))\cdot\frac{1}{\beta}\Delta\Phi(\mathbf{w}_t)$ and $\frac{1}{2}(2\eta/\beta)\theta'(\Phi(\mathbf{w}_t))\Delta\Phi(\mathbf{w}_t)$ cancel out. Thus our above bound becomes

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] &\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t))\lambda\Phi(\mathbf{w}_t) \\
&\quad + \frac{1}{2}\eta^2\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\right] + \frac{2C}{3}\eta^3\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\right] + 2C(\eta/\beta)^{3/2}d^{3/2}.
\end{aligned}$$

■

Next, we show a Lemma showing the iterates of GLD and SGLD are controlled. We will need this only when $s > 0$. Similar results have been shown in [29] and [4].

Lemma 20 *Suppose F satisfies Assumption 2 and Assumption 3. Consider the $\{\mathbf{w}_t\}_{t\ge0}$ generated by* GLD / SGLD *(for* SGLD *we need Assumption 5), run for T iterations for $T < \infty$ (we only use*

this for the T we set later). If the step size $\eta \in (0, 1 \wedge \frac{m}{4L^2} \wedge \frac{m}{4\max(L,B)} \wedge \frac{m}{4B^2} \wedge \frac{1}{6B})$, then we have the following bounds:

$$\mathbb{E}\left[\|\mathbf{w}_t\|^{2s}\right] \leq L_2 \max(\eta T, 1)^{s/2}, \mathbb{E}\left[\|\mathbf{w}_t\|^{3s}\right] \leq L_3 \max(\eta T, 1)^{3s/4},$$

where we define

$$L_2 := \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4\left(m + b + \frac{4d+2}{\beta}\right)}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right)^{s/2}, L_3 := \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4\left(m + b + \frac{4d+2}{\beta}\right)}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right)^{3s/4}.$$

Here $B = \max(L\max(1, \|\mathbf{w}^\star\|), \sigma_F)$, where σ_F comes from Assumption 4. (Recall we took $L \leftarrow \max(L, 1)$ if necessary earlier in the statement of Theorem 15.)

Moreover, if $s = 1$ (which implies F is L-smooth and (m, b) dissipative), we have the following uniform bounds:

$$\mathbb{E}\left[\|\mathbf{w}_t\|^2\right] \leq L_2, \mathbb{E}\left[\|\mathbf{w}_t\|^3\right] \leq L_3$$

for

$$L_2 := \|\mathbf{w}_0\|^2 + \frac{2}{m}\left(b + 2B^2 + \frac{d}{\beta}\right), L_3 := \left(\|\mathbf{w}_0\|^4 + C'' \vee \frac{2C''}{m}\right)^{3/4}$$

where

$$C'' = \frac{4}{m}\left(2C'^2\left(4 + \frac{1}{m}\right) \vee \frac{1}{m}\left(3mB + C'\right)^2\right), C' = m + b + \frac{4d+2}{\beta}.$$

Proof Our goal is to use Proposition 14 of Balasubramanian et al. [4] to control the second and fourth moments of the $\|\mathbf{w}_t\|$. Intuitively, our result should be the same as theirs except their V is replaced with βF, and then the relevant parameters change (except for d, the rest of them are all scaled by β). However, this gives some unnecessary β dependence which arises for technical reasons in their analysis (intuitively, they should cancel), so we need to modify their proof slightly to improve this dependence.

As done in the sampling literature [12], define the continuous-time interpolation of (1) by

$$\mathbf{w}_r = \mathbf{w}_t - (r - t\eta)\nabla F(\mathbf{w}_t) + \sqrt{\frac{2}{\beta}}(\mathbf{B}(r) - \mathbf{B}(t\eta)) \text{ for all } r \in [t\eta, (t+1)\eta).$$

This appears somewhat different than the interpolation defined in the literature, but it is actually the same. Our process (1) with step size η is equivalent to theirs with their $V = \beta F$ and their step size $h = \frac{\eta}{\beta}$. They index by 'time' where the subscript th corresponds to the t-th iterate whereas we index iterates simply by the iteration count (which is at 'time' $t\eta$ in the above interpolation) and are indexing time by r to avoid confusion.[8]

For the stochastic gradient case, this will be instead

$$\mathbf{w}_r = \mathbf{w}_t - (r - t\eta)\nabla f(\mathbf{w}_t; \mathbf{z}_t) + \sqrt{\frac{2}{\beta}}(\mathbf{B}(r) - \mathbf{B}(t\eta)) \text{ for all } r \in [t\eta, (t+1)\eta).$$

8. Using this correspondence one can actually carefully track the proof of Proposition 14 of Balasubramanian et al. [4] to show a similar result to what we show here.

Note for both these interpolations, all functions of quantities at time $t\eta$/iteration count t are constant (including $\mathbf{z}_t$), the randomness being over the Brownian motion $\mathbf{B}(r) - \mathbf{B}(t\eta)$.

We will do the proofs in the stochastic gradient case, and the proofs in the exact gradient case are the exact same.

First, we control the second moment. Define $\mathfrak{F}_t$ by the natural filtration with respect to $\mathbf{w}_j, \boldsymbol{\varepsilon}_j, \mathbf{z}_j$ for all $j \le t$. Analogously to the proof of Proposition 14 of [4], Itô's Lemma applied to $\|\mathbf{w}\|^2$ conditioned on $\mathfrak{F}_t$ yields for all $r \in [t\eta, (t+1)\eta]$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] &= 2\mathbb{E}[\langle \mathbf{w}_r, -\nabla f(\mathbf{w}_t;\mathbf{z}_t)\rangle|\mathfrak{F}_t] + \frac{1}{2}\cdot\sqrt{\frac{2}{\beta}}^2\cdot 2\mathrm{tr}(\mathbb{I}_d)\\
&= -2\mathbb{E}\left[\left\langle \mathbf{w}_t - (r-t\eta)\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{\frac{2}{\beta}}(\mathbf{B}(r)-\mathbf{B}(t\eta)), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\right\rangle|\mathfrak{F}_t\right] + \frac{2d}{\beta}\\
&= -2\mathbb{E}[\langle \mathbf{w}_t - (r-t\eta)\nabla f(\mathbf{w}_t;\mathbf{z}_t), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\rangle|\mathfrak{F}_t] + \frac{2d}{\beta}\\
&\le 2b - 2m\|\mathbf{w}_t\|^\gamma + 2(r-t\eta)\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2 + \frac{2d}{\beta}\\
&\le 2b - 2m\|\mathbf{w}_t\|^\gamma + 4\eta\cdot L^2\max(1,\|\mathbf{w}^\star\|)^{2s}\big(\|\mathbf{w}_t\|^{2s}+1\big) + \frac{2d}{\beta}\\
&\le 4m + 2b + \frac{2d}{\beta}. \qquad (12)
\end{aligned}$$

In the above we use that $\mathbb{E}[\langle(\mathbf{B}(r)-\mathbf{B}(t\eta)), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\rangle|\mathfrak{F}_t] = 0$, Assumption 5, Lemma 24, $\gamma \ge 2s$, $\eta \le \frac{m}{2B^2}$, and $r - t\eta \le \eta$. Integrating this over $r \in [t\eta, (t+1)\eta]$ and iterating yields

$$\mathbb{E}\big[\|\mathbf{w}_t\|^2\big] \le \|\mathbf{w}_0\|^2 + \left(4m + 2b + \frac{2d}{\beta}\right)\cdot\eta t \le \left(\|\mathbf{w}_0\|^2 + 4m + 2b + \frac{2d}{\beta}\right)\max(\eta T, 1).$$

We now control the fourth moment with the same idea. Applying Itô's Lemma to $\|\mathbf{w}\|^4 = \big(\|\mathbf{w}\|^2\big)^2$, we obtain

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4|\mathfrak{F}_t\big]\\
&= -4\mathbb{E}\big[\|\mathbf{w}_r\|^2\langle \mathbf{w}_r, \nabla f(\mathbf{w}_t;\mathbf{z}_t)\rangle|\mathfrak{F}_t\big] + \frac{1}{2}\cdot\sqrt{\frac{2}{\beta}}^2\cdot(4d+2)\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big]\\
&= -4\mathbb{E}\left[\|\mathbf{w}_r\|^2\left\langle \mathbf{w}_t - (r-t\eta)\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{\frac{2}{\beta}}(\mathbf{B}(r)-\mathbf{B}(t\eta)), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\right\rangle|\mathfrak{F}_t\right] + \frac{4d+2}{\beta}\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big].
\end{aligned}$$

Let $\mathbf{x} = \frac{\mathbf{B}(r)-\mathbf{B}(t\eta)}{\sqrt{r-t\eta}}$ be a standard Gaussian vector. Using Gaussian Integration by Parts on $h(\mathbf{x}) = \left\|\mathbf{w}_t - (r-t\eta)\nabla f(\mathbf{w}_t;\mathbf{z}_t) + \sqrt{\frac{2}{\beta}}\cdot\sqrt{r-t\eta}\mathbf{x}\right\|^2 = \|\mathbf{w}_r\|^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}_r\|^2\left\langle\sqrt{\frac{2}{\beta}}(\mathbf{B}(r)-\mathbf{B}(t\eta)), \nabla f(\mathbf{w}_t;\mathbf{z}_t)\right\rangle|\mathfrak{F}_t\right]\\
&= \sqrt{\frac{2}{\beta}}\cdot\sqrt{r-t\eta}\cdot\langle\mathbb{E}[\mathbf{x}h(\mathbf{x})|\mathfrak{F}_t], \nabla f(\mathbf{w}_t;\mathbf{z}_t)\rangle
\end{aligned}$$

$$= \sqrt{\frac{2}{\beta}} \cdot \sqrt{r - t\eta} \cdot \langle \mathbb{E}[\nabla h(\mathbf{x}) | \mathfrak{F}_t], \nabla f(\mathbf{w}_t; \mathbf{z}_t) \rangle$$
$$= \frac{4}{\beta}(r - t\eta)\langle \mathbf{w}_t - (r - t\eta)\nabla f(\mathbf{w}_t; \mathbf{z}_t), \nabla f(\mathbf{w}_t; \mathbf{z}_t) \rangle.$$

The above follows since $\nabla h(\mathbf{x}) = \sqrt{\frac{2}{\beta}} \cdot \sqrt{r - t\eta} \cdot \Big(\mathbf{w}_t - (r - t\eta)\nabla f(\mathbf{w}_t; \mathbf{z}_t) + \sqrt{\frac{2}{\beta}} \cdot \sqrt{r - t\eta}\mathbf{x} \Big)$ and as $\mathbf{x}$ is independent of $\mathfrak{F}_t$ and has mean of the 0 vector.

Hence, we have for all $r \in [t\eta, (t+1)\eta]$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4 | \mathfrak{F}_t\big] &= 4\mathbb{E}\big[\|\mathbf{w}_r\|^2 | \mathfrak{F}_t\big]\Big(-\langle \mathbf{w}_t, \nabla f(\mathbf{w}_t; \mathbf{z}_t)\rangle + (r - t\eta)\|\nabla f(\mathbf{w}_t; \mathbf{z}_t)\|^2 + \frac{4d+2}{\beta} \Big) \\
&\quad - \frac{16}{\beta}(r - t\eta)\langle \mathbf{w}_t - (r - t\eta)\nabla f(\mathbf{w}_t; \mathbf{z}_t), \nabla f(\mathbf{w}_t; \mathbf{z}_t)\rangle \\
&\le 4\Big(\mathbb{E}\big[\|\mathbf{w}_r\|^2 | \mathfrak{F}_t\big] + \frac{4}{\beta}(r - t\eta) \Big)\Big(-\langle \mathbf{w}_t, \nabla f(\mathbf{w}_t; \mathbf{z}_t)\rangle + (r - t\eta)\|\nabla f(\mathbf{w}_t; \mathbf{z}_t)\|^2 + \frac{4d+2}{\beta} \Big) \\
&\le 4\Big(\mathbb{E}\big[\|\mathbf{w}_r\|^2 | \mathfrak{F}_t\big] + \frac{4}{\beta}(r - t\eta) \Big)\Big(-m\|\mathbf{w}_t\|^\gamma + b + 2\eta \cdot L^2 \max(1, \|\mathbf{w}^\star\|)^{2s}\big(\|\mathbf{w}_t\|^{2s} + 1\big) + \frac{4d+2}{\beta} \Big) \\
&\le 4\Big(\mathbb{E}\big[\|\mathbf{w}_r\|^2 | \mathfrak{F}_t\big] + \frac{4}{\beta}(r - t\eta) \Big)\Big(-\frac{m}{2}\|\mathbf{w}_t\|^\gamma + m + b + \frac{4d+2}{\beta} \Big). \qquad (13)
\end{aligned}$$

The above follows as $r \ge t\eta$ and so the first factor in the above is always non-negative, as well as $\eta \le \frac{m}{4B^2}$ and $\gamma \ge 2s$.

Define $C' := m + b + \frac{4d+2}{\beta}$ for convenience. If $\|\mathbf{w}_t\| \ge \left(\frac{2C'}{m}\right)^{1/\gamma}$, this means $\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4 | \mathfrak{F}_t\big] \le 0$. Otherwise if $\|\mathbf{w}_t\| \le \left(\frac{2C'}{m}\right)^{1/\gamma}$, using our upper bound on $\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^2 | \mathfrak{F}_t\big]$ gives

$$\mathbb{E}\big[\|\mathbf{w}_r\|^2 | \mathfrak{F}_t\big] + \frac{4}{\beta}(r - t\eta) \le \|\mathbf{w}_t\|^2 + \Big(4m + 2b + \frac{2d}{\beta}\Big)(r - t\eta) + \frac{4}{\beta}(r - t\eta) \le \left(\frac{2C'}{m}\right)^{1/\gamma} + 4C',$$

as $r - t\eta \le \eta \le 1$. Note $\|\mathbf{w}_t\| \le \left(\frac{2C'}{m}\right)^{1/\gamma}$ implies the second factor in (13) is non-negative, and the second factor is at most C' clearly. Thus, in this case we have

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4 | \mathfrak{F}_t\big] \le 4C'\left(\left(\frac{2C'}{m}\right)^{1/\gamma} + 4C'\right) \le 8\left(\frac{4C'}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}.$$

Hence the above is an upper bound on $\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4 | \mathfrak{F}_t\big]$ in all cases, and iterating this gives the desired fourth moment bound

$$\mathbb{E}\big[\|\mathbf{w}_t\|^4\big] \le \|\mathbf{w}_0\|^4 + 8\left(\frac{4C'}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2} \eta t \le \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4C'}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right) \max(\eta T, 1).$$

From here, to obtain the desired conclusion, use monotonicity of moments (as $s \le 1$):

$$\mathbb{E}\big[\|\mathbf{w}_t\|^{2s}\big] \le \mathbb{E}\big[\|\mathbf{w}_t\|^4\big]^{2s/4} \le \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4C'}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right)^{s/2} \max(\eta T, 1)^{s/2}.$$

$$\mathbb{E}\big[\|\mathbf{w}_t\|^{3s}\big] \le \mathbb{E}\big[\|\mathbf{w}_t\|^4\big]^{3s/4} \le \left(\|\mathbf{w}_0\|^4 + 8\left(\frac{4C'}{m \wedge 1}\right)^{\frac{1+\gamma}{\gamma} \vee 2}\right)^{3s/4} \max(\eta T, 1)^{3s/4}.$$

When $s = 1$ and hence $\gamma = 2$ (which implies (m, b) dissipativeness), we can be tighter in the above analysis. First, using Lemma 18, we have

$$\mathbb{E}_{\mathbf{z}_t}\big[\|\nabla f(\mathbf{w}_t; \mathbf{z}_t) - \nabla F(\mathbf{w}_t)\|^2\big] \le \sigma_F^2 \le B^2.$$

With the above, identically as the steps of the proof of Lemma 3 of Raginsky et al. [29], using (m, b) dissipativeness and our constant upper bound on η, we can show a uniform bound on the second moment for both exact and stochastic gradients:

$$\mathbb{E}\big[\|\mathbf{w}_t\|^2\big] \le \|\mathbf{w}_0\|^2 + \frac{2}{m}\left(b + 2B^2 + \frac{d}{\beta}\right).$$

We also claim we have a uniform upper bound on the fourth moment. We break into two cases, both using a similar strategy.

1. $\|\mathbf{w}_t\| \le \left(\frac{2C'}{m}\right)^{1/2}$: In this case, the second factor in (13) is non-negative. Recall the upper bound we showed from (12):

   $$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] \le 4m + 2b + \frac{2d}{\beta}.$$

   This implies

   $$\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] \le \|\mathbf{w}_t\|^2 + \left(4m + 2b + \frac{2d}{\beta}\right)(r - t\eta).$$

   Now as the second factor in (13) is non-negative, we obtain using $r - t\eta \le \eta \le 1$,

   $$\begin{aligned}
   \frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4|\mathfrak{F}_t\big] &\le 4\left(\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] + \frac{4}{\beta}(r - t\eta)\right)\left(-\frac{m}{2}\|\mathbf{w}_t\|^2 + m + b + \frac{4d+2}{\beta}\right) \\
   &\le 4\left(\|\mathbf{w}_t\|^2 + \left(4m + 2b + \frac{2d}{\beta}\right)(r - t\eta) + \frac{4}{\beta}(r - t\eta)\right)\left(-\frac{m}{2}\|\mathbf{w}_t\|^2 + m + b + \frac{4d+2}{\beta}\right) \\
   &\le 4\big(\|\mathbf{w}_t\|^2 + 4C'\big)\left(-\frac{m}{2}\|\mathbf{w}_t\|^2 + C'\right) \\
   &\le 4\left(-\frac{m}{2}\|\mathbf{w}_t\|^4 + C'\|\mathbf{w}_t\|^2 + 4C'^2\right) \\
   &\le 4\left(-\frac{m}{4}\|\mathbf{w}_t\|^4 + C'^2\left(4 + \frac{1}{m}\right)\right) = -m\|\mathbf{w}_t\|^4 + 4C'^2\left(4 + \frac{1}{m}\right).
   \end{aligned}$$

   The last step uses AM-GM.

2. $\|\mathbf{w}_t\| > \left(\frac{2C'}{m}\right)^{1/2}$: This time, the second factor in (13) is negative, so we aim to lower bound $\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big]$. Recalling the intermediate steps in (12), we have

   $$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] = -2\mathbb{E}[\langle \mathbf{w}_t - (r - t\eta)\nabla f(\mathbf{w}_t; \mathbf{z}_t), \nabla f(\mathbf{w}_t; \mathbf{z}_t)\rangle|\mathfrak{F}_t] + \frac{2d}{\beta}$$

$$\begin{aligned}
&\geq -2\langle \mathbf{w}_t, \nabla f(\mathbf{w}_t; \mathbf{z}_t)\rangle \\
&\geq -2\|\mathbf{w}_t\|\|\nabla f(\mathbf{w}_t; \mathbf{z}_t)\| \\
&\geq -2B\|\mathbf{w}_t\|(\|\mathbf{w}_t\| + 1) \\
&\geq -3B\big(\|\mathbf{w}_t\|^2 + 1\big),
\end{aligned}$$

where we upper bound $\nabla f(\mathbf{w}_t; \mathbf{z}_t)$ via Lemma 24 and use AM-GM in the last step.

This implies

$$\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] \geq \|\mathbf{w}_t\|^2 - 3B\big(\|\mathbf{w}_t\|^2 + 1\big)(r - t\eta) \geq \frac{1}{2}\|\mathbf{w}_t\|^2 - 3B,$$

since $r - t\eta \leq \eta$, $\eta \leq \frac{1}{6B} \leq 1$.

The second factor in (13) is negative, so we may apply this in (13) to give

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4|\mathfrak{F}_t\big] &\leq 4\bigg(\mathbb{E}\big[\|\mathbf{w}_r\|^2|\mathfrak{F}_t\big] + \frac{4}{\beta}(r - t\eta)\bigg)\bigg(-\frac{m}{2}\|\mathbf{w}_t\|^2 + m + b + \frac{4d+2}{\beta}\bigg) \\
&\leq 4\bigg(\frac{1}{2}\|\mathbf{w}_t\|^2 - 3B\bigg)\bigg(-\frac{m}{2}\|\mathbf{w}_t\|^2 + C'\bigg) \\
&= 4\bigg(-\frac{m}{4}\|\mathbf{w}_t\|^4 + \bigg(\frac{3mB}{2} + \frac{C'}{2}\bigg)\|\mathbf{w}_t\|^2 - 3BC'\bigg) \\
&\leq 4\bigg(-\frac{m}{8}\|\mathbf{w}_t\|^4 + \frac{(3mB + C')^2}{2m}\bigg) = -\frac{m}{2}\|\mathbf{w}_t\|^4 + \frac{2}{m}\cdot(3mB + C')^2.
\end{aligned}$$

Again, the last step uses AM-GM.

From the above we see that in either case we have

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\big[\|\mathbf{w}_r\|^4|\mathfrak{F}_t\big] \leq -\frac{m}{2}\|\mathbf{w}_t\|^4 + C''$$

where $C'' = 4C'^2\big(4 + \frac{1}{m}\big) \vee \frac{2}{m}\cdot(3mB + C')^2$.

Iterating the above for one step and then taking full expectation yields the recursion

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}\|^4\big] \leq \mathbb{E}\big[\|\mathbf{w}_t\|^4\big] + \eta\bigg(-\frac{m}{2}\mathbb{E}\big[\|\mathbf{w}_t\|^4\big] + C''\bigg) = \bigg(1 - \frac{\eta m}{2}\bigg)\mathbb{E}\big[\|\mathbf{w}_t\|^4\big] + \eta C''.$$

If $1 - \frac{\eta m}{2} \leq 0$ we obtain $\mathbb{E}\big[\|\mathbf{w}_t\|^4\big] \leq C''$, and otherwise if $1 - \frac{\eta m}{2} \in (0, 1)$, iterating the above and summing the resulting geometric series gives

$$\mathbb{E}\big[\|\mathbf{w}_t\|^4\big] \leq \bigg(1 - \frac{\eta m}{2}\bigg)^t\|\mathbf{w}_0\|^4 + \frac{2\eta C''}{\eta m} = \|\mathbf{w}_0\|^4 + \frac{2C''}{m}.$$

The desired upper bound on the third moment in this case now just comes from monotonicity of moments. ■

We now are ready to prove Theorem 15. We do the proof when $0 < s \leq 1$ (when $s > 0$, $\gamma \geq 2s > 0$ so we can certainly use Lemma 20), and we discuss the simple extension to $s = 0$ and the tighter results when $s = 1$ at the end.

Proof Consider θ and $C = \frac{Ap^2+4Ap+1}{6}$ defined in terms of ρ_Φ in Lemma 16 for the $p \le 1$ case.

We set

$$C_0 = 50A\theta(\Phi(\mathbf{w}_0)) \vee 1, M = \max\left(\frac{1}{2}, 2C\right) \cdot \left(8\sigma_F^3 + 16\max(L,B)^3(\max(L_2,L_3)+1)\right),$$

$$\eta = \min\left(1, \frac{m}{4L^2}, \frac{m}{4\max(L,B)}, \frac{m}{4B^2}, \frac{1}{6B}, \frac{1}{120^2A^2B^2M^2} \cdot \frac{\beta^3\lambda^2}{d^3}, \frac{\lambda^{1+s/2}}{120AC_0M}\right),$$

$$T = \frac{C_0}{\eta\lambda}.$$

Here $\lambda \in \left[\frac{1}{8\beta}\min\left(\frac{1}{\mathsf{C}_{\mathrm{PI}}(\mu_\beta)}, \frac{1}{2}\right), \frac{1}{4\beta}\min\left(\frac{1}{\mathsf{C}_{\mathrm{PI}}(\mu_\beta)}, \frac{1}{2}\right)\right]$, as with Φ, comes from Theorem 3. Thus, using

$$T = C_0\max\left(\frac{1}{\lambda}\max\left\{1, \frac{4L^2}{m}, \frac{4\max(L,B)}{m}, \frac{4B^2}{m}, 6B\right\}, 120^2A^2B^2M^2 \cdot \frac{d^3}{\beta^2\lambda^3}, 120AC_0M\frac{1}{\lambda^{2+s/2}}\right)$$

and

$$\frac{1}{\lambda} \le 8\beta\max\left(\mathsf{C}_{\mathrm{PI}}(\mu_\beta), 2\right),$$

we see that our definition of T above is consistent with the statement of Theorem 15. Moreover, note $\eta T = \frac{C_0}{\lambda} \ge 1$.

As with before let $\mathfrak{F}_t$ be the natural filtration with respect to $\boldsymbol{\varepsilon}_{t'}, \mathbf{z}_{t'}$ for all $0 \le t' \le t$ in the SGLD case, and with respect to $\boldsymbol{\varepsilon}_{t'}$ for all $0 \le t' \le t$ in the GLD case.

Define

$$\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0) = \min(\tau_{\mathcal{A}_\varepsilon}(\mathbf{w}_0), T),$$

where in a slight abuse of notation, $\tau_{\mathcal{A}_\varepsilon}$ now denotes the hitting time of discrete-time GLD/SGLD to $\mathcal{A}_\varepsilon$ with the choice of η above. Note $\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)$ is a stopping time that is at most $T < \infty$.

Consider $\mathbf{w}_t$ for $t < \tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)$, thus $\mathbf{w}_t \in \mathcal{A}_\varepsilon^c$. By Theorem 3, this implies for this $\mathbf{w}_t$, (4) holds:

$$\langle \nabla F(\mathbf{w}), \nabla\Phi(\mathbf{w})\rangle \ge \lambda\Phi(\mathbf{w}) + \frac{1}{\beta}\Delta\Phi(\mathbf{w}).$$

Recall $\theta' > 0$ from Lemma 16, including in this case where $p \le 1$. By Lemma 17, which uses the geometric condition (3), we obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t)) \cdot \lambda\Phi(\mathbf{w}_t) + \frac{1}{2}\eta^2\|\nabla F(\mathbf{w}_t)\|^2 + C\eta^3\|\nabla F(\mathbf{w}_t)\|^3 + 2C\left(\frac{\eta d}{\beta}\right)^{3/2}.$$

This uses Lemma 16 in the $p \le 1$ case.

In the stochastic gradient case we have by Lemma 19 that

$$\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] \le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t)) \cdot \lambda\Phi(\mathbf{w}_t) + \frac{1}{2}\eta^2\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^2\right] + C\eta^3\mathbb{E}_{\mathbf{z}_t}\left[\|\nabla f(\mathbf{w}_t;\mathbf{z}_t)\|^3\right] + 2C\left(\frac{\eta d}{\beta}\right)^{3/2}.$$

(Note these results can be proved for either $\boldsymbol{\varepsilon}_t \sim \mathcal{S}^{d-1}$ or $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \mathbb{I}_d)$ by the exact same proof as Lemma 17, Lemma 19.)

Applying Lemma 18 and then Lemma 24, Young's Inequality, and $\|\mathbf{a}+\mathbf{b}\|^3 \le 4\|\mathbf{a}\|^3 + 4\|\mathbf{b}\|^3$, and noting $\sigma_F \ge 0$, we see in both the GLD and SGLD cases that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}_t, \mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] &\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t)) \cdot \lambda\Phi(\mathbf{w}_t) \\
&\quad + \frac{1}{2}\eta^2\big(2\sigma_F^2 + 2\|\nabla F(\mathbf{w}_t)\|^2\big) + C\eta^3\big(8\sigma_F^3 + 4\|\nabla F(\mathbf{w}_t\|^3\big) \\
&\quad + 2C\left(\frac{\eta d}{\beta}\right)^{3/2} \\
&\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t)) \cdot \lambda\Phi(\mathbf{w}_t) \\
&\quad + \frac{1}{2}\eta^2\big(2\sigma_F^2 + 4\max(L,B)^2\big(\|\mathbf{w}_t\|^{2s} + 1\big)\big) \\
&\quad + C\eta^3\big(8\sigma_F^3 + 16\max(L,B)^3\big(\|\mathbf{w}_t\|^{3s} + 1\big)\big) + 2C\left(\frac{\eta d}{\beta}\right)^{3/2}.
\end{aligned}$$

Recall that $\theta'(z) = \frac{1}{A(z+1)^p}$ where $p \le 1$, which is increasing on $z \ge 0$. Therefore, $z\theta'(z) = \frac{z}{A(z+1)^p} \ge \frac{1}{2A}$ for $z \ge 1$. Recall $\Phi(\mathbf{w}_t) \ge 1$ from Remark 4, because $t < \tau_{\mathcal{A}_\varepsilon, T}(\mathbf{w}_0)$ and so $\mathbf{w}_t \in \mathcal{A}_\varepsilon^c$. Thus, $\Phi(\mathbf{w}_t)\theta'(\Phi(\mathbf{w}_t)) \ge \frac{1}{2A}$. Therefore we can rearrange the above as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}_t, \mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))] &\le \theta(\Phi(\mathbf{w}_t)) - \eta\theta'(\Phi(\mathbf{w}_t)) \cdot \lambda\Phi(\mathbf{w}_t) + \frac{1}{2}\eta^2\big(2\sigma_F^2 + 4\max(L,B)^2\big(\|\mathbf{w}_t\|^{2s} + 1\big)\big) \\
&\quad + C\eta^3\big(8\sigma_F^3 + 16\max(L,B)^3\big(\|\mathbf{w}_t\|^{3s} + 1\big)\big) + 2C\left(\frac{\eta d}{\beta}\right)^{3/2} \\
&\le \theta(\Phi(\mathbf{w}_t)) - \frac{\eta\lambda}{2A} + \frac{1}{2}\eta^2\big(2\sigma_F^2 + 4\max(L,B)^2\big(\|\mathbf{w}_t\|^{2s} + 1\big)\big) \\
&\quad + C\eta^3\big(8\sigma_F^3 + 16\max(L,B)^3\big(\|\mathbf{w}_t\|^{3s} + 1\big)\big) + 2C\left(\frac{\eta d}{\beta}\right)^{3/2} \\
&= \theta(\Phi(\mathbf{w}_t)) - \frac{\eta\lambda}{2A} + \mathrm{err}(\mathbf{w}_t), \qquad (14)
\end{aligned}$$

where we define

$$\mathrm{err}(\mathbf{w}) := \frac{1}{2}\eta^2\big(2\sigma_F^2 + 4\max(L,B)^2\big(\|\mathbf{w}\|^{2s} + 1\big)\big) + C\eta^3\big(8\sigma_F^3 + 16\max(L,B)^3\big(\|\mathbf{w}\|^{3s} + 1\big)\big) + 2C\left(\frac{\eta d}{\beta}\right)^{3/2} > 0.$$

Now with (14), the idea is to sum and telescope this relations over $\tau_{\mathcal{A}_\varepsilon, T+1}$ time steps, as discussed in Subsection 2.1. The way to do this is using discrete-time Dynkin's Formula, stated in Theorem 11.3.1 of [23]:

Theorem 21 (Theorem 11.3.1 of [23]) *Let Z_t be any $\mathfrak{F}_t$-measurable function of $\mathbf{w}_0, \dots, \mathbf{w}_t$. Consider any stopping time τ and define $\tau^n := \min\{n, \tau, \inf(t \ge 0 : \mathbf{z}_t \ge n)\}$. Then we have for all $n \ge 0$ and $\mathbf{w}_0 \in \mathbb{R}^d$ that*

$$\mathbb{E}[Z_{\tau^n}] = \mathbb{E}[Z_0] + \mathbb{E}\left[\sum_{t=1}^{\tau^n} (\mathbb{E}[Z_t | \mathfrak{F}_t] - Z_{t-1})\right].$$

As a simple corollary of Theorem 21, we have the following, Proposition 11.3.2 of [23]. Unlike the above, it holds for *any stopping time*.

Corollary 22 (Proposition 11.3.2 of [23]) *Suppose there exists non-negative functions* s_t, f_t[9] *such that*

$$\mathbb{E}[Z_{t+1}|\mathfrak{F}_t] \leq Z_t - f_t(\mathbf{w}_t) + s_t(\mathbf{w}_t). \tag{15}$$

Then for any $\mathbf{w}_0 \in \mathbb{R}^d$ *and any stopping time* τ*,*

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1} f_t(\mathbf{w}_t)\right] \leq Z_0 + \mathbb{E}\left[\sum_{t=0}^{\tau-1} s_t(\mathbf{w}_t)\right].$$

Apply Corollary 22 for the stopping time $\tau = \tau_{\mathcal{A}_\varepsilon,T+1}$, $Z_t = \theta(\Phi(\mathbf{w}_t))$, and the functions f_t, s_t defined as follows. Take

$$f_t(\mathbf{w}) = \begin{cases} \frac{\eta\lambda}{2A} & \text{if } \mathbf{w} \in \mathcal{A}_\varepsilon^c \\ 0 & \text{otherwise} \end{cases}.$$

In the GLD case take

$$s_t(\mathbf{w}) = \begin{cases} \mathrm{err}(\mathbf{w}) & \text{if } \mathbf{w} \in \mathcal{A}_\varepsilon^c \\ \mathbb{E}_{\boldsymbol{\varepsilon},\mathbf{z}}\left[\theta\left(\Phi\left(\mathbf{w} - \eta\nabla F(\mathbf{w}) + \sqrt{\frac{2\eta}{\beta}}\boldsymbol{\varepsilon}\right)\right)\right] & \text{otherwise} \end{cases},$$

and in the SGLD case take

$$s_t(\mathbf{w}) = \begin{cases} \mathrm{err}(\mathbf{w}) & \text{if } \mathbf{w} \in \mathcal{A}_\varepsilon^c \\ \mathbb{E}_{\boldsymbol{\varepsilon},\mathbf{z}}\left[\theta\left(\Phi\left(\mathbf{w} - \eta\nabla f(\mathbf{w};\mathbf{z}) + \sqrt{\frac{2\eta}{\beta}}\boldsymbol{\varepsilon}\right)\right)\right] & \text{otherwise} \end{cases}.$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbb{I}_d)$ and $\mathbf{z}$ is an arbitrary data sample. Note the $\{f_t\}$, as well as the $\{s_t\}$, are the same function for all t. Since $\theta \geq 0$, the f_t and s_t are non-negative. As $\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\cdot]$ is the same as $\mathbb{E}[\cdot|\mathfrak{F}_t]$, (14) proves that (15) holds if $\mathbf{w}_t \in \mathcal{A}_\varepsilon^c$, and (15) holds for $\mathbf{w}_t \in \mathcal{A}_\varepsilon$ as the $Z_t \geq 0$ and as the $s_t(\mathbf{w}_t) = \mathbb{E}[Z_{t+1}|\mathfrak{F}_t]$[10]. Thus, Corollary 22 yields

$$\mathbb{E}\left[\sum_{t=0}^{\tau_{\mathbf{A}_\varepsilon,T}(\mathbf{w}_0)-1} \frac{\eta\lambda}{2A}\right] = \mathbb{E}\left[\sum_{t=0}^{\tau_{\mathbf{A}_\varepsilon,T}(\mathbf{w}_0)-1} f_t(\mathbf{w}_t)\right] \leq Z_0 + \mathbb{E}\left[\sum_{t=0}^{\tau_{\mathbf{A}_\varepsilon,T}(\mathbf{w}_0)-1} s_t(\mathbf{w}_t)\right] = \theta(\Phi(\mathbf{w}_0)) + \mathbb{E}\left[\sum_{t=0}^{\tau_{\mathbf{A}_\varepsilon,T}(\mathbf{w}_0)-1} \mathrm{err}(\mathbf{w}_t)\right],$$

since $\mathbf{w}_t \in \mathcal{A}_\varepsilon^c$ for all $t \leq \tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0) - 1$, and using the definition of f_t, s_t in that case.

Clearly we can simplify the left hand side as $\frac{\eta\lambda}{2A}\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]$. For the right hand side, note pointwise we have $\sum_{t=0}^{\tau_{\mathbf{A}_\varepsilon,T}(\mathbf{w}_0)-1} \mathrm{err}(\mathbf{w}_t) \leq \sum_{t=0}^{T-1} \mathrm{err}(\mathbf{w}_t)$ by definition of $\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)$ and as the $\mathrm{err}(\mathbf{w}) \geq 0$. Moreover, all the relevant expectations are finite (by Lemma 20 and as $\tau_{\mathcal{A}_\varepsilon,T+1} \leq T < \infty$). Therefore we see

$$\frac{\eta\lambda}{2A}\mathbb{E}\left[\tau_{\mathbf{A}_\varepsilon,T}(\mathbf{w}_0)\right] \leq \theta(\Phi(\mathbf{w}_0)) + \mathbb{E}\left[\sum_{t=0}^{T-1} \mathrm{err}(\mathbf{w}_t)\right].$$

We now show that the random variable $\tau_{\mathcal{A}_\varepsilon,T}$ is well-controlled.

9. The result in [23] states this for positive s_t, f_t, but it is clear their proof still works when the functions are non-negative.
10. But this is not relevant, since we apply Corollary 22 with $\tau = \tau_{\mathcal{A}_\varepsilon,T+1}$.

Lemma 23 *We have*

$$\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)] < \frac{T}{10}.$$

Proof Suppose otherwise that $\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)] \geq \frac{T}{10} > 0$. Rearranging the above gives

$$\begin{aligned}\frac{\lambda}{2A} &\leq \frac{\theta(\Phi(\mathbf{w}_0))}{\eta\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]} + \frac{1}{\eta\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]}\mathbb{E}\left[\sum_{t=0}^{T-1}\mathrm{err}(\mathbf{w}_t)\right] \\ &= \frac{\theta(\Phi(\mathbf{w}_0))}{\eta\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]} + \frac{1}{\eta\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]}\sum_{t=0}^{T-1}\mathbb{E}[\mathrm{err}(\mathbf{w}_t)]. \qquad (16)\end{aligned}$$

By Lemma 20, which we may apply as our choice of η is small enough, we have

$$\mathbb{E}\big[\|\mathbf{w}_t\|^{2s}\big] \leq L_2\max(\eta T,1)^{s/2}, \mathbb{E}\big[\|\mathbf{w}_t\|^{3s}\big] \leq L_3\max(\eta T,1)^{3s/4}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\mathrm{err}(\mathbf{w}_t)] &\leq \frac{1}{2}\eta^2\Big(2\sigma_F^2 + 4\max(L,B)^2(L_2\max(\eta T,1)^{s/2}+1)\Big) \\ &\qquad + C\eta^3\Big(8\sigma_F^3 + 16\max(L,B)^3(L_3\max(\eta T,1)^{3s/4}+1)\Big) + 2C\left(\frac{\eta d}{\beta}\right)^{3/2} \\ &\leq M\Big((\eta d/\beta)^{3/2} + \eta^2\cdot(\eta T)^{s/2} + \eta^3\cdot(\eta T)^{3s/4}\Big).\end{aligned}$$

The last line follows as $\eta T \geq 1$ and from definition of M (recall we took $\sigma_F \leftarrow \max(\sigma_F,1)$ if necessary); recall

$$M = \max\left(\frac{1}{2}, 2C\right)\cdot\big(8\sigma_F^3 + 16\max(L,B)^3(\max(L_2,L_3)+1)\big).$$

Recall our choice of T such that $\eta T = \frac{C_0}{\lambda}$, and also our choice of $C_0 = 50A\theta(\Phi(\mathbf{w}_0))\vee 1$. Therefore, (16) becomes

$$\begin{aligned}\frac{\lambda}{2A} &\leq \frac{\theta(\Phi(\mathbf{w}_0))}{\eta\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]} + \frac{1}{\eta\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)]}\sum_{t=0}^{T-1}\mathbb{E}[\mathrm{err}(\mathbf{w}_t)] \\ &\leq \frac{10\theta(\Phi(\mathbf{w}_0))}{\eta T} + \frac{10}{\eta T}\cdot T\cdot M\Big((\eta d/\beta)^{3/2} + \eta^2\cdot(\eta T)^{s/2} + \eta^3\cdot(\eta T)^{3s/4}\Big) \\ &= 10\left(\frac{\theta(\Phi(\mathbf{w}_0))\lambda}{C_0} + M\left(\eta^{1/2}(d/\beta)^{3/2} + \eta\cdot\frac{C_0^{s/2}}{\lambda^{s/2}} + \eta^2\cdot\frac{C_0^{3s/4}}{\lambda^{3s/4}}\right)\right) \\ &< 10\left(\frac{\lambda}{40A} + MC_0\left(\eta^{1/2}(d/\beta)^{3/2} + \frac{\eta}{\lambda^{s/2}} + \frac{\eta^2}{\lambda^{3s/4}}\right)\right) \\ &< 10\cdot\frac{\lambda}{20A} = \frac{\lambda}{2A}.\end{aligned}$$

In the second inequality we use $\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)] \geq \frac{T}{10}$ which we are supposing for contradiction. The last ienquality uses

$$\eta \leq \min\left(\frac{1}{120^2A^2B^2M^2}\cdot\frac{\beta^3\lambda^2}{d^3}, \frac{\lambda^{1+s/2}}{120AC_0M}\right).$$

Note as we have $\lambda \le 1$ and $A \ge 1$, $C_0 \ge 1$, $M \ge \frac{1}{2}$, this implies

$$\frac{\lambda^{1+s/2}}{120AC_0M} \le \frac{\lambda^{\frac{1}{2}+\frac{3s}{8}}}{(120AC_0M)^{1/2}},$$

which we also use to show $MC_0 \cdot \frac{\eta^2}{\lambda^{3s/4}} \le \frac{\lambda}{120A}$. This yields contradiction, and so we have the Lemma. ■

With Lemma 23, the finish is straightforward. By Markov's Inequality, with probability at least 0.8,

$$\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0) \le 5\mathbb{E}[\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0)] < \frac{T}{2}.$$

However, $\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0) < T$ implies $\tau_{\mathcal{A}_\varepsilon,T}(\mathbf{w}_0) = \tau_{\mathcal{A}_\varepsilon}(\mathbf{w}_0)$. Thus, with probability at least 0.8, we have $\tau_{\mathcal{A}_\varepsilon}(\mathbf{w}_0) < T$. That is, with probability at least 0.8 we hit $\mathcal{A}_\varepsilon = \{\mathbf{w} : F(\mathbf{w}) \le \varepsilon\}$ within T steps.

When $s = 0$ and $\gamma = 0$, we cannot use Lemma 20 anymore. But just note whenever $s = 0$, we can use the upper bound $\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^p] \le L^p \le L^3$ for $p = 2, 3$ in our upper bound of $\mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{z}_t}[\theta(\Phi(\mathbf{w}_{t+1}))]$. Defining instead

$$\text{err}(\mathbf{w}) := \frac{1}{2}\eta^2\left(2\sigma_F^2 + 4\max(L,B)^2\right) + C\eta^3\left(8\sigma_F^3 + 16\max(L,B)^3\right) + 2C\left(\frac{\eta d}{\beta}\right)^{3/2} > 0,$$

we see the rest of the proof goes through the same, with no use of Lemma 20.

The tighter results in the case when $s = 1$ (the L-smooth and (m, b)-dissipative setting) are also proved identically. They follow from plugging in the uniform moment bounds from Lemma 20 rather than the general ones into the proof of Lemma 23. Then, L_2, L_3 (which are different in this case) appear in the proof of Lemma 23 with no $\max(\eta T, 1)$ term present, and again we finish the same as above. ■

D.2. Details for Comparison to Literature

Here, we discuss how we derived optimization results using sampling results from literature, that we discussed in Subsection B.2. As mentioned there, we assume an $O(1)$ warm-start for all of the literature, which is the least favorable for us. Consider as an example how we obtained results for SGLD the smooth and dissipative case from Raginsky et al. [29], Xu et al. [35], and Zou et al. [39]. We analogously obtained results in this way from Yang and Wibisono [36], and directly cited the results of Kinoshita and Suzuki [19] as they were directly phrased in the same optimization setting as what we study.

Theorem 1 of Raginsky et al. [29] requires gradient noise δ to be exponentially small in d, which does not make sense (we only require gradient noise of constant order, which is more realistic). Theorem 3.6, Corollary 3.7, and Remark 3.9 of Xu et al. [35] reports an iteration count of $K = \widetilde{O}\left(\frac{d}{\varepsilon\lambda_*}\right)$ where λ_* is spectral gap of the discrete-time Markov Chain given by (1), however they do not count the iteration count B to compute each stochastic gradient from B data samples. Either they also require exponentially small gradient noise, or $B = \widetilde{O}\left(\frac{d^6}{\varepsilon^4\lambda_*^4}\right)$, and their total gradient complexity should be

$$K \cdot B = \widetilde{O}\left(\frac{d^7}{\varepsilon^5\lambda_*^5}\right).$$

Similarly, for the same paper's claimed runtime for Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD) in Theorem 3.10 and Corollary 3.11, noting the correct runtime should be $K \cdot B$, we obtain a runtime of

$$\widetilde{O}\left(\frac{Ld^5}{\lambda_*^4\varepsilon^4}\right) \geq \widetilde{O}\left(\frac{d^5}{\lambda_*^4\varepsilon^4}\right).$$

The last step simply follows from noting their $L \geq 1$, being the length of an inner loop.

This accounting must also for the result Theorem 4.5 and Corollary 4.7 of Zou et al. [39]. Accounting for $K \cdot B$, they obtain a rate of at least $\widetilde{O}\left(\frac{d^4\beta^2}{\rho^4\varepsilon^2}\right)$, where ρ is the Cheeger constant of μ_β, to obtain a TV distance of ε to the Gibbs measure. By Cheeger's Inequality, we have $\frac{1}{\rho^4} \geq \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^2$. However to convert from TV distance results to optimization results using Corollary 4.8 of their same paper Zou et al. [39], we need a TV distance of $\frac{\varepsilon}{d}$ (and this is necessary due to dissipativeness) to obtain an optimization result, which leads to additional dimension dependence. Combined with noting β is (at least) on the same order as $\frac{d}{\varepsilon}$ up to $\log$ factors, this gives a rate of at least

$$\widetilde{O}\left(\frac{d^8\mathsf{C}_{\mathrm{PI}}(\mu_\beta)^2}{\varepsilon^4}\right).$$

for optimizing F to $\widetilde{O}\left(\frac{d}{\beta} + \varepsilon\right) = \widetilde{O}(\varepsilon)$ tolerance.

We now discuss how we obtained results from the rest of literature. Generally the rest of literature handles exact gradients and so does not have the problem of those above two works. One point of note is that in some of the sampling literature, such as Balasubramanian et al. [4], Huang et al. [16], Vempala and Wibisono [33], sampling is done from e^{-f}/Z. That is, sampling is presumed to be done at constant temperature, a different setting than optimization. In our setting $f = \beta F$, and the smoothness parameter L or condition number in these works is that of f. Thus their smoothness parameter L scales like $\tilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. The rest of the rates from literature were then derived by converting KL divergence guarantees into TV distance guarantees via Pinkser's Inequality, and then using Corollary 4.8 of Zou et al. [39], analogously to the above example. In more detail, by Pinkser's Inequality, if F is s-Hölder continuous we need KL divergence to be at most $\frac{\varepsilon^2}{d^{s+1}}$.

Following Remark 9, it follows that the ε in the sampling results can be taken to be $\Theta(1)$. However, where ε denotes the desired optimization tolerance, the smoothness parameter L still scales like $\tilde{\Omega}\left(\frac{d}{\varepsilon}\right)$. Plugging in these choices, we obtained the results from Section 1.

As another example, we mention how we derived a rate from Corollary 19 of Balasubramanian et al. [4] (which still requires exact knowledge of gradient) in the GLD, Poincaré, and Lipschitz case. Taking $s = 0$ in Corollary 19 of Balasubramanian et al. [4], and even supposing a warm start of $K_0 = O(1)$ is possible, we see they obtain a TV distance of $\sqrt{\varepsilon}$ in

$$\widetilde{O}\left(\frac{\beta^6 d^3\mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3}{\varepsilon^5}\right).$$

However, since F is Lipschitz, we require a TV distance of $\frac{\varepsilon}{\sqrt{d}}$, the dimensionality again coming from Remark 4.6 of Zou et al. [39]. This yields a rate of

$$\widetilde{O}\left(\frac{\beta^6 d^8\mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3}{\varepsilon^{10}}\right).$$

We must have $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$, so in this case this gives a rate of at least $\widetilde{O}\left(\frac{\mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3 d^{14}}{\varepsilon^{16}}\right)$ for optimizing F to $\widetilde{O}(\varepsilon)$ tolerance. We can derive a faster rate from this result using Remark 9, which is also mentioned in Section 1.

Finally, we mention that we can compare the above results from Zou et al. [39], Chewi et al. [13], and Balasubramanian et al. [4] for general $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$; as mentioned in Section 1, to use the results of Chewi et al. [13] and Balasubramanian et al. [4] of optimization, their β dependence will be their stated dependence on the smoothness parameter L. Our dependence on β is always better than that of Zou et al. [39], and using $\beta = \widetilde{\Omega}\left(\frac{d}{\varepsilon}\right)$, we see for any such β our dependence in all parameters is better than that of Balasubramanian et al. [4], Chewi et al. [13] when $s \le \frac{1}{2}$.

Appendix E. Additional Proofs

E.1. Additional Helper Results

Here we establish many of the results we used in the main discretization proofs.

Lemma 24 *Suppose F satisfies Assumption 2. Then for all $\mathbf{w} \in \mathbb{R}^d$,*

$$\|\nabla F(\mathbf{w})\| \le L \max(1, \|\mathbf{w}^\star\|)^s (\|\mathbf{w}\|^s + 1),$$

where $\mathbf{w}^\star$ is any global minima of F. Moreover, if Assumption 5 holds, the above also holds for the stochastic gradient estimates $\|\nabla f(\mathbf{w}; \mathbf{z})\|$.

Proof Note $\nabla F(\mathbf{w}^\star) = 0$. By Triangle Inequality and Assumption 2,

$$\begin{aligned}\|\nabla F(\mathbf{w})\| &= \|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}^\star)\| \\ &\le L\|\mathbf{w} - \mathbf{w}^\star\|^s \\ &\le L(\|\mathbf{w}\| + \|\mathbf{w}^\star\|)^s \\ &\le L \max(1, \|\mathbf{w}^\star\|)^s (\|\mathbf{w}\|^s + 1).\end{aligned}$$

The last two steps used the following elementary inequalities:

$$(az + b)^s \le \max(a, b)^s (z + 1)^s \text{ for all } a, b, z \ge 0.$$

$$(z + 1)^{1/s'} \le z^{1/s'} + 1 \iff z + 1 \le (z^{1/s'} + 1)^{s'} \text{ for all } s' \ge 1.$$

The extension to stochastic gradients given Assumption 5 is immediate. ■

The following result is used to control the values of F using Assumption 2.

Lemma 25 *Suppose F satisfies Assumption 2. Then for all $\mathbf{w} \in \mathbb{R}^d$,*

$$F(\mathbf{w}) \le L\|\mathbf{w} - \mathbf{w}^\star\|^{s+1}.$$

Proof The proof is very similar to Lemma 3.4 of Bubeck et al. [6]. Let $\mathbf{w}^\star$ be any global minima of F, thus $F(\mathbf{w}^\star) = 0$ and $\nabla F(\mathbf{w}^\star) = 0$. We see from calculus and Cauchy-Schwartz that

$$F(\mathbf{w}) = |F(\mathbf{w}) - F(\mathbf{w}^\star) - \langle \nabla F(\mathbf{w}^\star), \mathbf{w} - \mathbf{w}^\star \rangle|$$

$$
\begin{aligned}
&= \left| \int_{t=0}^{1} \langle \nabla F(\mathbf{w}^\star + t(\mathbf{w} - \mathbf{w}^\star)) - \nabla F(\mathbf{w}^\star), \mathbf{w} - \mathbf{w}^\star \rangle \mathrm{d}t \right| \\
&\le \left| \int_{t=0}^{1} \|\nabla F(\mathbf{w}^\star + t(\mathbf{w} - \mathbf{w}^\star)) - \nabla F(\mathbf{w}^\star)\| \|\mathbf{w} - \mathbf{w}^\star\| \mathrm{d}t \right| \\
&\le \left| \int_{t=0}^{1} L t^s \|\mathbf{w} - \mathbf{w}^\star\|^s \|\mathbf{w} - \mathbf{w}^\star\| \mathrm{d}t \right| \\
&\le L \|\mathbf{w} - \mathbf{w}^\star\|^{s+1},
\end{aligned}
$$

where we apply Cauchy-Schwartz to obtain the first inequality and Assumption 2 for the second. ■

We also need the following simple integral to prove Lemma 14.

Lemma 26 *We have for any* $0 \le s \le 1$ *and* $M \ge 0$ *that*

$$
\int_{\mathbb{R}^d} e^{-M\|\mathbf{w}\|^{s+1}} \mathrm{d}\mathbf{w} = \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{s+1} \cdot M^{-\frac{d}{s+1}} \cdot \Gamma\left(\frac{d}{s+1}\right).
$$

Proof The surface area of $\mathcal{S}^{d-1}$ is $\frac{2\pi^{d/2}}{\Gamma(d/2)}$, which scales by r^{d-1} for an arbitrary radius r. Consider partitioning $\mathbb{R}^d$ into spheres of radius r: upon making this change of variables, which formally is $\mathrm{d}\mathbf{w} = \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{d-1} \mathrm{d}r$, we obtain

$$
\int_{\mathbb{R}^d} e^{-M\|\mathbf{w}\|^{s+1}} \mathrm{d}\mathbf{w} = \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_0^\infty e^{-Mr^{s+1}} r^{d-1} \mathrm{d}r.
$$

Let $u = r^{s+1}$, therefore $r = u^{\frac{1}{s+1}}$ and $\mathrm{d}r = \frac{1}{s+1} u^{-\frac{s}{s+1}} \mathrm{d}u$. Thus

$$
\begin{aligned}
\int_{\mathbb{R}^d} e^{-M\|\mathbf{w}\|^{s+1}} \mathrm{d}\mathbf{w} &= \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{s+1} \int_0^\infty e^{-Mu} u^{\frac{d-1-s}{s+1}} \mathrm{d}u \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{s+1} \cdot M^{-\frac{d}{s+1}} \cdot \Gamma\left(\frac{d}{s+1}\right).
\end{aligned}
$$

Here, the last equality is a well known integral essentially following from definition of the Gamma function, specifically

$$
M^{-t}\Gamma(t) = \int_0^\infty e^{-Mu} u^{t-1} \mathrm{d}u.
$$

It follows since $d \ge 1 \ge s$ and $s \ge 0$, hence $\frac{d-1-s}{s+1} = \frac{d}{s+1} - 1 \ge -1$, so we may apply these results regarding the Gamma function. ■

The last lemma is used to upper bound $z^p + 1$ for all $z \ge 0$ and any $p \ge 0$.

Lemma 27 *For all* $z \ge 0$ *and any* $p \ge 0$, $z^p + 1 \le 2(z+1)^p$.

Proof First suppose $p \ge 1$. Here we show $z^p + 1 \le (z+1)^p$, which clearly suffices. Letting $f(z) = (z+1)^p - (z^p + 1)$, we see $f'(z) \ge 0$ always. Therefore $f(z) \ge f(0) = 0$, proving this case.

Now suppose $0 \le p < 1$. Let $f(z) = \frac{(z+1)^p}{z^p+1}$. Then,

$$
f'(z) = \frac{p(z+1)^{p-1} \cdot (z^p+1) - (z+1)^p \cdot pz^{p-1}}{(z^p+1)^2} = \frac{p(z+1)^{p-1}\left(1 - z^{p-1}\right)}{(z^p+1)^2}.
$$

Therefore $f'(z) \leq 0$ for $z \in [0,1]$ and $f'(z) \geq 0$ for $z \in [1,\infty)$, so $f(z)$ is minimized on $[0,\infty)$ when $z = 1$. Hence, $f(z) \geq f(1) = 2^{p-1}$. Thus, $z^p + 1 \leq 2^{1-p}(z+1)^p \leq 2(z+1)^p$ as desired. ■