FLATNESS-AWARE STOCHASTIC GRADIENT LANGEVIN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalization in deep learning is closely tied to the pursuit of flat minima in the loss landscape, yet classical Stochastic Gradient Langevin Dynamics (SGLD) offers no mechanism to bias its dynamics toward such low-curvature solutions. This work introduces Flatness-Aware Stochastic Gradient Langevin Dynamics (fSGLD), designed to efficiently and provably seek flat minima in high-dimensional nonconvex optimization problems. At each iteration, fSGLD uses the stochastic gradient evaluated at parameters perturbed by isotropic Gaussian noise, commonly referred to as Random Weight Perturbation (RWP), thereby optimizing a randomized-smoothing objective that implicitly captures curvature information. Leveraging these properties, we prove that the invariant measure of fSGLD stays close to a stationary measure concentrated on the global minimizers of a loss function regularized by the Hessian trace whenever the inverse temperature and the scale of random weight perturbation are properly coupled. This result provides a rigorous theoretical explanation for the benefits of random weight perturbation. In particular, we establish non-asymptotic convergence guarantees in Wasserstein distance with the best known rate and derive an excess-risk bound for the Hessian-trace regularized objective. Extensive experiments on noisy-label and large-scale vision tasks, in both training-from-scratch and fine-tuning settings, demonstrate that fSGLD achieves superior or comparable generalization and robustness to baseline algorithms while maintaining the computational cost of SGD, about half that of SAM. Hessian-spectrum analysis further confirms that fSGLD converges to significantly flatter minima.

1 INTRODUCTION

Consider the overdamped Langevin dynamics governed by the stochastic differential equation (SDE)

$$\mathrm{d}Z_t = -\nabla u(Z_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \tag{1}$$

which admits a unique invariant (Gibbs) measure $\pi_\beta(\theta)$ proportional to $\exp(-\beta u(\theta))$, where $\beta > 0$ is the inverse temperature and $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. As β increases, this Gibbs measure concentrates on the global minimizers of u, establishing a direct link between Langevin dynamics and global optimization. Building on this property, Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011; Raginsky et al., 2017) was proposed as the Euler-Maruyama discretization of the Langevin SDE in which the exact gradient ∇u is replaced by a stochastic gradient. SGLD has attracted considerable attention as a prominent optimization algorithm for nonconvex problems, and under mild regularity conditions a series of works has established non-asymptotic global convergence guarantees (Raginsky et al., 2017; Xu et al., 2018; Majka et al., 2020; Chau et al., 2021; Zhang et al., 2023). Despite these elegant theoretical results, SGLD has not become a widely used optimizer in deep learning practice, largely because it lacks an intrinsic mechanism to favor flat minima, which are closely associated to strong generalization.

Alongside advances in SGLD, a separate line of work in deep learning has explored flatter solutions to improve generalization, inspired by the flat minima hypothesis (Hochreiter & Schmidhuber, 1997). As a result, numerous flatness-aware optimization algorithms have been developed, including Random Weight Perturbation (RWP) (Bisla et al., 2022; Li et al., 2024a), Entropy-SGD (Chaudhari et al., 2017), Entropy-MCMC (Li & Zhang, 2024), Sharpness-Aware Minimization (SAM) (Foret et al., 2021) and their variants (Xie et al., 2024; Li et al., 2024b; Tahmasebi et al., 2024; Luo et al.,

2024; Chen et al., 2024; Kang et al., 2025; Wei et al., 2025; Liu et al., 2022a;b; Du et al., 2022b; Li et al., 2025). In principle, flatness-aware optimization promotes exploration of flat regions by replacing the standard stochastic gradient with a perturbed gradient. For example, SAM applies a worst-case adversarial perturbation within a local neighborhood, whereas RWP uses symmetric random noise to generate the gradient perturbation and can be viewed as computing the stochastic gradient of a randomized-smoothing objective (Duchi et al., 2012). However, SAM's min–max formulation requires double gradient evaluations, leading to roughly twice the computational cost of standard SGD. On the theoretical side, recent studies have produced important advances in the analysis of SAM and related flatness-aware optimization methods, yielding valuable insights on generalization bounds, stability, and (local) convergence properties; e.g., see Andriushchenko & Flammarion (2022); Bartlett et al. (2023); Si & Yun (2023); Yu et al. (2024); Khanh et al. (2024); Oikonomou & Loizou (2025); Zhang et al. (2024); Li et al. (2024a). However, with a few notable exceptions (Ahn et al., 2024; Gatmiry et al., 2024), the global convergence properties of flatness-aware optimization in nonconvex settings, as well as a rigorous theoretical understanding of the role of RWP, remain relatively unexplored.

To address these challenges, we introduce Flatness-Aware Stochastic Gradient Langevin Dynamics (fSGLD), a principled synthesis of randomized smoothing and Langevin dynamics that efficiently explores flat minima. While randomized-smoothing surrogates are known to encode second-order information such as the Hessian trace, they also contain higher-order remainder terms of which effects are not negligible in high-dimensional nonconvex problems, weakening the intended flatness-aware regularization effect. Our key theoretical contribution is to show that when the two key hyperparameters, the inverse temperature parameter β and the perturbation scale σ, are properly balanced, the invariant measure of fSGLD concentrates on the global minimizers of the true Hessian-trace regularized objective, thereby isolating the genuine flatness-aware regularization effect. This principled coupling is crucial, as it ensures that the global exploration driven by Langevin dynamics is effectively guided across a landscape smoothed by the perturbation noise, steering the process toward genuinely flat regions. In particular, we establish non-asymptotic convergence guarantees in Wasserstein distance and an explicit excess-risk bound for the Hessian-trace-regularized objective, providing the rigorous evidence of the benefits of RWP in nonconvex settings. Our framework bridges and advances the theory and practice of flatness-aware stochastic optimization, opening new avenues to incorporate geometric smoothing into Langevin sampling and paving the way for more effective and principled flatness-regularized learning. To validate these results, we evaluate fSGLD on noisy-label datasets (CIFAR-10N/100N, WebVision) and large-scale vision fine-tuning (ViT-B/16). Extensive experiments demonstrate that fSGLD consistently matches or outperforms baselines including SGD, AdamW, SGLD, and SAM in generalization and robustness while maintaining the computational cost of standard SGD. Notably, using the theoretically prescribed coupling between β and σ yields substantially better performance than simply fixing a large β, which is the common SGLD practice. In summary, fSGLD is the first to combine the SGLD framework with the concept of flatness and to provide a global convergence analysis for flatness-aware optimization, thereby advancing the theoretical and practical foundations of both areas.

2 Problem Setting and FSGLD Algorithm

Notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a fixed probability space. We denote the probability law of a random variable Y by $\mathcal{L}(Y)$. Fix integers $d, m \geq 1$. Let I_d be the identity matrix of dimension d. The Euclidean scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $|\cdot|$ standing for the corresponding norm. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function, and we denote its gradient by ∇f. For any integer $q \geq 1$, let $\mathcal{P}(\mathbb{R}^q)$ be the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures Γ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ and ν. For any μ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, the Wasserstein distance of order $p \geq 1$ is defined as

$$W_p(\mu, \nu) = \left(\inf_{\Gamma \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x - y|^p \, \mathrm{d}\Gamma(x, y) \right)^{\frac{1}{p}}. \tag{2}$$

2.1 Intractable Hessian-based Regularization

We consider the following nonconvex stochastic optimization problem:

$$\min_{\theta \in \mathbb{R}^d} u(\theta) := \min_{\theta \in \mathbb{R}^d} \mathbb{E}\big[U(\theta, X)\big], \tag{3}$$

where $u : \mathbb{R}^d \to \mathbb{R}$ is a four-times continuously differentiable function with gradient $h := \nabla u$, $U : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is a measurable function satisfying $\mathbb{E}[|U(\theta, X)|] < \infty$ for all $\theta \in \mathbb{R}^d$, and X is a random variable with probability law $\mathcal{L}(X)$. In practice, the gradient h of u is usually unknown and one only has access to its unbiased estimate, i.e. $h(\theta) = \mathbb{E}[\nabla_\theta U(\theta, X)]$.

To improve generalization, we incorporate an inductive bias for flatness through a flatness-aware objective. More specifically, instead of optimizing the original objective u, we aim to solve the following *Hessian-trace regularized objective*:

$$v(\theta) := u(\theta) + \frac{\sigma^2}{2}\mathrm{tr}\left(H(\theta)\right), \tag{4}$$

where $\mathrm{tr}(H(\theta))$ is the trace of the Hessian of u evaluated at θ and $\sigma > 0$ controls the strength of the sharpness regularization. The global minimizers of this regularized objective v represent a trade-off between low loss from the original objective u and low curvature. For brevity, we will refer to these points as the *global flat minima* (i.e., $\arg\min_{\theta\in\mathbb{R}^d} v(\theta)$). However, computing $\mathrm{tr}\left(H(\theta)\right)$ is expensive in high dimension.

2.2 Randomized Smoothing as a Tractable Surrogate

To obtain a tractable alternative to the Hessian-trace regularized objective in 4, we introduce a Gaussian perturbation $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \in (0, 1)$, independent of X, and define the *randomized-smoothing surrogate objective:*

$$g_\epsilon(\theta) := \mathbb{E}\big[u(\theta + \epsilon)\big] = \mathbb{E}\big[\mathbb{E}_X\big[U(\theta + \epsilon, X)\big]\big]. \tag{5}$$

where the outer expectation is taken with respect to the noise ϵ and $\mathbb{E}_X[\cdot]$ denotes the conditional expectation given ϵ. This simple surrogate allows us to access curvature information. By Taylor's theorem, we have

$$u(\theta + \epsilon) = u(\theta) + \nabla u(\theta)^\top \epsilon + \tfrac{1}{2}\epsilon^\top H(\theta)\epsilon + \mathcal{R}(\theta, \epsilon),$$

where $\mathcal{R}(\theta, \epsilon)$ is the remainder term. Taking the expectation over $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ yields the key connection:

$$\begin{aligned} g_\epsilon(\theta) &= u(\theta) + \tfrac{\sigma^2}{2}\mathrm{tr}\left(H(\theta)\right) + \mathbb{E}[\mathcal{R}(\theta, \epsilon)] \\ &= v(\theta) + \mathbb{E}[\mathcal{R}(\theta, \epsilon)]. \end{aligned} \tag{6}$$

Thus, optimizing the tractable surrogate g_ϵ introduces the desired inductive bias toward flat minima by implicitly minimizing the Hessian–trace regularized objective v, provided that the remainder term $\mathbb{E}[\mathcal{R}(\theta, \epsilon)]$ is negligible.

2.3 FSGLD Algorithm

To optimize the surrogate objective g_ϵ in 5, we propose the Flatness-Aware Stochastic Gradient Langevin Dynamics (fSGLD) algorithm. Formally, let θ_0 be an $\mathbb{R}^d$-valued random variable representing the initial value, $(X_k)_{k\in\mathbb{N}}$ be an i.i.d sequence of data, $(\epsilon_k)_{k\in\mathbb{N}}$ be i.i.d copies of the Gaussian perturbation $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$, and $(\xi_k)_{k\in\mathbb{N}}$ be an independent sequence of standard d-dimensional Gaussian random variables. We assume that θ_0, $(\epsilon_k)_{k\in\mathbb{N}}$, and $(\xi_k)_{k\in\mathbb{N}}$ are all mutually independent. Then, the fSGLD algorithm is given by

$$\begin{cases} \theta_0^{\text{fSGLD}} & := \theta_0, \\ \theta_{k+1}^{\text{fSGLD}} & = \theta_k^{\text{fSGLD}} - \lambda\nabla_\theta U(\theta_k^{\text{fSGLD}} + \epsilon_{k+1}, X_{k+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{k+1}, \qquad k \in \mathbb{N} \end{cases} \tag{7}$$

where $\lambda > 0$ is the stepsize, $\beta > 0$ is the inverse temperature. We make three important remarks about this update rule. First, the gradient term in 7 is a unbiased stochastic gradient of g_ϵ, as its expectation over both the data X and the perturbation ϵ recovers the true gradient ∇g_ϵ:

$$\nabla g_\epsilon(\theta) = \mathbb{E}[\mathbb{E}_X[\nabla_\theta U(\theta + \epsilon, X)]]. \tag{8}$$

Second, the fSGLD can be interpreted as the standard SGLD for the original objective u combined with RWP. Third, under appropriate conditions, which will be introduced in the next section, the fSGLD algorithm generates a Markov chain that converges to a unique invariant (Gibbs) measure. This measure, denoted by π_β^{fSGLD}, is associated with the randomized-smoothing surrogate objective g_ϵ, i.e., $\pi_\beta^{\text{fSGLD}}(\theta) \propto \exp(-\beta g_\epsilon(\theta))$. The formal convergence guarantees are provided in Appendix C.2.

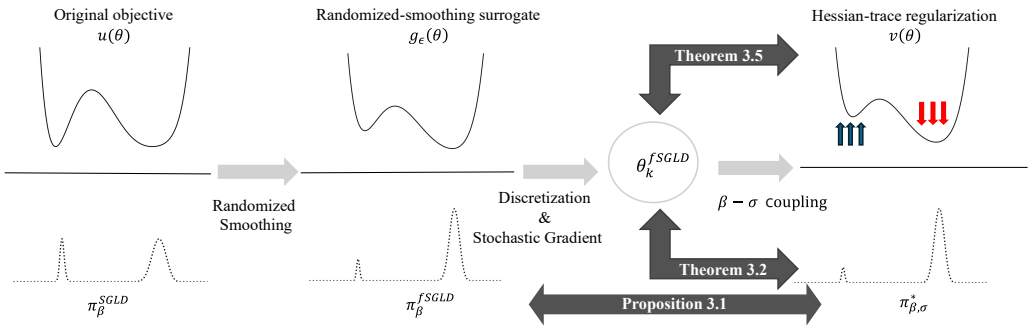


Figure 1: A schematic overview of the theoretical framework of fSGLD. The process begins with the **original objective** $u(\theta)$ and its associated Gibbs measure $\pi_\beta^{\mathrm{SGLD}}$ (left). **Randomized smoothing** transforms this into a tractable **surrogate objective**, $g_\epsilon(\theta)$, which is the basis for the fSGLD algorithm and its invariant measure, $\pi_\beta^{\mathrm{fSGLD}}$ (center). This highlights a key distinction: while the Gibbs measure of standard SGLD, $\pi_\beta^{\mathrm{SGLD}}$, is indifferent to the flatness of the minima, the fSGLD framework is designed such that its invariant measure, $\pi_\beta^{\mathrm{fSGLD}}$, targets the distribution over the flattest minima. Our ultimate goal is to target the **Hessian-trace regularized objective** $v(\theta)$ and its corresponding measure $\pi_{\beta,\sigma}^{\star}$, which concentrates on the desired global flat minima (right).

3 Theoretical Results

In this section, we present the main theoretical results that rigorously validate the fSGLD algorithm. We begin by stating the formal assumptions for our analysis. We then prove that the invariant measure of fSGLD converges to an ideal target distribution over flat minima when its key hyperparameters β and σ are properly coupled. Building on this, we derive non-asymptotic convergence guarantees for the fSGLD iterates in both Wasserstein distance and for the excess risk. The logical flow of our theoretical framework is summarized in the schematic illustration in Figure 1.

3.1 Assumptions

We first state the formal assumptions for our main theoretical results. Specifically, our assumptions impose standard conditions on: (i) moments of the initial parameters, stochastic gradient, and the noise processes; (ii) a Lipschitz condition on the stochastic gradient; and (iii) a dissipativity condition to ensure the stability of the Langevin dynamics.

Assumption 1 (Moments of the initial parameter, stochastic gradient, and independence of the data and noise perturbation)**.** *We assume the initial parameter θ_0 has a finite fourth moment, $\mathbb{E}[|\theta_0|^4] < \infty$, and that we have access to an unbiased stochastic gradient for the original objective u, $\mathbb{E}[\nabla_\theta U(\theta, X)] = h(\theta)$, where the data sequence $(X_k)_{k\in\mathbb{N}}$ is i.i.d. Furthermore, the perturbation noise $(\epsilon_k)_{k\in\mathbb{N}} \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \in (0, 1)$, $(X_k)_{k\in\mathbb{N}}$, $(\xi_k)_{k\in\mathbb{N}} \sim \mathcal{N}(0, I_d)$, and θ_0 are mutually independent.*

We discuss how the sampling model in Assumption 1 relates to the sampling scheme used in our numerical experiments (Section 4) in Remark B.4.

Assumption 2 (Lipschitzness)**.** *There exists $\varphi : \mathbb{R}^m \to [1, \infty)$ with $\mathbb{E}[|(1 + |X_0|)\varphi(X_0)|^4] < \infty$, and constants $L_1, L_2 > 0$ such that, for all x, $x' \in \mathbb{R}^m$ and θ, $\theta' \in \mathbb{R}^d$,*

$$|\nabla_\theta U(\theta, x) - \nabla_{\theta'} U(\theta', x)| \le L_1 \varphi(x)|\theta - \theta'|,$$
$$|\nabla_\theta U(\theta, x) - \nabla_\theta U(\theta, x')| \le L_2(\varphi(x) + \varphi(x'))(1 + |\theta|)|x - x'|,$$

Assumption 3 (Dissipativity)**.** *There exist a measurable function (symmetric matrix-valued) function $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$ and a measurable function $\hat{b} : \mathbb{R}^m \to \mathbb{R}$ such that for any $x \in \mathbb{R}^m$, $y \in \mathbb{R}^d$, $\langle y, A(x)y\rangle \ge 0$ and for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,*

$$\langle \nabla_\theta U(\theta, x), \theta\rangle \ge \langle \theta, A(x)\theta\rangle - \hat{b}(x).$$

The smallest eigenvalue of $\mathbb{E}[A(X_0)]$ is a positive real number $\bar{a} > 0$ and $\mathbb{E}[\hat{b}(X_0))] = \bar{b} > 0$.

Note that this dissipativity condition is a standard requirement for analysis of SGLD in the literature; e.g., see Raginsky et al. (2017); Xu et al. (2018); Deng et al. (2020a;b; 2022); Futami & Fujisawa (2023). In particular, our version in Assumption 3 follows the more general formulation of Zhang et al. (2023), which allows for dependency on the data X. Moreover, several direct consequences of these assumptions, which are useful for our subsequent analysis, are detailed in Appendix B.

3.2 Target Gibbs Measure for Global Flat Minima

Our analysis begins by defining the ideal target distribution which concentrates on the global flat minima. The natural choice is the Gibbs measure associated with v, which we define as $\pi^{\star}_{\beta,\sigma}$:

$$\pi^{\star}_{\beta,\sigma}(\mathrm{d}\theta) \propto \exp(-\beta v(\theta))\mathrm{d}\theta. \tag{9}$$

By construction, as the inverse temperature $\beta \to \infty$, this measure concentrates on the global flat minima.

The central question is whether the invariant measure of fSGLD, $\pi^{\mathrm{fSGLD}}_{\beta}$, converges to this ideal Gibbs measure $\pi^{\star}_{\beta,\sigma}$. For these two Gibbs measures to align, the remainder term $\mathbb{E}[\mathcal{R}(\theta,\epsilon)]$ in 6 must be negligible. In high-dimensional nonconvex problems, this is a non-trivial condition, as higher-order terms can be substantial and unpredictable, potentially corrupting the intended regularization effect. For this reason, in the low-temperature limit ($\beta \to \infty$), a careful interplay between the inverse temperature β and the noise scale σ becomes essential. Please refer to Appendix A for the formal relationship between two Gibbs measures $\pi^{\mathrm{fSGLD}}_{\beta}$ and $\pi^{\star}_{\beta,\sigma}$.

The following proposition shows that when the perturbation scale σ and inverse temperature β are properly coupled, the invariant measure of fSGLD converges to the ideal target measure in the Wasserstein distance of order two.

Proposition 3.1. *Let Assumptions 1, 2, and 3 hold, and let $\sigma = \beta^{-\frac{1+\eta}{4}}$ for $\eta \in (0,1)$. Then*

$$W_2(\pi^{fSGLD}_{\beta}, \pi^{\star}_{\beta,\sigma}) \leq \underline{D},$$

where $\underline{D} = O(\beta^{-\frac{\eta}{2}} d \log d)$, whose explicit expression is given in 57. Moreover,

$$\lim_{\beta\to\infty} W_2(\pi^{fSGLD}_{\beta}, \pi^{\star}_{\beta,\sigma}) = 0. \tag{10}$$

The proof of Proposition 3.1 is postponed to Appendix C.2. This proposition rigorously shows how RWP induces the desired Hessian-trace regularization effect through a theoretically-prescribed coupling of the two key hyperparameters, σ and β. As demonstrated in our experiments, this coupling yields meaningful improvements in generalization.

3.3 Convergence Guarantees for fSGLD

Having established that fSGLD correctly targets the ideal distribution for flat minima, our first main result provides non-asymptotic error bounds on the Wasserstein-1 and -2 distances between the law of the k-th fSGLD iterate $\mathcal{L}(\theta^{\mathrm{fSGLD}}_{k})$ and the target Gibbs measure $\pi^{\star}_{\beta,\sigma}$. All proofs for the results in this section are provided in Appendix C.2.

Theorem 3.2. *Let Assumptions 1, 2, and 3 hold, and let $\sigma = \beta^{-\frac{1+\eta}{4}}$ for $\eta \in (0,1)$. Then, there exist constants $\dot{c}, D_1, D_2, D_3, \underline{D} > 0$ such that, for every $\beta > 0$, for $0 < \lambda \leq \lambda_{max}$ with λ_{max} given in 24, and $k \in \mathbb{N}$,*

$$W_1(\mathcal{L}(\theta^{fSGLD}_{k}), \pi^{\star}_{\beta,\sigma}) \leq D_1 e^{-\dot{c}\lambda k/2}(1 + \mathbb{E}[|\theta_0|^4]) + (D_2 + D_3)\sqrt{\lambda} + \underline{D}, \tag{11}$$

where $\dot{c}$ is given in Lemma C.6, and

$$D_1 = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}}}\right)\right), \quad \text{with } D_\star > 0 \text{ independent of } d, \beta, k,$$

$$D_2 = O\left(1 + \sqrt{\frac{d}{\beta}}\right), \quad D_3 = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}}}\right)\right), \quad \underline{D} = O(\beta^{-\frac{\eta}{2}} d \log d).$$

The explicit expressions of D_1, D_2, D_3 are given in 72, and $\underline{D}$ is given in 57. Furthermore, let $\beta_{\bar{\delta}}$, $\lambda_{\bar{\delta}}$, $k_{\bar{\delta}}$ be as in 76, 77, and 78 respectively. For any $\bar{\delta} > 0$, if we choose $\beta \geq \beta_{\bar{\delta}}$, $\lambda \leq \lambda_{\bar{\delta}}$, and $k \geq k_{\bar{\delta}}$, then

$$W_1(\mathcal{L}(\theta_k^{fSGLD}), \pi^\star_{\beta,\sigma}) \leq \bar{\delta}.$$

Corollary 3.3. *Let Assumption 1, 2 and 3 hold, and let $\sigma = \beta^{-\frac{1+\eta}{4}}$ for $\eta \in (0,1)$. Then, there exists constants $\dot{c}, D_4, D_5, D_6, \underline{D} > 0$ such that, for every $\beta > 0$, $0 < \lambda \leq \lambda_{max}$ with λ_{max} given in 24, and $k \in \mathbb{N}$,*

$$W_2(\mathcal{L}(\theta_k^{fSGLD}), \pi^\star_{\beta,\sigma}) \leq D_4 e^{-\dot{c}\lambda k/4}(\mathbb{E}[|\theta_0|^4] + 1) + (D_5 + D_6)\lambda^{1/4} + \underline{D}, \tag{12}$$

where $\dot{c}$ is given in Lemma C.6, and,

$$D_4 = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right), \quad \text{with } D_\star > 0 \text{ independent of } d,\ \beta,\ k,$$

$$D_5 = O\left(1 + \sqrt{\frac{d}{\beta}}\right), \quad D_6 = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right), \quad \underline{D} = O(\beta^{-\frac{\eta}{2}} d \log d).$$

The explicit expressions of D_4, D_5, D_6 are given in 74, and $\underline{D}$ is given in 57. In addition, let $\beta_{\widetilde{\delta}}$, $\lambda_{\widetilde{\delta}}$, $k_{\widetilde{\delta}}$ be as in 80, 81, and 82 respectively. For any $\widetilde{\delta} > 0$, if we choose $\beta \geq \beta_{\widetilde{\delta}}$, $\lambda \leq \lambda_{\widetilde{\delta}}$, and $k \geq k_{\widetilde{\delta}}$, then

$$W_2(\mathcal{L}(\theta_k^{fSGLD}), \pi^\star_{\beta,\sigma}) \leq \widetilde{\delta}.$$

Remark 3.4. *We emphasize that Theorem 3.2 and Corollary 3.3 recover the best known convergence results for SGLD under comparable assumptions, see e.g. Zhang et al. (2023). Unfortunately, the constants D_1, D_3, D_4, D_6 have exponential dependence on d and β due to the coupling arguments of Eberle et al. (2019), commonly used in the SGLD literature (Chau et al., 2021; Zhang et al., 2023). In this setting, any improvement in the dimension dependence would necessitate substantially strengthening the contraction-rate estimates in Eberle et al. (2019, Theorem 2.2).*

Remark 3.5. *The proofs of Theorem 3.2 and Corollary 3.3 rely on the following decomposition:*

$$\begin{aligned} W_p(\mathcal{L}(\theta_k^{fSGLD}), \pi^\star_{\beta,\sigma}) \leq\ & W_p(\mathcal{L}(\theta_k^{fSGLD}), \mathcal{L}(Z_t^{\lambda,fSGLD})) + W_p(\mathcal{L}(Z_t^{\lambda,fSGLD}), \pi_\beta^{fSGLD}) \\ & + W_p(\pi_\beta^{fSGLD}, \pi^\star_{\beta,\sigma}), \quad p = \{1,2\}, \quad t \in (kT, (k+1)T]. \end{aligned} \tag{13}$$

The first term on the right-hand side of 13 corresponds to the discretization error between the fSGLD recursion 7 and the time-rescaled version of flatness Langevin SDE 26 associated with the randomized-smoothing surrogate objective g_ϵ defined in 5. The second term captures the convergence error between this SDE and its invariant measure π_β^{fSGLD}. The third term is the distance between the two measures provided in Proposition 3.1. The proofs of the first two error terms follow the general structure of Chau et al. (2021); Zhang et al. (2023), but require substantial adaptation to handle the fSGLD update (instead of SGLD) as well as the surrogate objective function g_ϵ (instead of the original objective u). The proof of the third error term is entirely new.

While the previous results guarantee convergence from a sampling perspective, our final result analyzes fSGLD as an optimizer. The following theorem provides a non-asymptotic bound on the expected excess risk with respect to the Hessian-trace regularized objective v.

Theorem 3.6. *Let Assumption 1, 2 and 3 hold, and let $\sigma = \beta^{-\frac{1+\eta}{4}}$ for $\eta \in (0,1)$. Then, there exist constants $\dot{c}$, $D_1^\#$, $D_2^\#$, $D_3^\# > 0$ such that, for every $\beta > 0$, $0 < \lambda \leq \lambda_{max}$ with λ_{max} given in 24, $k \in \mathbb{N}$,*

$$\mathbb{E}[g_\epsilon(\theta_k^{fSGLD})] - \inf_{\theta \in \mathbb{R}^d} v(\theta) \leq D_1^\# e^{-\dot{c}\lambda k/4} + D_2^\# \lambda^{1/4} + D_3^\#, \tag{14}$$

where $\dot{c}$ is given in Lemma C.6, and

$$D_1^\# = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right),$$

$$D_2^\# = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right),$$

$$D_3^\# = O\left((d/\beta)\log(D_\star(\beta^{(1-\eta)/2}+1))\right).$$

The explicit expressions of $D_1^{\#}$ and $D_2^{\#}$ are given in 85, $D_3^{\#}$ is defined in 90. Moreover, let $\beta_{\underline{\delta}}$, $\lambda_{\underline{\delta}}$, $k_{\underline{\delta}}$ be as in 91, 92, and 93 respectively. For any $\underline{\delta} > 0$, if we choose $\beta \geq \beta_{\underline{\delta}}$, $\lambda \leq \lambda_{\underline{\delta}}$, and $k \geq k_{\underline{\delta}}$, then

$$\mathbb{E}[g_\epsilon(\theta_k^{fSGLD})] - \inf_{\theta \in \mathbb{R}^d} v(\theta) \leq \underline{\delta}.$$

Remark 3.7. *In this special case when the perturbation scale $\sigma \to 0$, the fSGLD update reduces exactly to SGLD algorithm and the constants in the non-asymptotic bounds in Theorem 3.2, Corollary 3.3, and Theorem 3.6 coincide with the ones in (Zhang et al., 2023).*

This result provides a rigorous guarantee that fSGLD finds global flat minima by effectively solving the Hessian-trace regularized objective.

4 Numerical Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three challenging noisy label datasets including CIFAR-10N and CIFAR-100N (Wei et al., 2022), and WebVision (Li et al., 2017). CIFAR-10N and CIFAR-100N include real-world annotation errors introduced by human annotators, offering realistic yet standardized benchmarks for noisy label learning. For CIFAR-10N, we use the aggregate noise setting. WebVision is a large-scale, in-the-wild benchmark, consisting of more than 2.4 million images with labels automatically collected from Google and Flickr based on the 1,000 ImageNet ILSVRC2012 categories. Following standard protocol Li et al. (2020); Ortego et al. (2021); Li et al. (2022), we use the first 50 classes from its Google image subset and report Top-1 (WV-1) and Top-5 (WV-5) accuracy on the official validation set.

Models. We use ResNet-34 and ResNet-50 for training from scratch. For fine-tuning experiments, we use the pre-trained ViT-B/16 (Dosovitskiy et al., 2021) architecture, which has been trained on the ImageNet-1K (Deng et al., 2009) dataset as the backbone on CIFAR-10N and CIFAR-100N.

Baselines and Implementation Details. We compare fSGLD against four baselines: SGD with momentum, AdamW (Loshchilov & Hutter, 2019), SGLD (Welling & Teh, 2011), and SAM (Foret et al., 2021). To ensure a fair comparison, all optimizer hyperparameters are tuned using Optuna (Akiba et al., 2019) with 20 trials of Bayesian optimization. For each optimizer, the search spaces were carefully chosen to include previously reported optimal hyperparameters from the literature, ensuring that all baselines are strongly tuned. For fSGLD, we search for the optimal noise scale σ, while the inverse temperature β is determined by our theoretically-prescribed coupling, $\beta = \sigma^{-4/(1+\eta)}$ with $\eta = 0.01$. For experiments with training from scratch, all experiments are trained for 150 epochs with a batch size of 128. The learning rate decays by a factor of 0.1 in the 50th and 100th epochs. For fine tuning, models are trained for 75 epochs with a batch size of 128, decaying the rate by a factor of 0.1 at the 50th epoch. The detailed hyperparameter search spaces for each optimizer and experimental settings are provided in Appendix D.1.

4.2 Empirical performance on real-world noisy label datasets

Training from scratch. We first evaluate the performance of all optimizers when training ResNet models from scratch. Table 1 presents the results across all dataset-architecture combinations. Our proposed method, fSGLD (β-σ coupled), consistently achieves the best or second-best performance on every benchmark. Notably, on the CIFAR-100N dataset which presents significant challenges due to its higher noise ratio and larger number of classes, fSGLD significantly outperforms all baselines.

In terms of computational cost, the wall-clock time per iteration (s/iter) shows that fSGLD has a training speed comparable to standard optimizers like SGD, AdamW, and SGLD. In contrast, SAM incurs nearly double the computational overhead due to its min-max formulation requiring two gradient evaluations per step. This highlights a key advantage of our method: fSGLD matches or surpasses SAM's strong performance with a computational budget similar to standard SGD.

Table 1: Performance comparison on ResNet-34 and ResNet-50. Results are reported as mean±std over five different random seeds. Within each model block, the best result is **bold** and the second-best is underlined. WV-1/WV-5 denote Top-1/Top-5 accuracy on WebVision. The wall-clock time per iteration (s/iter) measured on CIFAR-10N for each model architecture.

Model	Optimizer	CIFAR-10N	CIFAR-100N	WV-1	WV-5	(s/iter)
ResNet-34	SGD	$89.31_{\pm 0.84}$	$58.47_{\pm 0.20}$	$71.87_{\pm 0.44}$	$89.33_{\pm 0.30}$	22.0
	AdamW	$89.25_{\pm 0.66}$	$56.77_{\pm 0.47}$	$68.69_{\pm 0.32}$	$87.01_{\pm 0.24}$	22.5
	SAM	$91.53_{\pm 0.22}$	$59.18_{\pm 0.33}$	$\underline{73.49}_{\pm 0.36}$	$\underline{90.32}_{\pm 0.31}$	41.3
	SGLD	$88.77_{\pm 0.51}$	$57.33_{\pm 0.36}$	$70.87_{\pm 0.67}$	$88.06_{\pm 0.30}$	22.2
	fSGLD (β-σ coupled)	$\mathbf{91.72}_{\pm 0.20}$	$\mathbf{62.02}_{\pm 0.29}$	$\mathbf{73.55}_{\pm 0.27}$	$89.86_{\pm 0.12}$	23.7
	fSGLD (β fixed)	$\underline{91.56}_{\pm 0.19}$	$\underline{61.55}_{\pm 0.45}$	$73.23_{\pm 0.34}$	$\mathbf{90.63}_{\pm 0.38}$	23.7
ResNet-50	SGD	$89.41_{\pm 0.26}$	$57.52_{\pm 0.17}$	$71.11_{\pm 0.59}$	$88.31_{\pm 0.40}$	31.9
	AdamW	$89.26_{\pm 0.31}$	$57.28_{\pm 0.90}$	$69.92_{\pm 0.67}$	$87.97_{\pm 0.34}$	32.3
	SAM	$\underline{90.88}_{\pm 0.49}$	$59.01_{\pm 0.60}$	$72.52_{\pm 0.46}$	$89.53_{\pm 0.44}$	60.7
	SGLD	$88.89_{\pm 0.40}$	$56.90_{\pm 0.65}$	$69.43_{\pm 0.40}$	$87.17_{\pm 0.22}$	32.1
	fSGLD (β-σ coupled)	$\mathbf{91.26}_{\pm 0.08}$	$\mathbf{62.08}_{\pm 0.45}$	$\mathbf{73.31}_{\pm 0.50}$	$\mathbf{90.07}_{\pm 0.20}$	34.1
	fSGLD (β fixed)	$90.72_{\pm 0.29}$	$\underline{61.56}_{\pm 1.08}$	$\underline{72.87}_{\pm 0.64}$	$\underline{89.59}_{\pm 0.41}$	34.1

Fine-tuning. We also evaluate performance in the fine-tuning setting, using a pre-trained ViT-B/16 model on CIFAR-10N and CIFAR-100N. The results are presented in Table 2. Our method, fSGLD (β-σ coupled), consistently outperforms standard optimizers like SGD and SGLD, and achieves performance competitive with or superior to SAM at roughly half the computational overhead.

4.3 Ablation Study: The Effect of the β-σ Coupling

To empirically validate our theoretical claim, we examine the effect of the theoretically-prescribed β-σ coupling. We compare fSGLD (β-σ coupled) against fSGLD (β fixed) which reflects a common heuristic of setting a large, fixed β for optimization. The results, summarized in Table 1 and Table 2, show that the coupled version consistently outperforms the fixed version in all settings, with the single exception of the WV-5 metric on ResNet-34. This provides strong empirical evidence that our theoretically-prescribed coupling is crucial for improving performance.

Table 2: Fine-tuning performance comparison on ViT-B/16.

Model	ViT-B/16		
Dataset	CIFAR-10N	CIFAR-100N	(s/epoch)
SGD	94.64	71.80	343.2
AdamW	95.57	72.30	344.5
SAM	**96.75**	74.66	656.7
SGLD	94.13	71.36	344.8
fSGLD (β fixed)	96.70	$\underline{75.16}$	345.8
fSGLD (β-σ coupled)	$\underline{96.72}$	**75.18**	345.8

4.4 Sensitivity Analysis

Table 3: Performance with respect to the number of random perturbations n used in fSGLD.

	CIFAR-10N	(s/epoch)
$n = 1$	$91.72_{\pm 0.18}$	23.7
$n = 2$	$91.57_{\pm 0.18}$	41.8
$n = 3$	$91.79_{\pm 0.17}$	60.4
$n = 4$	$92.04_{\pm 0.13}$	78.5
$n = 5$	$91.83_{\pm 0.19}$	97.0

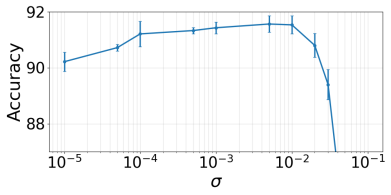
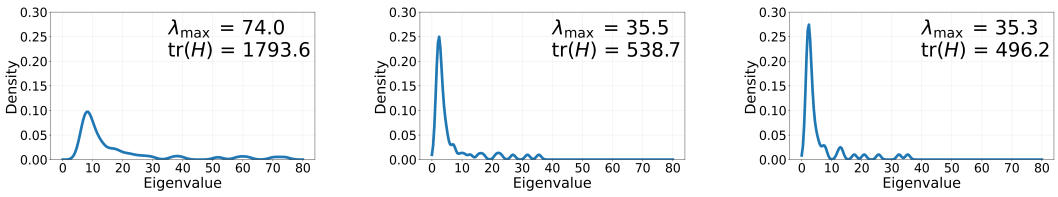


Figure 2: Sensitive analysis of noise standard deviation σ on CIFAR-10N with ResNet-34.

While our fSGLD algorithm uses a single perturbation per iteration ($n = 1$), we examine how performance is affected by using multiple perturbations, which can provide a more accurate estimation of the Hessian trace. As shown in Table 3, increasing n can improve accuracy, but this comes at a nearly linear increase in computational cost. Remarkably, fSGLD already achieves strong performance with just a single perturbation, making $n = 1$ a practical and efficient choice.

Next, we analyze the effect of the perturbation scale σ, as illustrated in Figure 2. The performance on CIFAR-10N remains stable and robust across a wide range of small to moderate values of σ. However, performance degrades sharply when σ becomes excessively large, as the strong perturbations begin to destabilize the training process.

4.5 Hessian spectrum



(a) SGD (b) SAM (c) fSGLD

Figure 3: The distribution of the leading eigenvalues and Hessian trace of ResNet-34 trained on CIFAR-10N with SGD, SAM, and fSGLD.

To empirically verify our theoretical insight that fSGLD finds flat minima by implicitly regularizing the Hessian trace, we analyze the curvature of the loss landscape at the solutions found by SGD, SAM, fSGLD. Note that we use the best hyperparameter configuration for each optimizer.

We compute two standard measures of sharpness for a ResNet-34 trained on CIFAR-10N: the maximum eigenvalue ($\lambda_{\max}$) of the Hessian and its trace ($\text{tr}(H(\theta))$). Since exact computation is intractable, we estimate the top 50 eigenvalues using the Lanczos algorithm (Lin et al., 2016; Ghorbani et al., 2019) and approximate the trace with Hutchinson's method (Avron & Toledo, 2011; Ubaru et al., 2017). Detailed settings are described in Appendix D.2.

The results, presented in Figure 3, confirm our hypothesis. fSGLD converges to solutions with a significantly smaller maximum eigenvalue and Hessian trace compared to standard SGD. Remarkably, the degree of flatness achieved by fSGLD is comparable to SAM in terms of $\lambda_{\max}$ and even lower in terms of $\text{tr}(H(\theta))$. This result is achieved at roughly half the computational cost of SAM. These results empirically validate our theoretical analysis, confirming that the proposed algorithm effectively promotes convergence to flatter minima.

5 Related Work and Discussions

We review the most relevant literature on SAM, RWP, Hessian-based optimization, and SGLD.

Flat Minima and Generalization. Empirical studies (Keskar et al., 2017; Jastrzkebski et al., 2017; Jiang et al., 2020) and theoretical analyses (Dziugaite & Roy, 2017; Neyshabur et al., 2017) consistently show that flatter minima are strongly correlated with better generalization in deep neural networks. However, elucidating precise notions of sharpness and their relationship to generalization remains an open and active area of research (Andriushchenko & Flammarion, 2022; Ding et al., 2024; Wen et al., 2023; Tahmasebi et al., 2024).

SAM and RWP. The success of SAM (Foret et al., 2021) has produced a wide range of follow-up work to improve its efficiency, effectiveness, and applicability. Extensions include algorithmic improvements to approximate the inner maximization more efficiently (Liu et al., 2022a; Du et al., 2022a; Kwon et al., 2021; Xie et al., 2024; Li et al., 2024b; Chen et al., 2024; Kang et al., 2025). Beyond these, several Hessian-based regularization approaches have explored flatness from a different angle. For example, Zhang et al. (2024) propose Noise-Stability Optimization, and Li et al.

(2024a) studies random weight perturbation with explicit Hessian penalties. Both works focus on PAC-Bayes generalization bounds and local convergence to stationary points, providing algorithm-agnostic guarantees about the perturbed loss rather than the training dynamics of a specific optimizer. By contrast, we show that the invariant measure of fSGLD yields global, non-asymptotic convergence guarantees and an explicit link between random weight perturbation and Hessian-trace regularization. Lastly, the concept of using noise for regularization was formalized through the framework of randomized smoothing (Duchi et al., 2012), and our work makes this connection explicit for Langevin dynamics, differing fundamentally from explicit Hessian-penalty methods that rely on costly approximations (Sankar et al., 2021).

SGLD and its Convergence Rate. Following the seminal works of Welling & Teh (2011); Raginsky et al. (2017), numerous variants of SGLD have been developed to improve its practical performance, such as variance reduction techniques (Kinoshita & Suzuki, 2022; Dubey et al., 2016; Huang & Becker, 2021), preconditioned SGLD (Li et al., 2016), replica exchange SGLD (Dong & Tong, 2021; Deng et al., 2020a). A parallel line of research has focused on its theoretical properties, particularly its non-asymptotic convergence rate. Early results (Raginsky et al., 2017; Xu et al., 2018) showed convergence in the Wasserstein-2 distance at a rate dependent on the number of iterations. More recently, the state-of-the art analyses have established convergence rates of $O(\lambda^{1/2})$ in Wasserstein-1 and $O(\lambda^{1/4})$ in Wasserstein-2 distance (Zhang et al., 2023). Our convergence rates are consistent with these best-known results. However, a crucial distinction is that prior work proves convergence to the minimizers of the original objective u, whereas our guarantees are for convergence to global flat minima.

6 Conclusion and Limitations

In this work, we introduced Flatness-Aware Stochastic Gradient Langevin Dynamics (fSGLD), a novel algorithm that synthesizes randomized smoothing with Langevin dynamics to efficiently target flat minima. By evaluating the gradient at parameters perturbed by Gaussian noise, a technique known as Random Weight Perturbation (RWP), fSGLD optimizes a surrogate objective that provably incorporates Hessian trace information without explicit computation.

Our main theoretical contribution is a rigorous non-asymptotic analysis of this process. We establish convergence guarantees in Wasserstein distance and provide the explicit excess risk bound for this class of flatness-aware optimizers. Crucially, our theory shows that the desired regularization effect emerges from a precise coupling of the noise scale σ and the inverse temperature β.

Empirically, fSGLD demonstrates superior or competitive performance against strong baselines, including SAM, on challenging noisy-label and fine-tuning benchmarks. These gains are achieved at a computational cost comparable to standard SGD, roughly half that of SAM. Our analysis of the Hessian spectrum further confirms that fSGLD converges to significantly flatter minima, providing a direct validation of its mechanism. Ultimately, our work provides one of the provable links between an efficient algorithmic design (RWP within SGLD) and quantifiable generalization benefits, bridging the gap between heuristic flatness-seeking methods and rigorous convergence theory.

Limitations and Future Directions. Applying fSGLD to diffusion-based generative models is a particularly promising direction; investigating whether its bias towards flatter regions of the loss landscape can lead to more diverse or higher-quality samples is a compelling open question. On the theoretical side, we leave for future work the extension of our analysis to the case where u is semiconvex (i.e., its gradient is one-sided Lipschitz), rather than satisfying Assumption 2.

References

Kwangjun Ahn, Ali Jadbabaie, and Suvrit Sra. How to escape sharp minima with random perturbations. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International conference on machine learning*, pp. 639–668. PMLR, 2022.

Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.

Christina Baek, J Zico Kolter, and Aditi Raghunathan. Why is SAM robust to label noise? In *The Twelfth International Conference on Learning Representations*, 2024.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: bouncing across ravines and drifting towards wide minima. *J. Mach. Learn. Res.*, 24 (1), 2023. ISSN 1532-4435.

Devansh Bisla, Jing Wang, and Anna Choromanska. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics*, pp. 8299–8339. PMLR, 2022.

François Bolley and Cédric Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

Huy N Chau, Chaman Kumar, Miklós Rásonyi, and Sotirios Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23:217–244, 2019.

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017.

Simiao Chen, Xiaoge Deng, Dongpo Xu, Tao Sun, and Dongsheng Li. Decentralized stochastic sharpness-aware minimization algorithm. *Neural Networks*, 176:106325, 2024. ISSN 0893-6080.

Jiarui Chu and Ludovic Tangpi. Nonasymptotic estimation of risk measures using stochastic gradient langevin dynamics. *SIAM Journal on Financial Mathematics*, 15(2):503–536, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Wei Deng, Qi Feng, Liyao Gao, Faming Liang, and Guang Lin. Non-convex learning via replica exchange stochastic gradient mcmc. In *International Conference on Machine Learning*, pp. 2474–2483. PMLR, 2020a.

Wei Deng, Guang Lin, and Faming Liang. A contour stochastic gradient Langevin dynamics algorithm for simulations of multi-modal distributions. *Advances in neural information processing systems*, 33:15725–15736, 2020b.

Wei Deng, Siqi Liang, Botao Hao, Guang Lin, and Faming Liang. Interacting contour stochastic gradient Langevin dynamics. In *International Conference on Learning Representations*, 2022.

Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2), 2024.

Jing Dong and Xin T Tong. Replica exchange for non-convex optimization. *Journal of Machine Learning Research*, 22(173):1–59, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022a.

Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23439–23451, 2022b.

Kumar Avinava Dubey, Sashank J. Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 884–893, 2017.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Quantitative harris-type theorems for diffusions and mckean–vlasov processes. *Transactions of the American Mathematical Society*, 371 (10):7135–7173, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Futoshi Futami and Masahiro Fujisawa. Time-independent information-theoretic generalization bounds for sgld. *Advances in Neural Information Processing Systems*, 36:8173–8185, 2023.

Khashayar Gatmiry, Zhiyuan Li, Sashank J. Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. In *Forty-first International Conference on Machine Learning*, 2024.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Jing Huang, Yin Dai, Yulin Jiao, Lican Kang, and Xiliang Lu. Nonasymptotic convergence analysis for the tamed unadjusted stochastic langevin algorithm. 2025.

Zhishen Huang and Stephen Becker. Stochastic gradient langevin dynamics with variance reduction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Stanisław Jastrzkebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

Helei Kang, Yiming Jiang, Jinlan Liu, and Dongpo Xu. Sharpness-aware minimization method with momentum acceleration for deep neural networks. *Knowledge-Based Systems*, 326:113967, 2025. ISSN 0950-7051.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Pham Duy Khanh, Hoang-Chau Luong, Boris S. Mordukhovich, and Dat Ba Tran. Fundamental convergence analysis of sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13149–13182, 2024.

Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization. In *Advances in Neural Information Processing Systems*, 2022.

Sangamesh Kodge. MiniWebVision, February 2024. URL https://github.com/sangamesh-kodge/Mini-WebVision.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914, 2021.

Bolian Li and Ruqi Zhang. Entropy-MCMC: Sampling from flat basins with ease. In *The Twelfth International Conference on Learning Representations*, 2024.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Junnan Li, Richard Socher, and Steven C.H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 316–325, June 2022.

Tao Li, Qinghua Tao, Weihao Yan, Yingwen Wu, Zehao Lei, Kun Fang, Mingzhen He, and Xiaolin Huang. Revisiting random weight perturbation for efficiently improving generalization. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5631–5640, 2024b.

Tian Li, Tianyi Zhou, and Jeff Bilmes. Tilted sharpness-aware minimization. In *Forty-second International Conference on Machine Learning*, 2025.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, Jesse Berent, Abhinav Gupta, Rahul Sukthankar, and Luc Van Gool. Webvision challenge: Visual learning and understanding with web data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2243–2251, 2017.

Lin Lin, Yousef Saad, and Chao Yang. Approximating spectral densities of large matrices. *SIAM review*, 58(1):34–65, 2016.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12360–12370, 2022a.

Yong Liu, Siqi Mai, Minhao Cheng, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Random sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24543–24556, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Haocheng Luo, Tuan Truong, Tung Pham, Mehrtash Harandi, Dinh Phung, and Trung Le. Explicit eigenvalue regularization improves sharpness-aware minimization. *Advances in Neural Information Processing Systems*, 37:4424–4453, 2024.

Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *Annals of Applied Probability*, 30(4):1534–1581, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *The Thirteenth International Conference on Learning Representations*, 2025.

Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6606–6615, 2021.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1674–1703. PMLR, 2017.

Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9481–9488, 2021.

Dongkuk Si and Chulhee Yun. Practical sharpness-aware minimization cannot converge all the way to optima. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26190–26228, 2023.

Behrooz Tahmasebi, Ashkan Soleymani, Dara Bahri, Stefanie Jegelka, and Patrick Jaillet. A universal class of sharpness-aware minimization algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

Chengli Tan, Jiangshe Zhang, Junmin Liu, Yicheng Wang, and Yunda Hao. Stabilizing sharpness-aware minimization through a simple renormalization strategy. *Journal of Machine Learning Research*, 26(68):1–35, 2025.

Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of tr(f(a)) via stochastic lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, 2017.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Zheng Wei, Xingjun Zhang, and Zhendong Tan. Unifying and revisiting sharpness-aware minimization with noise-injected micro-batch scheduler for efficiency improvement. *Neural Networks*, 185:107205, 2025. ISSN 0893-6080.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning*, pp. 681–688, 2011.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wanyun Xie, Fabian Latorre, Kimon Antonakopoulos, Thomas Pethick, and Volkan Cevher. Improving SAM requires rethinking its optimization formulation. In *Forty-first International Conference on Machine Learning*, 2024.

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.

Runsheng Yu, Youzhi Zhang, and James Kwok. Improving sharpness-aware minimization by lookahead. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

Hongyang R. Zhang, Dongyue Li, and Haotian Ju. Noise stability optimization for finding flat minima: A hessian-based regularization approach. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25, 2023.

A Relationship between $\pi^{\star}_{\beta,\sigma}$ and $\pi^{\mathrm{fSGLD}}_{\beta}$

We derive the relationship between the target measure $\pi^{\star}_{\beta,\sigma}$ and the invariant measure $\pi^{\mathrm{fSGLD}}_{\beta}$ of the fSGLD algorithm, which will be used to prove Proposition 3.1, Theorem 3.2, and Corollary 3.3. By Taylor's theorem, we obtain

$$u(\theta+\epsilon) = u(\theta) + \nabla u(\theta)^T\epsilon + \frac{1}{2}\epsilon^T H(\theta)\epsilon + \mathcal{R}(\theta,\epsilon), \tag{15}$$

where $\mathcal{R}(\theta,\epsilon)$ denotes the remainder term. Taking the expectation over $\epsilon \sim \mathcal{N}(0,\sigma^2 I_d)$ in 15, we have

$$\begin{aligned} g_\epsilon(\theta) &= u(\theta) + \frac{1}{2}\mathbb{E}[\epsilon^T H(\theta)\epsilon] + \mathbb{E}[\mathcal{R}(\theta,\epsilon)] \\ &= u(\theta) + \frac{1}{2}\mathrm{tr}\left(H(\theta)\cdot\mathbb{E}[\epsilon^T\epsilon]\right) + \mathbb{E}[\mathcal{R}(\theta,\epsilon)] \\ &= v(\theta) + \mathbb{E}[\mathcal{R}(\theta,\epsilon)], \end{aligned} \tag{16}$$

where

$$v(\theta) = u(\theta) + \frac{\sigma^2}{2}\mathrm{tr}\left(H(\theta)\right),$$

and

$$\begin{aligned} \mathbb{E}[\mathcal{R}(\theta,\epsilon)] &= \frac{1}{6}\sum_{i,j,k=1}^{d}\frac{\partial^3 u}{\partial\theta_i\partial\theta_j\partial\theta_k}(\theta)\,\mathbb{E}[\epsilon_i\epsilon_j\epsilon_k] + \frac{1}{24}\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta)\,\mathbb{E}[\epsilon_i\epsilon_j\epsilon_k\epsilon_l] \\ &\quad + \sum_{j=5}^{\infty}\frac{1}{j!}\sum_{i_1,i_2,\ldots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_j}}(\theta)\,\mathbb{E}[\epsilon_{i_1}\epsilon_{i_2}\ldots\epsilon_{i_j}] \\ &= \frac{1}{24}\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta)\,\sigma^4(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}) \\ &\quad + \sum_{j=6}^{\infty}\frac{1}{j!}\sum_{i_1,i_2,\ldots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_j}}(\theta)\,\mathbb{E}[\epsilon_{i_1}\epsilon_{i_2}\ldots\epsilon_{i_j}], \end{aligned} \tag{17}$$

where δ_{ij} denotes the Kronecker delta. Since $\sigma \in (0,1)$ by Assumption 1, we have $\mathbb{E}[\mathcal{R}(\theta,\epsilon)] = O(\sigma^4)$. Let the normalization constant of $\pi^{\mathrm{fSGLD}}_{\beta}$ be given by

$$Z_\beta := \int_{\mathbb{R}^d} e^{-\beta g_\epsilon(\theta)}\,\mathrm{d}\theta, \tag{18}$$

and let the normalization constant of $\pi^{\star}_{\beta,\sigma}$ be given by

$$Z_{\beta,\sigma} := \int_{\mathbb{R}^d} e^{-\beta v(\theta)}\,\mathrm{d}\theta. \tag{19}$$

Using 16, 18, and 19, we obtain

$$\begin{aligned} \pi^{\mathrm{fSGLD}}_{\beta}(\mathrm{d}\theta) &= Z_\beta^{-1}\exp(-\beta g_\epsilon(\theta))\,\mathrm{d}\theta \\ &= Z_\beta^{-1}Z_{\beta,\sigma}\exp(-\beta\,\mathbb{E}[\mathcal{R}(\theta,\epsilon)])\,\pi^{\star}_{\beta,\sigma}(\mathrm{d}\theta). \end{aligned} \tag{20}$$

B Additional results for Section 3.1

This section collects several technical remarks and direct consequences of the assumptions presented in Section 3.1.

Remark B.1. *By Assumption 1 and 2, the gradient $h(\theta) = \mathbb{E}[\nabla_\theta U(\theta, X)]$ for all $\theta \in \mathbb{R}^d$, is well-defined. In addition, one obtains for all θ, $\theta' \in \mathbb{R}^d$,*

$$|h(\theta) - h(\theta')| \le L_1\mathbb{E}[\varphi(X_0)]|\theta - \theta'|.$$

As a consequence of Assumption 2, one obtains, for fixed $\widetilde{\epsilon} \in \mathbb{R}^d$,

$$|\nabla_\theta U(\theta+\widetilde{\epsilon},x) - \nabla_{\theta'} U(\theta'+\widetilde{\epsilon},x)| \le L_1\varphi(x)|\theta-\theta'|,$$
$$|\nabla_\theta U(\theta+\widetilde{\epsilon},x) - \nabla_\theta U(\theta+\widetilde{\epsilon},x')| \le L_2(\varphi(x)+\varphi(x'))(1+|\theta+\widetilde{\epsilon}|)|x-x'|.$$

Also, Assumption 2 implies

$$|\nabla_\theta U(\theta+\widetilde{\epsilon},x)| \le L_1\varphi(x)|\theta| + L_2\bar{\varphi}(x)(1+|\widetilde{\epsilon}|) + \widetilde{G}(\widetilde{\epsilon}),$$

where $\bar{\varphi}(x) := (\varphi(x)+\varphi(0))|x|$, *and* $\widetilde{G}(\widetilde{\epsilon}) := |\nabla_{\theta'} U(\widetilde{\epsilon},0)|$. *Moreover,*

$$|\nabla g_\epsilon(\theta)| \le L_1\mathbb{E}[\varphi(X_0)]|\theta| + L_2\mathbb{E}[\bar{\varphi}(X_0)](1+\mathbb{E}[|\epsilon|]) + \mathbb{E}[\widetilde{G}(\epsilon)].$$

Remark B.2. *By Assumption 1 and 3, one obtains a dissipativity condition of* h, *i.e., for any* $\theta \in \mathbb{R}^d$, $\langle \nabla h(\theta), \theta\rangle \ge \bar{a}|\theta|^2 - \bar{b}$. *Let* $\zeta \in (0, \bar{a}L_1^{-2}(\mathbb{E}[\varphi^2(X_0)])^{-1})$. *As a consequence of Assumptions 2 and 3, one obtains, for any* $\theta \in \mathbb{R}^d$

$$\langle \nabla g_\epsilon(\theta), \theta\rangle \ge a|\theta|^2 - b, \tag{21}$$

where

$$\begin{aligned} a &:= \bar{a} - \zeta L_1^2\mathbb{E}[\varphi^2(X_0)] > 0, \\ b &:= (2\zeta)^{-1}\sigma^2 d + 4\zeta L_2^2\mathbb{E}[\bar{\varphi}^2(X_0)](1+\sigma^2 d) + 2\zeta\mathbb{E}[\widetilde{G}^2(\epsilon)] + \bar{b} > 0, \end{aligned} \tag{22}$$

and $\widetilde{G}$ *and* $\bar{\varphi}$ *are given in Remark B.1.*

Proof of Remark B.2. Using Assumption 3 and Remark B.1, and Young's inequality, one obtains, for fixed $\widetilde{\epsilon} \in \mathbb{R}^d$

$$\begin{aligned} \langle \nabla_\theta U(\theta+\widetilde{\epsilon},x),\theta\rangle &= \langle \nabla_\theta U(\theta+\widetilde{\epsilon},x),\theta+\widetilde{\epsilon}\rangle - \langle \nabla_\theta U(\theta+\widetilde{\epsilon},x),\widetilde{\epsilon}\rangle \\ &\ge \langle \theta+\widetilde{\epsilon}, A(x)\theta+\widetilde{\epsilon}\rangle - \hat{b}(x) - \zeta 2^{-1}|\nabla_\theta U(\theta+\widetilde{\epsilon},x)|^2 - (2\zeta)^{-1}|\widetilde{\epsilon}|^2 \\ &\ge \langle \theta, (A(x)-\zeta L_1^2\varphi^2(x))\theta\rangle + \langle \theta, A(x)\widetilde{\epsilon}\rangle + \langle \widetilde{\epsilon}, A(x)\theta\rangle + \langle \widetilde{\epsilon}, A(x)\widetilde{\epsilon}\rangle \\ &\quad - 4\zeta L_2^2\bar{\varphi}^2(x)(1+|\widetilde{\epsilon}|^2) - 2\zeta\widetilde{G}^2(\widetilde{\epsilon}) - \hat{b}(x) - (2\zeta)^{-1}|\widetilde{\epsilon}|^2. \end{aligned} \tag{23}$$

Therefore,

$$\begin{aligned} \nabla g_\epsilon(\theta) &= \mathbb{E}[\mathbb{E}_X[\nabla_\theta U(\theta+\epsilon,X)]] \\ &\ge (\bar{a} - \zeta L_1^2\mathbb{E}[\varphi^2(X_0)])|\theta|^2 + (\bar{a} - (2\zeta)^{-1})\sigma^2 d - 4\zeta L_2^2\mathbb{E}[\bar{\varphi}^2(X_0)](1+\sigma^2 d) \\ &\quad - 2\zeta\mathbb{E}[\widetilde{G}^2(\epsilon)] - \bar{b} \\ &\ge a|\theta|^2 - b, \end{aligned}$$

where a and b are defined in 22.

□

Lemma B.3. *Let Assumption 2 and 3 hold. Then* $\pi^\star_{\beta,\sigma}$ *has finite second moments.*

Proof of Lemma B.3. As a consequence of Assumption 2, $\nabla v(\theta)$ is Lipschitz continuous. Let $\bar{\zeta} \in (0, 4\bar{a}\sigma^{-2})$. Using Assumption 2, Assumption 3, and Young's inequality, one obtains

$$\begin{aligned} \langle \nabla v(\theta),\theta\rangle &= \langle \nabla u(\theta),\theta\rangle + \frac{\sigma^2}{2}\langle \nabla\left(\mathrm{tr}\left(H(\theta)\right)\right),\theta\rangle \\ &\ge \left(\bar{a} - \frac{\bar{\zeta}\sigma^2}{4}\right)|\theta|^2 - \bar{b} - \frac{\sigma^2}{4\bar{\zeta}}|\nabla\left(\mathrm{tr}\left(H(\theta)\right)\right)|^2, \end{aligned}$$

which implies that $\nabla v(\theta)$ is dissipative. Therefore, $\pi^\star_{\beta,\sigma}$ has finite second moment. □

Remark B.4. *Our theoretical results are established under the standard i.i.d. sampling model (Assumption 1) where a sample* X_k *is drawn at each iteration. In contrast, our experiments in Section 4 follow the common practice in deep learning, training over multiple epochs using uniform sampling without replacement. These two sampling regimes are not identical because sampling*

without replacement is not strictly i.i.d. However, it is well established in stochastic optimization that both regimes induce gradient noise with very similar statistical properties, particularly in large-dataset settings. Sampling with replacement would align exactly with our theoretical setup. In application domains where sample generators are available, such as financial modeling Chu & Tangpi (2024), the online sampling arises naturally. Since our empirical evaluation focuses on settings that are standard in the deep learning community, we adopt the widely used multi-pass training protocol for benchmarking.

Remark B.5. *Controlling the remainder term $\mathbb{E}[\mathcal{R}(\theta,\epsilon)]$ in 17 could in principle require very strong smoothness assumptions such as globally bounded fourth-order derivatives to ensure uniform control of higher-order terms. These are not standard in SGLD analyses, and our approach does not impose any such extra conditions, Instead, by leveraging only the dissipativity conditon (Assumption 3) together with local Lipschitz continuity (Assumption 2), we establish all convergence results without any global C^4 boundedness or similar strong regularity. This distinction highlights a key theoretical contribution of our work: rigorous non-asymptotic analysis for nonconvex high-dimensional objectives under significantly weaker and more realistic assumptions.*

C Overview of the non-asymptotic Wasserstein analysis and error bound for the expected excess risk

In this section, we derive the results introduced in Sections 3.2 and 3.3. We begin by presenting the framework behind these two sections.

The 'data' process $(X_k)_{k\in\mathbb{N}}$ in 7 is adapted to a given filtration $(\mathcal{X}_k)_{k\in\mathbb{N}}$ representing the flow of past information, and we denote the sigma-algebra of $\cup_{k\in\mathbb{N}}\mathcal{X}_k$ by $\mathcal{X}_\infty$. In addition, we assume that θ_0 , $\mathcal{X}_\infty$, $(\epsilon_k)_{k\in\mathbb{N}}$, and $(\xi_k)_{k\in\mathbb{N}}$ are all independent among themselves.

We define

$$\lambda_{\max} := \min\left\{\frac{\min\{a, a^{\frac{1}{3}}\}}{16(1+L_1)^2(\mathbb{E}[(1+\varphi(X_0))^4])^{1/2}}, \frac{1}{a}\right\}, \tag{24}$$

where L_1, φ and a are defined in Assumptions 2 and Remark B.2, respectively.

C.1 Auxiliary processes

We start by defining the process $(Z_t^{\text{fSGLD}})_{t\in\mathbb{R}_+}$ as the solution of the *flatness* Langevin SDE

$$\begin{aligned} Z_0^{\text{fSGLD}} &:= \theta_0 \in \mathbb{R}^d, \\ \mathrm{d}Z_t^{\text{fSGLD}} &:= -\nabla g_\epsilon(Z_t^{\text{fSGLD}})\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}B_t, \end{aligned} \tag{25}$$

where B_t is a standard d-dimensional Brownian motion. Denote by $(\mathcal{F}_t)_{t\geq 0}$ the natural filtration of $(B_t)_{t\geq 0}$ and by Σ_{θ_0} the sigma-algebra generated by θ_0, and we assume that $(\mathcal{F}_t)_{t\geq 0}$ is independent of $\mathcal{X}_\infty \vee \Sigma_{\theta_0}$. Furthermore, denote by $\mathcal{F}_\infty$ the sigma-algebra of $\bigcup_{t\geq 0}\mathcal{F}_t$.

Remark C.1. *By Remark B.1, SDE 25 has a unique solution adapted to $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$.*

To facilitate the convergence analysis, we introduce the time-rescaled version of the process 25. For each $\lambda > 0$, $Z_t^{\lambda,\text{fSGLD}} := Z_{\lambda t}^{\text{fSGLD}}$, $t \in \mathbb{R}_+$, and let $\widetilde{B}_t^\lambda := B_{\lambda t}/\sqrt{\lambda}$, $t \geq 0$. We observe that $(\widetilde{B}_t)_{t\geq 0}$ is a Brownian motion and

$$\begin{aligned} Z_0^{\lambda,\text{fSGLD}} &:= \theta_0 \\ \mathrm{d}Z_t^{\lambda,\text{fSGLD}} &:= -\lambda\nabla g_\epsilon(Z_t^{\lambda,\text{fSGLD}})\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\,\mathrm{d}\widetilde{B}_t^\lambda. \end{aligned} \tag{26}$$

The natural filtration of $(\widetilde{B}_t)_{t\geq 0}$ is denoted by $(\mathcal{F}_t^\lambda)_{t\geq 0}$ with $\mathcal{F}_t^\lambda := \mathcal{F}_{\lambda t}$, $t \in \mathbb{R}_+$. For a positive real number a, we denote its integer part by $\lfloor a \rfloor$. Then, we define $(\bar{\theta}_t^{\text{fSGLD}})_{t\in\mathbb{R}_+}$, the continuous-time interpolation of fSGLD 7, as

$$\begin{aligned} \bar{\theta}_0^{\text{fSGLD}} &:= \theta_0, \\ \mathrm{d}\bar{\theta}_t^{\text{fSGLD}} &:= -\lambda\nabla_\theta U(\bar{\theta}_{\lfloor t\rfloor}^{\text{fSGLD}} + \epsilon_{\lceil t\rceil}, X_{\lceil t\rceil})\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\widetilde{B}_t. \end{aligned} \tag{27}$$

At grid-points, we note that the law of the interpolated process is the same as the law of the fSGLD algorithm 7, i.e. $\mathcal{L}(\bar{\theta}_k^{\text{fSGLD}}) = \mathcal{L}(\theta_k^{\text{fSGLD}})$, for each $k \in \mathbb{N}$. Moreover, we introduce the following continuous-time process $(\Phi_t^{s,u,\lambda,\text{fSGLD}})_{t\geq s}$, which is beneficial for our analysis, and define it as the solution of the following SDE

$$\begin{aligned}\Phi_s^{s,u,\lambda,\text{fSGLD}} &:= v \in \mathbb{R}^d \\ \mathrm{d}\Phi_t^{s,u,\lambda,\text{fSGLD}} &:= -\lambda \nabla g_\epsilon(\Phi_t^{s,u,\lambda,\text{fSGLD}})\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\,\mathrm{d}\widetilde{B}_t^\lambda.\end{aligned}$$

Definition C.2. *Fix $k \in \mathbb{N}$. For any $t \geq kT$, define $\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} := \Phi_t^{kT,\bar{\theta}_{kT}^{\text{fSGLD}},\lambda,\text{fSGLD}}$, where $T := \lfloor 1/\lambda \rfloor$.*

In other words, $\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}$ in Definition C.2 is a process started from the value of the continuous-time interpolation fSGLD process 27 at time kT and run until time $t \geq kT$ with the continuous-time flatness Langevin dynamics.

C.2 Proofs of the results in Sections 3.2 and 3.3

To prove Proposition 3.1, Theorem 3.2, and Corollary 3.3, we will use the following results in Corollary C.3 and Lemma C.4 below.

Recall that for any μ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, then Kullbak-Leibler divergence (or relative entropy) between μ and ν is defined as

$$\text{KL}(\mu||\nu) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu, & \text{if } \mu \ll \nu, \\ \infty, & \text{otherwise.} \end{cases} \tag{28}$$

Corollary C.3. *(Bolley & Villani, 2005, Corollary 2.3) For any two Borel probability measures μ and ν with finite second moments, one obtains*

$$W_2(\mu,\nu) \leq C_\nu \left[\sqrt{KL(\mu||\nu)} + \left(\frac{KL(\mu||\nu)}{2}\right)^{1/4}\right],$$

where

$$C_\nu := 2 \inf_{\widetilde{\kappa}>0} \left(\frac{1}{\widetilde{\kappa}}\left(\frac{3}{2} + \log \int_{\mathbb{R}^d} e^{\widetilde{\kappa}|\theta|^2}\,\nu(\mathrm{d}\theta)\right)\right)^{1/2}. \tag{29}$$

Lemma C.4. *Let Assumption 3 hold. Then, the following set*

$$A := \left\{\theta \in \mathbb{R}^d : |\theta| \leq \sqrt{\frac{b}{a}}\right\}, \tag{30}$$

contains all the minimizers of $u(\theta)$, $v(\theta)$, and $g_\epsilon(\theta)$, where a and b are given in 22.

Proof of Lemma C.4. Let $\theta_{g_\epsilon}^\star$, $\theta_u^\star$, $\theta_v^\star$ be a minimizer of $g_\epsilon(\theta)$, $u(\theta)$, and $v(\theta)$, respectively. By Assumption 3, we have

$$0 = \langle \nabla v(\theta_v^*), \theta_v^\star\rangle = \langle \nabla v(\theta_u^*), \theta_u^\star\rangle = \langle \nabla u(\theta_u^\star), \theta_u^\star\rangle \geq \overline{a}|\theta_u^\star|^2 - \overline{b}, \tag{31}$$

which implies

$$|\theta_u^\star| = |\theta_v^\star| \leq \sqrt{\frac{\overline{b}}{\overline{a}}} \leq \sqrt{\frac{b}{a}}.$$

Due to Remark B.2, we have

$$0 = \langle \nabla g_\epsilon(\theta_{g_\epsilon}^\star), \theta_{g_\epsilon}^\star\rangle \geq a|\theta_{g_\epsilon}^\star|^2 - b, \tag{32}$$

which implies

$$|\theta_{g_\epsilon}^\star| \leq \sqrt{\frac{b}{a}}.$$

□

Proof of Proposition 3.1. Using 20 with 17, we have

$$\begin{aligned}\mathrm{KL}(\pi_\beta^{\mathrm{fSGLD}}||\pi^\star_{\beta,\sigma}) &= \int_{\mathbb{R}^d} \log\left(\frac{\pi_\beta^{\mathrm{fSGLD}}(\mathrm{d}\theta)}{\pi^\star_{\beta,\sigma}(\mathrm{d}\theta)}\right)\, \pi_\beta^{\mathrm{fSGLD}}(\mathrm{d}\theta) \\ &= \int_{\mathbb{R}^d} \log\left(Z_\beta^{-1} Z_{\beta,\sigma}\exp(-\beta\,\mathbb{E}[\mathcal{R}(\theta,\epsilon)])\right)\pi_\beta^{\mathrm{fSGLD}}(\mathrm{d}\theta) \\ &= \log\left(\frac{Z_{\beta,\sigma}}{Z_\beta}\right) - \beta\int_{\mathbb{R}^d}\mathbb{E}[\mathcal{R}(\theta,\epsilon)]\,\pi_\beta^{\mathrm{fSGLD}}(\mathrm{d}\theta).\end{aligned} \tag{33}$$

We focus on the first term on the right-hand side of 33. We denote the complementary set of A in Lemma C.4 by A^c. Using 18 and 19, one obtains

$$\begin{aligned}\log\left(\frac{Z_{\beta,\sigma}}{Z_\beta}\right) &= \log\left(\frac{\int_A e^{-\beta v(\theta)}d\theta + \int_{A^c} e^{-\beta v(\theta)}d\theta}{\int_A e^{-\beta g_\epsilon(\theta)}d\theta + \int_{A^c} e^{-\beta g_\epsilon(\theta)}d\theta}\right) \\ &= \log\left(\frac{\frac{\int_A e^{-\beta v(\theta)}d\theta}{\int_A e^{-\beta g_\epsilon(\theta)}d\theta} + \frac{\int_{A^c} e^{-\beta v(\theta)}d\theta}{\int_A e^{-\beta g_\epsilon(\theta)}d\theta}}{1+\frac{\int_{A^c} e^{-\beta g_\epsilon(\theta)}d\theta}{\int_A e^{-\beta g_\epsilon(\theta)}d\theta}}\right).\end{aligned} \tag{34}$$

We provide a bound on the first term of the numerator in 34, i.e., $\frac{\int_A e^{-\beta v(\theta)}\mathrm{d}\theta}{\int_A e^{-\beta g_\epsilon(\theta))}\mathrm{d}\theta}$. By 17 and the extreme value theorem, there exists a constant $C_A > 0$:

$$|g_\epsilon(\theta) - v(\theta)| \le C_A\sigma^4, \quad \forall\,\theta\in A, \tag{35}$$

where

$$C_A = \max_{\theta\in A}\left(\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta) + \sum_{j=6}^{\infty}\sum_{i_1,i_2,\ldots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_j}}(\theta)\right). \tag{36}$$

This leads to

$$e^{-C_A\beta\sigma^4}\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta \le \int_A e^{-\beta v(\theta)}\mathrm{d}\theta \le e^{C_A\beta\sigma^4}\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta,$$

which implies

$$e^{-C_A\beta\sigma^4} \le \frac{\int_A e^{-\beta v(\theta)}\mathrm{d}\theta}{\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta} \le e^{C_A\beta\sigma^4}. \tag{37}$$

We provide a bound on the second term of the numerator on the right-hand side of 34, i.e., $\frac{\int_{A^c} e^{-\beta v(\theta)}\mathrm{d}\theta}{\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta}$. We note that, for any $\theta\in A^c$, there exists $\delta_v > 0$ such that

$$v(\theta) > v(\theta_v^*) + \delta_v, \quad \text{for } \theta\in A^c. \tag{38}$$

For $0 < \beta_0 < \beta$, we have

$$-(\beta-\beta_0)v(\theta) < -(\beta-\beta_0)(v(\theta_v^*)+\delta_v),$$

which implies

$$\int_{A^c} e^{-\beta v(\theta)}\mathrm{d}\theta \le e^{-(\beta-\beta_0)(v(\theta_v^*)+\delta_v)}\int_{A^c} e^{-\beta_0 v(\theta)}\mathrm{d}\theta. \tag{39}$$

By 17 and the extreme value theorem, we obtain

$$\exp(-\beta\,\mathbb{E}[\mathcal{R}(\theta,\epsilon)]) \ge \exp(-\beta C_A\sigma^4), \qquad \theta\in A. \tag{40}$$

Using 40, we have

$$\begin{aligned}\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta &\ge \exp(-\beta C_A\sigma^4)\int_A e^{-\beta v(\theta)}\,\mathrm{d}\theta \\ &= \exp(-\beta C_A\sigma^4)\left(\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}\left(\frac{b}{a}\right)^{\frac{d}{2}} + \int_A\sum_{i=1}^{\infty}\frac{(-\beta v(\theta))^i}{i!}\mathrm{d}\theta\right) \\ &\ge \frac{(1+\bar{c}_A(\beta))}{\exp(\beta C_A\sigma^4)\Gamma(\frac{d}{2}+1)},\end{aligned} \tag{41}$$

where $\bar{c}_A(\beta) := \min_{\theta\in A}\sum_{i=1}^{\infty}\frac{(-\beta v(\theta))^i}{i!}$. Combining 39 and 41, we obtain

$$\frac{\int_{A^c} e^{-\beta v(\theta)}\mathrm{d}\theta}{\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta} \leq \frac{\exp(\beta C_A\sigma^4)\,\Gamma(\frac{d}{2}+1)\int_{A^c} e^{-\beta_0 v(\theta)}\mathrm{d}\theta}{e^{(\beta-\beta_0)(v(\theta_v^*)+\delta_v)}\,(1+\bar{c}_A(\beta))}. \tag{42}$$

We provide a bound on the ratio in the denominator on the right-hand side of 34. By Taylor's theorem,

$$g_\epsilon(\theta) = g_\epsilon(\theta^\star_{g_\epsilon}) + \frac{1}{2}(\theta-\theta^\star_{g_\epsilon})^T\nabla^2 g_\epsilon(\theta^\star_{g_\epsilon})(\theta-\theta^\star_{g_\epsilon}) + R_2(\theta), \tag{43}$$

where $R_2(\theta)$ is the remainder term accounting for the residual error above second order. By the extreme value theorem, there exists $\overline{m}$, $\overline{M} > 0$ such that

$$\overline{m} \leq e^{-\beta R_2(\theta)} \leq \overline{M}, \quad \forall\,\theta\in A. \tag{44}$$

Using 43 and 44, one obtains

$$\begin{aligned}\int_A e^{-\beta g_\epsilon(\theta)}d\theta &= e^{-\beta g_\epsilon(\theta^\star_{g_\epsilon})}\int_A e^{-\frac{\beta}{2}(\theta-\theta^\star_{g_\epsilon})^T\nabla^2 g_\epsilon(\theta^\star_{g_\epsilon})(\theta-\theta^\star_{g_\epsilon})-\beta R_2(\theta)}d\theta \\ &\leq \overline{M}e^{-\beta g_\epsilon(\theta^\star_{g_\epsilon})}\left(\frac{2\pi}{\beta}\right)^{d/2}\frac{1}{\sqrt{\det\nabla^2 g_\epsilon(\theta^\star_{g_\epsilon})}}.\end{aligned} \tag{45}$$

Using 45, it follows

$$1+\frac{\int_{A^c} e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta}{\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta} \geq 1+\frac{\int_{A^c} e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta}{\overline{M}e^{-\beta g_\epsilon(\theta^\star_{g_\epsilon})}\frac{1}{\sqrt{\det\nabla^2 g_\epsilon(\theta^\star_{g_\epsilon})}}} \geq 1.$$

Thus,

$$\frac{1}{1+\frac{\int_{A^c} e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta}{\int_A e^{-\beta g_\epsilon(\theta)}\mathrm{d}\theta}} \leq \frac{1}{1+\frac{\int_{A^c} e^{-\beta g_\epsilon(\theta))}\mathrm{d}\theta}{\overline{M}e^{-\beta g_\epsilon(\theta^\star_{g_\epsilon})}\frac{1}{\sqrt{\det\nabla^2 g_\epsilon(\theta^\star_{g_\epsilon})}}}} \leq 1. \tag{46}$$

Using 46, 42, and 37 in 34 yields

$$\log\left(\frac{Z_{\beta,\sigma}}{Z_\beta}\right) \leq \log\left(e^{C_A\beta\sigma^4} + \frac{\exp(\beta C_A\sigma^4)\,\Gamma(\frac{d}{2}+1)\int_{A^c} e^{-\beta_0 v(\theta)}d\theta}{e^{(\beta-\beta_0)(v(\theta_v^*)+\delta_v)}\,(1+\bar{c}_A(\beta))}\right). \tag{47}$$

Since $\sigma^4 = \beta^{-(1+\eta)}$, then $\mathbb{E}[\mathcal{R}(\theta,\epsilon)] = O(\beta^{-(1+\eta)})$. Therefore, we can bound 33 using 47, so that

$$\begin{aligned}\mathrm{KL}(\pi_\beta^{\mathrm{fSGLD}}||\pi^\star_{\beta,\sigma}) \leq \log&\left(e^{C_A\beta^{-\eta}} + \frac{\exp(C_A\beta^{-\eta})\,\Gamma(\frac{d}{2}+1)\int_{A^c} e^{-\beta_0 v(\theta)}d\theta}{e^{(\beta-\beta_0)(v(\theta_v^*)+\delta_v)}\,(1+\bar{c}_A(\beta))}\right) \\ &- \beta\int_{\mathbb{R}^d}\mathbb{E}[\mathcal{R}(\theta,\epsilon)]\,\pi_\beta^{\mathrm{fSGLD}}(\theta)\,\mathrm{d}\theta =: C_1.\end{aligned} \tag{48}$$

By the Stirling's formula, we have $\log(\Gamma(\frac{d}{2}+1))$ is $O(d\log d)$, and the right-hand side of 48 is $O(\beta^{-\eta}d\log d)$. Then, one obtains

$$\lim_{\beta\to\infty}\mathrm{KL}(\pi_\beta^{\mathrm{fSGLD}}||\pi^\star_{\beta,\sigma}) = 0. \tag{49}$$

We apply Corollary C.3 with $\widetilde{\kappa}=1$, to prove the asymptotic convergence in Wasserstein distance of order two between $\pi_\beta^{\mathrm{fSGLD}}$ and $\pi^\star_{\beta,\sigma}$. First, we provide a bound on the constant $C_{\pi^\star_{\beta,\sigma}}$ in Corollary C.3 using $\log(x+y)\leq\log 2+\max(\log(x),\log(y))$ for all $x,y>0$

$$\begin{aligned}&C^2_{\pi^\star_{\beta,\sigma}} \\ &\leq 6+4\log\left(\int_{\mathbb{R}^d} e^{|\theta|^2-\beta u(\theta)-\frac{\beta\sigma^2}{2}\mathrm{tr}(H(\theta))}\,\mathrm{d}\theta\right) \\ &\leq 6+4\log 2+4\max\left(\log\left(\int_A e^{|\theta|^2-\beta u(\theta)-\frac{\beta\sigma^2}{2}\mathrm{tr}(H(\theta))}\,\mathrm{d}\theta\right),\log\left(\int_{A^c} e^{|\theta|^2-\beta u(\theta)-\frac{\beta\sigma^2}{2}\mathrm{tr}(H(\theta))}\,\mathrm{d}\theta\right)\right).\end{aligned} \tag{50}$$

From 22, recall that b is $O(d\beta^{-\frac{1+\eta}{2}})$. We can control the first integral on the right-hand side of 50 using Remark C.4, i.e.

$$\log\left(\int_A e^{|\theta|^2-\beta u(\theta)-\frac{\beta\sigma^2}{2}\mathrm{tr}(H(\theta))}\,\mathrm{d}\theta\right) \leq \frac{b}{a}+\log(\widetilde{c}_{A,\beta})+\frac{d}{2}\log\left(\frac{\pi b}{a}\right). \tag{51}$$

where $\widetilde{c}_{A,\beta} := \max_{\theta\in A} e^{-\beta u(\theta)-\frac{\beta\sigma^2}{2}\mathrm{tr}(H(\theta))}$. For $\theta \in A^c$ and $\widetilde{c} \in (0,1)$, we have, by Assumption 3,

$$\begin{aligned} u(\theta) &= u(\widetilde{c}\theta) + \int_{\widetilde{c}}^1 \langle \theta, \nabla u(t\theta)\rangle\,\mathrm{d}t \\ &\geq u(\theta_u^\star) + \int_{\widetilde{c}}^1 t^{-1}\langle t\theta, \nabla u(t\theta)\rangle\,\mathrm{d}t \\ &\geq u(\theta_u^\star) + \int_{\widetilde{c}}^1 t^{-1}(\bar{a}|t\theta|^2-\bar{b})\,\mathrm{d}t \\ &\geq \frac{\bar{a}(1-\widetilde{c}^2)}{2}|\theta|^2 + \bar{b}\log\widetilde{c} + u(\theta_u^\star) \\ &= \bar{c}|\theta|^2+\bar{p}, \end{aligned} \tag{52}$$

where $\bar{c} := \frac{\bar{a}(1-\widetilde{c}^2)}{2} > 0$, and $\bar{p} := \bar{b}\log\widetilde{c} + u(\theta_u^\star)$. For any $\theta \in A^c$, there exists $\delta_u > 0$ such that

$$u(\theta) > u(\theta_u^*) + \delta_u, \quad \text{for} \quad \theta \in A^c. \tag{53}$$

For any $\beta_0 \in (\frac{1}{\bar{c}}, \beta) = (\frac{2}{\bar{a}(1-\widetilde{c}^2)}, \beta)$, we have

$$-(\beta-\beta_0)u(\theta) < -(\beta-\beta_0)(u(\theta_u^*)+\delta_u), \quad \text{for} \quad \theta \in A^c. \tag{54}$$

Using 54, Assumption 3, 52, and $\beta_0 > \frac{1}{\bar{c}}$, one obtains

$$\begin{aligned} \log\left(\int_{A^c} e^{|\theta|^2-\beta u(\theta)-\frac{\beta\sigma^2}{2}\mathrm{tr}(H(\theta))}\mathrm{d}\theta\right) &\leq \log\left(e^{-(\beta-\beta_0)(u(\theta_u^*)+\delta_u)-\beta\sigma^2 d\bar{a}/2}\int_{A^c} e^{|\theta|^2-\beta_0 u(\theta)}\mathrm{d}\theta\right) \\ &\leq \log\left(e^{-(\beta-\beta_0)(u(\theta_u^*)+\delta_u)-\beta_0\bar{p}-\beta\sigma^2 d\bar{a}/2}\int_{A^c} e^{(1-\beta_0\bar{c})|\theta|^2}\mathrm{d}\theta\right) \\ &\leq \log\left(e^{-(\beta-\beta_0)(u(\theta_u^*)+\delta_u)-\beta_0\bar{p}-\beta\sigma^2 d\bar{a}/2}+2\right) + \frac{d}{2}\log\left(\frac{\pi}{\beta_0\bar{c}-1}\right). \end{aligned} \tag{55}$$

Plugging 51 and 55 in 50 with $\sigma^4 = \beta^{-(1+\eta)}$, yields

$$\begin{aligned} C^2_{\pi^\star_{\beta,\sigma}} \leq\ & 6+4\log 2 \\ & + 4\max\left(\frac{b}{a}+\log(\widetilde{c}_{A,\beta})+\frac{d}{2}\log\left(\frac{\pi b}{a}\right),\right. \\ & \qquad \left.\log\left(e^{-(\beta-\beta_0)(u(\theta_u^*)+\delta_u)-\beta_0\bar{p}-\beta^{\frac{1-\eta}{2}}d\bar{a}/2}+2\right)+\frac{d}{2}\log\left(\frac{\pi}{\beta_0\bar{c}-1}\right)\right), \end{aligned} \tag{56}$$

which is $O(d\log(d\beta^{-\frac{1+\eta}{2}}))$. Therefore, applying Corollary C.3 with 56 and 48, we obtain

$$\begin{aligned} & W_2(\pi_\beta^{\mathrm{fSGLD}}, \pi^\star_{\beta,\sigma}) \\ & \leq \left[6+4\log 2\right. \\ & \quad + 4\max\left(\frac{b}{a}+\log(\widetilde{c}_{A,\beta})+\frac{d}{2}\log\left(\frac{\pi b}{a}\right),\right. \\ & \qquad \left.\left.\log\left(e^{-(\beta-\beta_0)(u(\theta_u^*)+\delta_u)-\beta_0\bar{p}-\beta^{\frac{1-\eta}{2}}d\bar{a}/2}+2\right)+\frac{d}{2}\log\left(\frac{\pi}{\beta_0\bar{c}-1}\right)\right)\right]^{\frac{1}{2}} \\ & \quad \times \left[\left(\sqrt{C_1}+2^{-\frac{1}{4}}(C_1)^{1/4}\right)\right] =: \underline{D}, \end{aligned} \tag{57}$$

Therefore, 57 is $O(\beta^{-\frac{\eta}{2}} d \log d)$. Taking the limit for $\beta \to \infty$ in 57 and using 49, we arrive at

$$\lim_{\beta \to \infty} W_2(\pi_\beta^{\text{fSGLD}}, \pi^\star_{\beta,\sigma}) = 0. \tag{58}$$

□

We use the following triangle inequality to establish a non-asymptotic bound for $W_1(\mathcal{L}(\theta_k^{\text{fSGLD}}), \pi^\star_{\beta,\sigma})$:

$$\begin{aligned} W_1(\mathcal{L}(\theta_k^{\text{fSGLD}}), \pi^\star_{\beta,\sigma}) \leq W_1(\mathcal{L}(\bar{\theta}_t^{\text{fSGLD}}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}})) + W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}), \mathcal{L}(Z_t^{\lambda,\text{fSGLD}})) \\ + W_1(\mathcal{L}(Z_t^{\lambda,\text{fSGLD}}), \pi_\beta^{\text{fSGLD}}) + W_1(\pi_\beta^{\text{fSGLD}}, \pi^\star_{\beta,\sigma}). \end{aligned} \tag{59}$$

We control the four terms on the right-hand side of 59 separately. The bounds for the first three terms follow from Zhang et al. (2023), with Zhang et al. (2023, Assumptions 1) replaced by Assumptions 1, and using the Lipschitzness and dissipativity of g_ϵ established in Remark B.1 and Remark B.2. For completeness, we reproduce these proofs here to make the convergence analysis of fSGLD self-contained.

We define, for each $p \geq 1$, the Lyapunov function $\widetilde{V}_p$ by $\widetilde{V}_p(\theta) := (1 + |\theta|^2)^{p/2}$, $\theta \in \mathbb{R}^d$, and similarly $\widetilde{v}_p(\omega) := (1 + \omega^2)^{p/2}$, for any real $\omega \geq 0$. These functions are twice continuously differentiable and

$$\sup_\theta (|\nabla \widetilde{V}_p(\theta)| / \widetilde{V}_p(\theta)) < \infty, \qquad \lim_{|\theta| \to \infty} (|\nabla \widetilde{V}_p(\theta)| / \widetilde{V}_p(\theta)) = 0. \tag{60}$$

Let $\mathcal{P}_{\widetilde{V}_p}$ denote the set of $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} \widetilde{V}_p(\theta)\, \mu(\mathrm{d}\theta) < \infty$. Then, we define a functional that plays a central role in establishing the convergence rate in the Wasserstein-1 distance. For $\mu, \nu \in \mathcal{P}_{\widetilde{V}_2}$, let

$$w_{1,2}(\mu, \nu) := \inf_{\Gamma \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} [1 \wedge |\theta - \theta'|](1 + \widetilde{V}_2(\theta) + \widetilde{V}_2(\theta'))\, \Gamma(\mathrm{d}\theta, \mathrm{d}\theta'). \tag{61}$$

Moreover, it holds that $W_1(\mu, \nu) \leq w_{1,2}(\mu, \nu)$.

Proposition C.5. *Let Assumptions 1, 2, and 3 hold. Let $(\tilde{Z}_t^{fSGLD})_{t \in \mathbb{R}_+}$ be the solution of 25 with initial condition $\tilde{Z}_0^{fSGLD} = \tilde{\theta}_0$ which is independent of $\mathcal{F}_\infty$ and satisfies $\mathbb{E}[|\tilde{\theta}_0|^2] < \infty$. Then,*

$$w_{1,2}(\mathcal{L}(Z_t^{fSGLD}), \mathcal{L}(\tilde{Z}_t^{fSGLD})) \leq \hat{c} e^{-\dot{c}t} w_{1,2}(\mathcal{L}(\theta_0), \mathcal{L}(\tilde{\theta}_0)),$$

where the constants $\dot{c}$ and $\hat{c}$ are given in Lemma C.6.

Proof. From Remark B.1, one can deduce

$$\begin{aligned} |\nabla_\theta g_\epsilon(\theta) - \nabla_{\theta'} g_\epsilon(\theta')| &\leq \mathbb{E}[|\nabla_\theta u(\theta + \epsilon) - \nabla_{\theta'} u(\theta' + \epsilon)|] \\ &\leq L_1 \mathbb{E}[\varphi(X_0)] |\theta - \theta'|. \end{aligned} \tag{62}$$

The rest of the proof follows using Assumption 1, 2, and 3, 62, Lemma C.15, and 60 in Zhang et al. (2023, Proof of Proposition 4.6). □

The constants $\dot{c}$ and $\hat{c}$ from Proposition C.5 are given in an explicit form.

Lemma C.6. *The contraction constant $\dot{c} > 0$ in Proposition C.5 is given by*

$$\dot{c} := \min\left\{\bar{\phi}, \bar{c}(2), 4\tilde{c}(2)\varepsilon\bar{c}(2)/2\right\}/2 \tag{63}$$

where $\bar{c}(2) = a/2$, $\tilde{c}(2) = (3/2)a v_2(\bar{M}_2)$ with $\bar{M}_2$ given in Lemma C.15, $\bar{\phi}$ is given by

$$\bar{\phi} := \left(\bar{r}\sqrt{8\pi/(\beta L_1 \mathbb{E}[\varphi(X_0)])} \exp\left(\left(\bar{r}\sqrt{\beta L_1 \mathbb{E}[\varphi(X_0)]/8} + \sqrt{8/(\beta L_1 \mathbb{E}[\varphi(X_0)])}\right)^2\right)\right)^{-1}, \tag{64}$$

and moreover, $\varepsilon > 0$ can be chosen such that the following inequality is satisfied

$$\varepsilon \leq 1 \wedge \left(4\tilde{c}(2)\sqrt{2\beta\pi/(L_1 \mathbb{E}[\varphi(X_0)])} \int_0^{\tilde{r}} \exp\left(s\sqrt{\beta L_1 \mathbb{E}[\varphi(X_0)]/8} + \sqrt{8/(\beta L_1 \mathbb{E}[\varphi(X_0)])}\right)^2 \mathrm{d}s\right)^{-1}, \tag{65}$$

where $\tilde{r} := 2\sqrt{2\tilde{c}(2)/\bar{c}(2) - 1}$ and $\bar{r} := 2\sqrt{4\tilde{c}(2)(1 + \bar{c}(2))/\bar{c}(2) - 1}$. The constant $\hat{c} > 0$ is given by $\hat{c} := 2(1 + \bar{r}) \exp(\beta L_1 \mathbb{E}[\varphi(X_0)]\bar{r}^2/8 + 2\bar{r})/\varepsilon$.

Proof. This follows by adapting the arguments of Zhang et al. (2023, Proof of Lemma 4.11) to the *flatness* Langevin SDE 25, using 62 together with Lemma C.15. □

From the definition of $\lambda_{\max}$ given in 24, it follows that $0 < \lambda \leq \lambda_{\max} \leq 1$, and hence $1/2 < \lambda T \leq 1$. We now proceed to bound the first term in 59.

Lemma C.7. *Let Assumptions 1, 2, and 3 hold. For any $0 < \lambda < \lambda_{max}$ given in 24, $t \in (kT, (k+1)T]$,*

$$W_2(\mathcal{L}(\bar{\theta}_t^{fSGLD}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD})) \leq \sqrt{\lambda}\left(e^{-ak/4}\bar{D}_{2,1}\mathbb{E}[\widetilde{V}_2(\theta_0)] + \bar{D}_{2,2}\right)^{1/2},$$

where

$$\begin{aligned}\bar{D}_{2,1} &:= 4e^{4L_1^2\mathbb{E}[\varphi^2(X_0)]}(L_1^2\mathbb{E}[\varphi^2(X_0)]\bar{\psi}_Y + \bar{\psi}_Z),\\ \bar{D}_{2,2} &:= 4e^{4L_1^2\mathbb{E}[\varphi^2(X_0)]}(L_1^2\mathbb{E}[\varphi^2(X_0)]\widetilde{\psi}_Y + \widetilde{\psi}_Z),\end{aligned} \tag{66}$$

with $\bar{\psi}_Y$, $\widetilde{\psi}_Y$ given in 96, and $\bar{\psi}_Z$, $\widetilde{\psi}_Z$ given in 97.

Proof. This follows by applying Lemma C.17 together with the argument used in Zhang et al. (2023, Proof of Lemma 4.7). We summarize the main steps in the following. Using synchronous coupling together with 27, Definition C.2, Remark B.1, and it follows that for any $t \in (kT, (k+1)T]$,

$$\begin{aligned}\left|\bar{\Phi}_t^{\lambda,k,\mathrm{fSGLD}} - \bar{\theta}_t^{\mathrm{fSGLD}}\right| &\leq \lambda\left|\int_{kT}^t \left[\nabla_\theta U(\bar{\theta}_{\lfloor s\rfloor}^{\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil}) - \nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}})\right]\mathrm{d}s\right|\\
&\leq \lambda\left|\int_{kT}^t \left[\nabla_\theta U(\bar{\theta}_{\lfloor s\rfloor}^{\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|\\
&\quad + \lambda\left|\int_{kT}^t \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|\\
&\leq \lambda L_1\int_{kT}^t \varphi(X_{\lceil s\rceil})\left|\bar{\theta}_{\lfloor s\rfloor}^{\mathrm{fSGLD}} - \bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}\right|\mathrm{d}s\\
&\quad + \lambda\left|\int_{kT}^t \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|.\end{aligned} \tag{67}$$

Squaring both sides of 67 and taking expectations, we obtain using Assumption 1

$$\begin{aligned}\mathbb{E}\left[\left|\bar{\Phi}_t^{\lambda,k,\mathrm{fSGLD}} - \bar{\theta}_t^{\mathrm{fSGLD}}\right|^2\right] &\leq 2\lambda L_1^2\int_{kT}^t \mathbb{E}\left[\varphi^2(X_0)\right]\mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\mathrm{fSGLD}} - \bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}\right|^2\right]\mathrm{d}s\\
&\quad + 2\lambda^2\mathbb{E}\left[\left|\int_{kT}^t \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right].\end{aligned}$$

From $\lambda T \leq 1$ and Lemma C.17, we get

$$\begin{aligned}&\mathbb{E}\left[\left|\bar{\Phi}_t^{\lambda,k,\mathrm{fSGLD}} - \bar{\theta}_t^{\mathrm{fSGLD}}\right|^2\right]\\
&\leq 4\lambda L_1^2\mathbb{E}\left[\varphi^2(X_0)\right]\int_{kT}^t \mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\mathrm{fSGLD}} - \bar{\theta}_s^{\mathrm{fSGLD}}\right|^2\right]\mathrm{d}s\\
&\quad + 4\lambda L_1^2\mathbb{E}\left[\varphi^2(X_0)\right]\int_{kT}^t \mathbb{E}\left[\left|\bar{\theta}_s^{\mathrm{fSGLD}} - \bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}\right|^2\right]\mathrm{d}s\\
&\quad + 2\lambda^2\mathbb{E}\left[\left|\int_{kT}^t \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right]\\
&\leq 4\lambda L_1^2\mathbb{E}\left[\varphi^2(X_0)\right]\left(e^{-\lambda akT}\bar{\psi}_Y\mathbb{E}[\widetilde{V}_2(\theta_0)] + \widetilde{\psi}_Y\right)\\
&\quad + 4\lambda L_1^2\mathbb{E}\left[\varphi^2(X_0)\right]\int_{kT}^t \mathbb{E}\left[\left|\bar{\theta}_s^{\mathrm{fSGLD}} - \bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}\right|^2\right]\mathrm{d}s\\
&\quad + 2\lambda^2\mathbb{E}\left[\left|\int_{kT}^t \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\mathrm{fSGLD}} + \epsilon_{\lceil s\rceil}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right],\end{aligned} \tag{68}$$

We now bound the last term in 68 by splitting the final integral. Let $kT + N < t \leq kT + N + 1$ with $N + 1 \leq T, N \in \mathbb{N}$. It follows that

$$\left| \int_{kT}^{t} \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{\lceil s \rceil}, X_{\lceil s \rceil}) \right] \mathrm{d}s \right| = \left| \sum_{n=1}^{N} I_n + R_N \right|,$$

where $I_n := \int_{kT+(n-1)}^{kT+n} [\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n})] \mathrm{d}s$, and $R_N := \int_{kT+N}^{t} [\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+N+1}, X_{kT+N+1})] \mathrm{d}s$. Squaring both sides, we obtain

$$\left| \sum_{n=1}^{N} I_n + R_N \right|^2 = \sum_{n=1}^{N} |I_n|^2 + 2 \sum_{n=2}^{N} \sum_{j=1}^{n-1} \langle I_n, I_j \rangle + 2 \sum_{n=1}^{N} \langle I_n, R_N \rangle + |R_N|^2.$$

Let $\mathcal{H}_\epsilon$ denote the sigma-algebra generated by ϵ. We define the filtration $\mathcal{J}_t = \mathcal{F}_\infty^\lambda \vee \mathcal{X}_{\lfloor t \rfloor} \vee \mathcal{H}_{\lfloor \epsilon \rfloor}$ and we take expectations of both sides. Observe that for any $n = 2, \ldots, N$, $j = 1, \ldots, n-1$,

$$\begin{aligned}
&\mathbb{E}\left[\langle I_n, I_j \rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}[\langle I_n, I_j \rangle | \mathcal{J}_{kT+j}]\right], \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \int_{kT+(n-1)}^{kT+n} [\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n})] \mathrm{d}s, \right.\right.\right. \\
&\qquad \left.\left.\left. \int_{kT+(j-1)}^{kT+j} [\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+j}, X_{kT+j})] \mathrm{d}s \right\rangle \middle| \mathcal{J}_{kT+j} \right] \right] \\
&= \mathbb{E}\left[\left\langle \int_{kT+(n-1)}^{kT+n} \mathbb{E}\left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n}) \middle| \mathcal{J}_{kT+j} \right] \mathrm{d}s, \right.\right. \\
&\qquad \left.\left. \int_{kT+(j-1)}^{kT+j} [\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+j}, X_{kT+j})] \mathrm{d}s \right\rangle \right] = 0.
\end{aligned}$$

By the same reasoning, $\mathbb{E}\langle I_n, R_N \rangle = 0$ for all $1 \leq n \leq N$. Combining these results, we can bound the last term on the right-hand side of 68 using Lemma C.18

$$\begin{aligned}
&2\lambda^2 \mathbb{E}\left[\left| \int_{kT}^{t} \left[\nabla g_\epsilon(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_s^{\lambda,k,\text{fSGLD}} + \epsilon_{\lceil s \rceil}, X_{\lceil s \rceil}) \right] \mathrm{d}s \right|^2 \right] \\
&= 2\lambda^2 \sum_{n=1}^{N} \mathbb{E}\left[|I_n|^2 \right] + 2\lambda^2 \mathbb{E}\left[|R_N|^2 \right] \\
&\leq 4e^{-a\lambda kT/2} \lambda (\bar{\psi}_Z \mathbb{E}[\widetilde{V}_2(\theta_0)] + \widetilde{\psi}_Z).
\end{aligned}$$

Consequently, 68 is bounded as follows

$$\begin{aligned}
\mathbb{E}\left[\left| \bar{\Phi}_t^{\lambda,k,\text{fSGLD}} - \bar{\theta}_t^{\text{fSGLD}} \right|^2 \right] \leq\ & 4\lambda L_1^2 \mathbb{E}\left[\varphi^2(X_0) \right] \int_{kT}^{t} \mathbb{E}\left[\left| \bar{\theta}_s^{\text{fSGLD}} - \bar{\Phi}_s^{\lambda,k,\text{fSGLD}} \right|^2 \right] \mathrm{d}s \\
&+ 4e^{-a\lambda kT/2} \lambda (L_1^2 \mathbb{E}\left[\varphi^2(X_0) \right] \bar{\psi}_Y + \bar{\psi}_Z) \mathbb{E}[\widetilde{V}_2(\theta_0)] \\
&+ 4\lambda (L_1^2 \mathbb{E}\left[\varphi^2(X_0) \right] \widetilde{\psi}_Y + \widetilde{\psi}_Z).
\end{aligned}$$

Applying Grönwall's inequality yields

$$\begin{aligned}
\mathbb{E}\left[\left| \bar{\Phi}_t^{\lambda,k,\text{fSGLD}} - \bar{\theta}_t^{\text{fSGLD}} \right|^2 \right] \leq \lambda e^{4L_1^2 \mathbb{E}[\varphi^2(X_0)]} \Big[& 4e^{-a\lambda kT/2} (L_1^2 \mathbb{E}\left[\varphi^2(X_0) \right] \bar{\psi}_Y + \bar{\psi}_Z) \mathbb{E}[\widetilde{V}_2(\theta_0)] \\
& + 4(L_1^2 \mathbb{E}\left[\varphi^2(X_0) \right] \widetilde{\psi}_Y + \widetilde{\psi}_Z) \Big].
\end{aligned}$$

Finally, we obtain using $\lambda T \geq 1/2$,

$$\begin{aligned}
W_2^2(\mathcal{L}(\bar{\theta}_t^{\text{fSGLD}}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}})) &\leq \mathbb{E}\left| \bar{\Phi}_t^{\lambda,k,\text{fSGLD}} - \bar{\theta}_t^{\text{fSGLD}} \right|^2 \\
&\leq \lambda (e^{-an/4} \bar{C}_{2,1} \mathbb{E}[\widetilde{V}_2(\theta_0)] + \bar{C}_{2,2}),
\end{aligned} \tag{69}$$

where

$$\bar{D}_{2,1} := 4e^{4L_1^2\mathbb{E}[\varphi^2(X_0)]}(L_1^2\mathbb{E}\left[\varphi^2(X_0)\right]\bar{\psi}_Y + \bar{\psi}_Z),$$
$$\bar{D}_{2,2} := 4e^{4L_1^2\mathbb{E}[\varphi^2(X_0)]}(L_1^2\mathbb{E}\left[\varphi^2(X_0)\right]\widetilde{\psi}_Y + \widetilde{\psi}_Z).$$

□

The bound for the second term on the right-hand side of 59 is established in the following lemma.

Lemma C.8. *Let Assumptions 1, 2, and 3 hold. For any $0 < \lambda < \lambda_{max}$ given in 24, $t \in (kT, (k+1)T]$,*

$$W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD}), \mathcal{L}(Z_t^{\lambda,fSGLD})) \leq \sqrt{\lambda}(e^{-\dot{c}k/2}\bar{D}_{2,3}\mathbb{E}[\widetilde{V}_4(\theta_0)] + \bar{D}_{2,4}),$$

where

$$\begin{aligned}\bar{D}_{2,3} &= \hat{c}\left(1 + \frac{2}{\dot{c}}\right)(e^{a/2}\bar{D}_{2,1} + 12),\\ \bar{D}_{2,4} &= \frac{\hat{c}}{1-\exp(-\dot{c})}(\bar{D}_{2,2} + 12c_3(\lambda_{max} + a^{-1}) + 9\widetilde{v}_4(\bar{M}_4) + 15),\end{aligned} \tag{70}$$

with $\bar{D}_{2,1}$, $\bar{D}_{2,2}$ given in 66, $\hat{c}$, $\dot{c}$ given in Lemma C.6, c_3 given in 95, and $\bar{M}_4$ given in Lemma C.15.

Proof. This follows by applying Proposition C.5, Lemma C.7, Corollary C.14, and Lemma C.16 together with the arguments in Zhang et al. (2023, Proof of Lemma 4.8). □

Adapting the reasoning of Lemma C.8, we establish a non-asymptotic W_2 bound between $\mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}})$ and $\mathcal{L}(Z_t^{\lambda,\text{fSGLD}})$, presented in the next corollary.

Corollary C.9. *Let Assumptions 1, 2, and 3 hold. For any $0 < \lambda < \lambda_{max}$ given in 24, $t \in (kT, (k+1)T]$,*

$$W_2(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD}), \mathcal{L}(Z_t^{\lambda,fSGLD})) \leq \lambda^{1/4}(e^{-\dot{c}/4}\bar{D}_{2,3}^{\star}(\mathbb{E}[\widetilde{V}_4(\theta_0)])^{1/2} + \bar{D}_{2,4}^{\star}),$$

where

$$\begin{aligned}\bar{D}_{2,3}^{\star} &:= \sqrt{2\hat{c}}(1 + 4/\dot{c})(e^{a/8}\bar{D}_{2,1}^{1/2} + 2\sqrt{2}),\\ \bar{D}_{2,4}^{\star} &:= \frac{\sqrt{2\hat{c}}}{1-\exp(-\dot{c}/2)}(\bar{D}_{2,2}^{1/2} + 2\sqrt{2c_3}(\lambda_{max} + a^{-1})^{1/2} + \sqrt{3}\widetilde{v}_4^{1/2}(\bar{M}_4) + \sqrt{15}),\end{aligned} \tag{71}$$

with $\bar{D}_{2,1}$, $\bar{D}_{2,2}$ given in 66, $\hat{c}$, $\dot{c}$ given in Lemma C.6, c_3 given in 95, and $\bar{M}_4$ given in Lemma C.15.

Proof. This follows using Proposition C.5, Lemma C.7, Corollary C.14, and Lemma C.16 in Zhang et al. (2023, Proof of Corollary 4.9). □

We can now derive a non-asymptotic bound for the first three terms on the right-hand side of 59 in W_1 distance.

Theorem C.10. *Let Assumptions 1, 2, and 3 hold. Then, there exist constants $\dot{c}, D_1, D_2, D_3 > 0$ such that, for every $\beta > 0$, for $0 < \lambda < \lambda_{max}$, any $t \in (kT, (k+1)T]$, and $k \in \mathbb{N}$,*

$$\begin{aligned}&W_1(\mathcal{L}(\bar{\theta}_t^{fSGLD}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD})) + W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD}), \mathcal{L}(Z_t^{\lambda,fSGLD})) + W_1(\mathcal{L}(Z_t^{\lambda,fSGLD}), \pi_\beta^{fSGLD})\\ &\leq D_1 e^{-\dot{c}\lambda k/2}(1 + \mathbb{E}[|\theta_0|^4]) + (D_2 + D_3)\sqrt{\lambda},\end{aligned}$$

where

$$\begin{aligned}D_1 &:= 2e^{\dot{c}/2}\left[(\lambda_{max}^{1/2}(\bar{D}_{2,1}^{1/2} + \bar{D}_{2,2}^{1/2} + \bar{D}_{2,3} + \bar{D}_{2,4}) + \hat{c}) + \hat{c}\left(1 + \int_{\mathbb{R}^d}\widetilde{V}_2(\theta)\pi_{\beta,\sigma}(\mathrm{d}\theta)\right)\right]\\ &= O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}}}\right)\right),\\ D_2 &:= \bar{D}_{2,1}^{1/2} + \bar{D}_{2,2}^{1/2} = O\left(1 + \sqrt{\frac{d}{\beta}}\right),\\ D_3 &:= \bar{D}_{2,3} + \bar{D}_{2,4} = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}}}\right)\right),\end{aligned} \tag{72}$$

with $\hat{c}$, $\dot{c}$ given in Lemma C.6, $\bar{D}_{2,1}$, $\bar{D}_{2,2}$ given in 66 (Lemma C.7), $\bar{D}_{2,3}$, $\bar{D}_{2,4}$ given in 70 (Lemma C.8), $D_\star > 0$ is independent of d, β, k.

Proof. Using Lemma C.7, and Lemma C.8 in Zhang et al. (2023, Proof of Lemma 4.10), we obtain for $t \in (kT, (k+1)T]$,

$$\begin{aligned} &W_1(\mathcal{L}(\bar{\theta}_t^{\text{fSGLD}}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}})) + W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}), \mathcal{L}(Z_t^{\lambda,\text{fSGLD}})) \\ &\leq (\bar{D}_{2,1}^{1/2} + + \bar{D}_{2,2}^{1/2} + \bar{D}_{2,3} + \bar{D}_{2,4})\sqrt{\lambda}[(e^{-\dot{c}k/2}\mathbb{E}[\widetilde{V}_4(\theta_0)] + 1)], \end{aligned} \tag{73}$$

where $\bar{D}_{2,1}$, $\bar{D}_{2,2}$ are given in 66 (Lemma C.7), and $\bar{D}_{2,3}$, $\bar{D}_{2,4}$ are given in 70 (Lemma C.8). The remainder of the proof follows by applying 73 and Proposition C.5 in Zhang et al. (2023, Proof of Theorem 2.4). □

An analogous result to Theorem C.10 holds in Wasserstein-2 distance, as stated in the next corollary.

Corollary C.11. *Let Assumption 1, 2 and 3 hold. Then, there exists constants $\dot{c}, D_4, D_5, D_6 > 0$ such that, for every $\beta > 0$, $0 < \lambda \leq \lambda_{max}$, any $t \in (kT, (k+1)T]$, and $k \in \mathbb{N}$,*

$$\begin{aligned} &W_2(\mathcal{L}(\bar{\theta}_t^{fSGLD}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD})) + W_2(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,fSGLD}), \mathcal{L}(Z_t^{\lambda,fSGLD})) + W_2(\mathcal{L}(Z_t^{\lambda,fSGLD}), \pi_\beta^{fSGLD}) \\ &\leq D_4 e^{-\dot{c}\lambda k/4}(\mathbb{E}[|\theta_0|^4] + 1) + (D_5 + D_6)\lambda^{1/4}, \end{aligned}$$

where

$$\begin{aligned} D_4 &:= 2(\lambda_{max}^{1/2}(\bar{D}_{2,1}^{1/2} + \bar{D}_{2,2}^{1/2}) + \lambda_{max}^{1/4}(\bar{D}_{2,3}^\star + \bar{D}_{2,4}^\star) + \sqrt{2}\hat{c}^{1/2}) \\ &\quad + \sqrt{2}\hat{c}^{1/2}\left(1 + \int_{\mathbb{R}^d} \widetilde{V}_2(\theta)\pi_\beta^{fSGLD}(\mathrm{d}\theta)\right) \\ &= O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1 - e^{-\dot{c}/2}}\right)\right) \\ D_5 &:= \lambda_{max}^{1/4}\bar{D}_{2,1}^{1/2} + \lambda_{max}^{1/4}\bar{D}_{2,2}^{1/2} = O\left(1 + \sqrt{\frac{d}{\beta}}\right) \\ D_6 &:= \bar{D}_{2,3}^\star + \bar{D}_{2,4}^\star = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1 - e^{-\dot{c}/2}}\right)\right), \end{aligned} \tag{74}$$

where $\hat{c}$, $\dot{c}$ given in Lemma C.6, $\bar{D}_{2,1}$, $\bar{D}_{2,2}$ given in 66 (Lemma C.7), $\bar{D}_{2,3}^\star$, $\bar{D}_{2,4}^\star$ given in 71 (Corollary C.9), $D_\star > 0$ is independent of d, β, k.

Proof. This follows by applying Lemma C.7, Corollary C.9, and Proposition C.5 in Zhang et al. (2023, Proof of Corollary 2.5). □

Proof of Theorem 3.2. Using Theorem C.10 and Proposition 3.1 in 59, we get

$$\begin{aligned} W_1(\mathcal{L}(\theta_k^{\text{fSGLD}}), \pi_{\beta,\sigma}^\star) &\leq W_1(\mathcal{L}(\bar{\theta}_t^{\text{fSGLD}}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}})) + W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}), \mathcal{L}(Z_t^{\lambda,\text{fSGLD}})) \\ &\quad + W_1(\mathcal{L}(Z_t^{\lambda,\text{fSGLD}}), \pi_\beta^{\text{fSGLD}}) + W_1(\pi_\beta^{\text{fSGLD}}, \pi_{\beta,\sigma}^\star) \\ &\leq D_1 e^{-\dot{c}\lambda k/2}(1 + \mathbb{E}[|\theta_0|^4]) + (D_2 + D_3)\sqrt{\lambda} + W_2(\pi_\beta^{\text{fSGLD}}, \pi_{\beta,\sigma}^\star) \\ &\leq D_1 e^{-\dot{c}\lambda k/2}(1 + \mathbb{E}[|\theta_0|^4]) + (D_2 + D_3)\sqrt{\lambda} + \underline{D}. \end{aligned} \tag{75}$$

In addition, for any $\bar{\delta} > 0$, if we choose λ, k and β in 75 such that $\lambda \leq \lambda_{\max}$, and

$$D_1 e^{-\dot{c}\lambda k/2}(1 + \mathbb{E}[|\theta_0|^4]) \leq \frac{\bar{\delta}}{3}, \qquad (D_2 + D_3)\sqrt{\lambda} \leq \frac{\bar{\delta}}{3}, \qquad \underline{D} \leq \frac{\bar{\delta}}{3},$$

then $W_1(\mathcal{L}(\theta_k^{\text{fSGLD}}), \pi_{\beta,\sigma}^\star) \leq \bar{\delta}$. This yields

$$\beta \geq \beta_{\bar{\delta}} := \left(3\underline{D}^0/\bar{\delta}\right)^{\frac{1}{\eta}}, \tag{76}$$

where $\underline{D}^0$ contains the terms independent of β in the right-hand side of 57, and

$$\lambda \leq \lambda_{\bar{\delta}} := \frac{\bar{\delta}^2}{9(D_2 + D_3)^2} \wedge \lambda_{\max}, \tag{77}$$

and $\lambda k \geq \frac{2}{\dot{c}} \ln\left(\frac{3D_1(1+\mathbb{E}[|\theta_0|^4])}{\bar{\delta}}\right)$. From 72, it follows that

$$k \geq k_{\bar{\delta}} := \frac{D_\star e^{D_\star(1+d/\beta)(1+\beta)}}{\bar{\delta}^2 \dot{c}} \left(1 + \frac{1}{(1-e^{-\dot{c}})^2}\right) \ln\left(\frac{D_\star e^{D_\star(1+d/\beta)(1+\beta)}}{\bar{\delta}} \left(1 + \frac{1}{1-e^{-\dot{c}}}\right)\right). \tag{78}$$

□

Proof of Corollary 3.3. Using triangle inequality, Corollary C.11, and Proposition 3.1 in 59, we get, for any $t \in (kT, (k+1)T]$, and $k \in \mathbb{N}$,

$$\begin{aligned} W_2(\mathcal{L}(\theta_k^{\text{fSGLD}}), \pi^\star_{\beta,\sigma}) &\leq W_2(\mathcal{L}(\bar{\theta}_t^{\text{fSGLD}}), \mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}})) + W_2(\mathcal{L}(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}), \mathcal{L}(Z_t^{\lambda,\text{fSGLD}})) \\ &\quad + W_2(\mathcal{L}(Z_t^{\lambda,\text{fSGLD}}), \pi_\beta^{\text{fSGLD}}) + W_2(\pi_\beta^{\text{fSGLD}}, \pi^\star_{\beta,\sigma}) \\ &\leq D_4 e^{-\dot{c}\lambda k/4}(\mathbb{E}[|\theta_0|^4] + 1) + (D_5 + D_6)\lambda^{1/4} + \underline{D}, \end{aligned} \tag{79}$$

In addition, for any $\widetilde{\delta} > 0$, λ, k and β such that $\lambda \leq \lambda_{\max}$, and

$$D_4 e^{-\dot{c}\lambda k/4}(\mathbb{E}[|\theta_0|^4] + 1) \leq \frac{\widetilde{\delta}}{3}, \qquad (D_5 + D_6)\lambda^{1/4} \leq \frac{\widetilde{\delta}}{3}, \qquad \underline{D} \leq \frac{\widetilde{\delta}}{3},$$

then $W_2(\mathcal{L}(\theta_k^{\text{fSGLD}}), \pi^\star_{\beta,\sigma}) \leq \widetilde{\delta}$. This yields

$$\beta \geq \beta_{\widetilde{\delta}} := \left(3\underline{D}^0/\widetilde{\delta}\right)^{\frac{1}{\eta}}, \tag{80}$$

where $\underline{D}^0$ is the same as in the proof of Theorem 3.2,

$$\lambda \leq \lambda_{\widetilde{\delta}} := \frac{\bar{\delta}^4}{81(D_5 + D_6)^4} \wedge \lambda_{\max}, \tag{81}$$

and $\lambda k \geq \frac{4}{\dot{c}} \ln\left(\frac{3D_4(1+\mathbb{E}[|\theta_0|^4])}{\bar{\delta}}\right)$. From 74, it follows that

$$k \geq k_{\widetilde{\delta}} := \frac{D_\star e^{D_\star(1+d/\beta)(1+\beta)}}{\bar{\delta}^4 \dot{c}} \left(1 + \frac{1}{(1-e^{-\dot{c}/2})^4}\right) \ln\left(\frac{D_\star e^{D_\star(1+d/\beta)(1+\beta)}}{\bar{\delta}} \left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right). \tag{82}$$

□

Remark C.12. *The convergence rate in Corollary 3.3 can be improved to $O(\lambda^{\frac{1}{2}})$ under substantially stronger assumptions than Assumption 1, 2 and 3. For example, one may assume that the π_β^{fSGLD} satisfies a log-Sobolev inequality, as in Huang et al. (2025). However, such assumptions go beyond the scope of our work.*

Proof of Theorem 3.6. We begin by decomposing the expected excess risk using the random variable Z_∞^{fSGLD}, for which $\mathcal{L}(Z_\infty^{\text{fSGLD}}) = \pi_\beta^{\text{fSGLD}}$, and obtain

$$\begin{aligned} &\mathbb{E}[g_\epsilon(\theta_k^{\text{fSGLD}})] - \inf_{\theta \in \mathbb{R}^d} g_\epsilon(\theta) \\ &= (\mathbb{E}[g_\epsilon(\theta_k^{\text{fSGLD}})] - \mathbb{E}[g_\epsilon(Z_\infty)]) + (\mathbb{E}[g_\epsilon(Z_\infty)] - \inf_{\theta \in \mathbb{R}^d} g_\epsilon(\theta)). \end{aligned} \tag{83}$$

We proceed by controlling the two terms on the right-hand side of 83 separately. By using Raginsky et al. (2017, Lemma 3.5), Remark B.1 with $\sigma^2 = \beta^{-\frac{1+\eta}{2}}$ for $\eta \in (0,1)$, Lemma C.13, and Corollary C.11, the first term on the RHS of 83 can be bounded by

$$\mathbb{E}[g_\epsilon(\theta_k^{\text{fSGLD}})] - \mathbb{E}[g_\epsilon(Z_\infty)] \leq D_1^\# e^{-\dot{c}\lambda k/4} + D_2^\# \lambda^{1/4}, \tag{84}$$

where

$$D_1^{\#} := D_4(L_1\mathbb{E}[\varphi(X_0)](\mathbb{E}[|\theta_0|^2] + c_1(\lambda_{\max} + a^{-1})) + L_2\mathbb{E}[\bar{\varphi}(X_0)](1 + d\beta^{-(1+\eta)/2}) + \mathbb{E}[\widetilde{G}(\epsilon)]) \times (\mathbb{E}[|\theta_0|^4] + 1),$$
$$D_2^{\#} := (D_5 + D_6) \times (L_1\mathbb{E}[\varphi(X_0)](\mathbb{E}[|\theta_0|^2] + c_1(\lambda_{\max} + a^{-1})) + L_2\mathbb{E}[\bar{\varphi}(X_0)](1 + d\beta^{-(1+\eta)/2}) + \mathbb{E}[\widetilde{G}(\epsilon)]), \tag{85}$$

with $\dot{c}$ given in 63, D_4, D_5, D_6 given in 74, and c_1 given in 94. The second term on the RHS of 83 can be controlled via Raginsky et al. (2017, Proposition 11), which leads to

$$\mathbb{E}[g_\epsilon(Z_\infty)] - \inf_{\theta\in\mathbb{R}^d} g_\epsilon(\theta) \le D_\diamond^{\#}, \tag{86}$$

where

$$D_\diamond^{\#} := \frac{d}{2\beta}\log\left(\frac{eL_1\mathbb{E}[\varphi(X_0)]}{a}\left(\frac{b\beta}{d} + 1\right)\right). \tag{87}$$

Using the estimates from 84 and 86 in 83, we obtain

$$\mathbb{E}[g_\epsilon(\theta_k^{\text{fSGLD}})] - \inf_{\theta\in\mathbb{R}^d} g_\epsilon(\theta) \le D_1^{\#}e^{-\dot{c}\lambda k/4} + D_2^{\#}\lambda^{1/4} + D_\diamond^{\#}. \tag{88}$$

Applying 16 on the LHS of 88, along with 17, and choosing $\sigma^4 = \beta^{-(1+\eta)}$, it follows that

$$\mathbb{E}[g_\epsilon(\theta_k^{\text{fSGLD}})] - \inf_{\theta\in\mathbb{R}^d} v(\theta) \le D_1^{\#}e^{-\dot{c}\lambda k/4} + D_2^{\#}\lambda^{1/4} + D_3^{\#}, \tag{89}$$

where

$$D_1^{\#} = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right),$$
$$D_2^{\#} = O\left(e^{D_\star(1+d/\beta)(1+\beta)}\left(1 + \frac{1}{1-e^{-\dot{c}/2}}\right)\right),$$
$$D_3^{\#} := D_\diamond^{\#} + \beta^{-(1+\eta)}\inf_{\theta\in\mathbb{R}^d}\left(\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta) + \sum_{j=6}^{\infty}\sum_{i_1,i_2,\ldots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_j}}(\theta)\right)$$
$$= O\left((d/\beta)\log(D_\star(\beta^{(1-\eta)/2}+1))\right), \tag{90}$$

with $D_\star > 0$ a constant independent of d, β, k. In addition, for $\underline{\delta} > 0$, if we choose β such that $D_3^{\#} \le \underline{\delta}/3$, then choose λ such that $\lambda \le \lambda_{\max}$ and $D_2^{\#}\lambda^{1/4} \le \underline{\delta}/3$, and choose k such that $D_1^{\sharp}e^{-\dot{c}\lambda k/4} \le \underline{\delta}/3$, we obtain

$$\mathbb{E}[g_\epsilon(\theta_k^{\text{fSGLD}})] - \inf_{\theta\in\mathbb{R}^d} v(\theta) \le \underline{\delta}.$$

This yields

$$\beta \ge \beta_{\underline{\delta}} := \beta_c \vee \frac{9d}{2\underline{\delta}}\log\left(\frac{eL_1\mathbb{E}[\varphi(X_0)]}{ad}(b+1)(d+1)\right) \vee \left[\frac{9}{\underline{\delta}}\inf_{\theta\in\mathbb{R}^d}\left(\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta) + \sum_{j=6}^{\infty}\sum_{i_1,i_2,\ldots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_j}}(\theta)\right)\right]^{\frac{1}{1+\eta}}, \tag{91}$$

where β_c is the root of the function $f^\sharp(\beta) = \frac{\log(\beta+1)}{\beta} - \frac{2\underline{\delta}}{9d}$, with $\beta > 0$. Since

$$D_3^{\sharp} \le \frac{d}{2\beta}\log\left(\frac{eL_1\mathbb{E}[\varphi(X_0)]}{ad}(b+1)(d+1)(\beta+1)\right) + \beta^{-(1+\eta)}\inf_{\theta\in\mathbb{R}^d}\left(\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta) + \sum_{j=6}^{\infty}\sum_{i_1,i_2,\ldots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_j}}(\theta)\right),$$

we can ensure $D_3^\sharp \le \underline{\delta}/3$ by imposing

$$\frac{d}{2\beta}\log\left(\frac{eL_1\mathbb{E}[\varphi(X_0)]}{ad}(b+1)(d+1)\right) \le \frac{\underline{\delta}}{9}, \qquad \frac{d}{2\beta}\log(\beta+1) \le \frac{\underline{\delta}}{9},$$
$$\beta^{-(1+\eta)} \inf_{\theta\in\mathbb{R}^d}\left(\sum_{i,j,k,l=1}^{d}\frac{\partial^4 u}{\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_l}(\theta) + \sum_{j=6}^{\infty}\sum_{i_1,i_2,\dots,i_j=1}^{d}\frac{\partial^j u}{\partial\theta_{i_1}\partial\theta_{i_2}\dots\partial\theta_{i_j}}(\theta)\right) \le \frac{\underline{\delta}}{9}.$$

Moreover, one can verify that

$$\lambda \le \lambda_{\underline{\delta}} := \frac{\underline{\delta}^4}{81(D_2^\sharp)^4} \wedge \lambda_{\max}, \tag{92}$$

and $\lambda k \ge \frac{4}{\dot{c}}\ln\frac{3D_1^\sharp}{\underline{\delta}}$, where $\dot{c}$ is given explicitly in Lemma C.6. This leads to

$$\begin{aligned} k \ge k_{\underline{\delta}} := \frac{D_\star e^{D_\star(1+d/\beta)(1+\beta)}}{\underline{\delta}^4\dot{c}}\left(1+\frac{1}{(1-e^{-\dot{c}/2})^4}\right) \\ \times \ln\left(\frac{D_\star e^{D_\star(1+d/\beta)(1+\beta)}}{\underline{\delta}}\left(1+\frac{1}{1-e^{-\dot{c}/2}}\right)\right). \end{aligned} \tag{93}$$

□

C.3 Auxiliary Results

We present the auxiliary results required for the convergence analysis in Appendix C.2. Their proofs follow the same lines as Zhang et al. (2023), with Zhang et al. (2023, Assumptions 1–3) replaced by Assumptions 1, together with the properties established in Remark B.1 and Remark B.2. For completeness, we include their statements to make the convergence analysis of fSGLD self-contained.

Lemma C.13 (Moment bounds of 27)**.** *Let Assumption 1, 2 and 3 hold. For any $0 < \lambda \le \lambda_{max}$ given in 24, $k \in \mathbb{N}$, $t \in (k, k+1]$,*

$$\mathbb{E}\left[|\bar{\theta}_t^{fSGLD}|^2\right] \le (1-a\lambda(t-k))(1-a\lambda)^k\,\mathbb{E}[|\theta_0|^2] + c_1(\lambda_{max}+a^{-1}),$$

where a and b are given in Remark B.2, and

$$c_1 := c_0 + 2d\beta^{-1}, \quad c_0 := 2b + 8\lambda_{\max}L_2^2\mathbb{E}\left[\bar{\varphi}^2(X_0)\right](1+\sigma^2 d) + 4\lambda_{\max}\mathbb{E}[\widetilde{G}^2(\epsilon)]. \tag{94}$$

Moreover, $\sup_{t>0}\mathbb{E}[|\bar{\theta}_t^{fSGLD}|^2] \le \mathbb{E}[|\theta_0|^2] + c_1(\lambda_{max}+a^{-1}) < \infty$. By a similar argument, one obtains

$$\mathbb{E}\left[|\bar{\theta}_t^{fSGLD}|^4\right] \le (1-a\lambda(t-k))(1-a\lambda)^k\,\mathbb{E}[|\bar{\theta}_0^{fSGLD}|^4 + c_3(\lambda_{max}+a^{-1}),$$

where

$$\begin{aligned} M &:= \max\{(8ba^{-1} + 48a^{-1}\lambda_{\max}(L_2^2\mathbb{E}\left[\bar{\varphi}^2(X_0)\right](1+\sigma^2 d) + \mathbb{E}[\widetilde{G}^2(\epsilon)]))^{1/2}, \\ &\qquad (128a^{-1}\lambda_{\max}^2(L_2^3\mathbb{E}\left[\bar{\varphi}^3(X_0)\right]\mathbb{E}[(1+|\epsilon|)^3] + \mathbb{E}[\widetilde{G}^3(\epsilon)]))^{1/3}\}, \\ c_2 &:= 4bM^2 + 152(1+\lambda_{\max})^3 \\ &\qquad \times\left((1+L_2)^4\mathbb{E}\left[(1+\bar{\varphi}(X_0))^4\right]\mathbb{E}[(1+|\epsilon|)^4] + \mathbb{E}[(1+\widetilde{G}(\epsilon))^4]\right)(1+M)^2, \\ c_3 &:= (1+a\lambda_{max})c_2 + 12d^2\beta^{-2}(\lambda_{max}+9a^{-1}). \end{aligned} \tag{95}$$

Moreover, this implies $\sup_{t>0}\mathbb{E}[|\bar{\theta}_t^{fSGLD}|^4] < \infty$.

Proof. This follows along the same lines as Zhang et al. (2023, Lemma 4.2) under our own Assumptions 1, 2, and 3, and using the estimates in Remark B.1 and B.2. □

Lemma C.13 provides a uniform fourth-moment bound for the process $(\bar{\theta}_t^{fSGLD})_{t\ge0}$ which in turn yields a uniform bound for $\widetilde{V}_4(\bar{\theta}_t^{fSGLD})$, as given in the next corollary.

Corollary C.14. *Let Assumption 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{max}$, $k \in \mathbb{N}$, $t \in (k, k+1]$,

$$\mathbb{E}[\widetilde{V}_4(\bar{\theta}_t^{fSGLD})] \leq 2(1-a\lambda)^{\lfloor t \rfloor}\mathbb{E}[\widetilde{V}_4(\bar{\theta}_0^{fSGLD})] + 2c_3(\lambda_{max} + a^{-1}) + 2,$$

where c_3 *is given in Lemma C.13.*

Proof. This follows from the definition of the Lyapunov function $\widetilde{V}_4$ together with Lemma C.13. □

We establish a drift condition for the flatness Langevin SDE 25, which will be instrumental in deriving moment bounds for the continuous-time proces $\bar{\Phi}_t^{\lambda,k,\text{ fSGLD}}$ in Lemma C.16.

Lemma C.15. *(Chau et al., 2021, Lemma 3.5) Let Assumption 1 and 3 hold. Then, for each* $p \geq 2$, $\theta \in \mathbb{R}^d$,

$$\Delta\widetilde{V}_p(\theta)\beta^{-1} - \langle \nabla g_\epsilon(\theta), \nabla\widetilde{V}_p(\theta)\rangle \leq -\bar{c}(p)\widetilde{V}_p(\theta) + \tilde{c}(p),$$

where $\bar{c}(p) := ap/4$ *and* $\tilde{c}(p) := (3/4)ap\,\widetilde{v}_p(\bar{M}_p)$ *with* $\bar{M}_p := (1/3 + 4b/(3a) + 4d/(3a\beta) + 4(p-2)/(3a\beta))^{1/2}$.

Lemma C.16. *Let Assumption 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{max}$, $t \geq kT$, *with* $k \in \mathbb{N}$, *the following inequality holds*

$$\mathbb{E}[\widetilde{V}_2(\bar{\Phi}_t^{\lambda,k,fSGLD})] \leq e^{-\lambda t a/2}\mathbb{E}[\widetilde{V}_2(\theta_0)] + c_1(\lambda_{max} + a^{-1}) + 3\widetilde{v}_2(\bar{M}_2) + 1,$$

where c_1 *is given in Lemma C.13. In addition, the following inequality holds*

$$\mathbb{E}[\widetilde{V}_4(\bar{\Phi}_t^{\lambda,k,fSGLD})] \leq 2e^{-a\lambda t}\mathbb{E}[\widetilde{V}_4(\bar{\theta}_0^{fSGLD})] + 3\widetilde{v}_4(\bar{M}_4) + 2c_3(\lambda_{max} + a^{-1}) + 2,$$

where $\bar{M}_2$ *and* $\bar{M}_4$ *are given in Lemma C.15, and* c_3 *is given in Lemma C.13.*

Proof. This follows by applying Lemma C.13, Corollary C.14, and Lemma C.15 in Zhang et al. (2023, Proof of Lemma 4.5). □

Lemma C.17. *Let Assumption 1, 2 and 3 hold, and let* λ_{max} *be given in 24. Then, for any* $t > 0$,

$$\mathbb{E}\left[|\bar{\theta}_{\lfloor t \rfloor}^{fSGLD} - \bar{\theta}_t^{fSGLD}|^2\right] \leq \lambda\left[e^{-\lambda a \lfloor t \rfloor}\bar{\psi}_Y\mathbb{E}[\widetilde{V}_2(\theta_0)] + \widetilde{\psi}_Y\right],$$

where

$$\begin{aligned}
\bar{\psi}_Y &:= 2\lambda_{max}L_1^2\mathbb{E}[\varphi^2(X_0)], \\
\widetilde{\psi}_Y &:= 2c_1L_1^2\lambda_{max}\mathbb{E}[\varphi^2(X_0)](\lambda_{max} + a^{-1}) + 4\lambda_{max}L_2^2\mathbb{E}[\bar{\varphi}^2(X_0)] + 4\lambda_{max}\mathbb{E}[\widetilde{G}^2(\epsilon)] + 2d\beta^{-1},
\end{aligned} \tag{96}$$

with c_1 *given in Lemma C.13.*

Proof. This follows by applying Remark B.1 and Lemma C.13 in Zhang et al. (2023, Proof of Lemma A.2). □

Lemma C.18. *Let Assumption 1, 2 and 3 hold. For any* $t \in (kT, (k+1)T]$, *with* $k, N \in \mathbb{N}$ *and* $n = 1, \ldots, N+1$, *where* $N + 1 \leq T$, *one obtains*

$$\mathbb{E}[|\nabla g_\epsilon(\bar{\Phi}_t^{\lambda,k,fSGLD}) - \nabla_\theta U(\bar{\Phi}_t^{\lambda,k,fSGLD} + \epsilon_{kT+n}, X_{kT+n})|^2] \leq e^{-a\lambda t/2}\bar{\psi}_Z\mathbb{E}[\widetilde{V}_2(\theta_0)] + \widetilde{\psi}_Z,$$

where

$$\begin{aligned}
\bar{\psi}_Z &= 8L_2^2\mathbb{E}[(\varphi(X_0) + \varphi(\mathbb{E}[X_0]))^2|X_0 - \mathbb{E}[X_0]|^2], \\
\widetilde{\psi}_Z &= 8L_2^2E[(\varphi(X_0) + \varphi(\mathbb{E}[X_0]))^2|X_0 - \mathbb{E}[X_0]|^2](3\widetilde{v}_2(\bar{M}_2) + c_1(\lambda_{max} + a^{-1}) + 1 + \sigma^2 d),
\end{aligned} \tag{97}$$

with $\bar{M}_2$ *and* c_1 *given in Lemma C.15 and Lemma C.13, respectively.*

Proof. We adapt the Zhang et al. (2023, Proof of Lemma A.1). First, we define the filtration $\mathcal{J}_t = \mathcal{F}_\infty^\lambda \vee \mathcal{X}_{\lfloor t \rfloor} \vee \mathcal{H}_{\lfloor \epsilon \rfloor}$. Then, the result follows by an application of Lemma C.19, Remark B.1, and Lemma C.16

$$
\begin{aligned}
&\mathbb{E}\left[\left|\nabla g_\epsilon(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n})\right|^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left|\nabla g_\epsilon(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}) - \nabla_\theta U(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n})\right|^2 \middle| \mathcal{J}_{kT}\right]\right] \\
&= \mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg|\mathbb{E}\left[\nabla_\theta U(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n}) \middle| \mathcal{J}_{kT}\right] \\
&\qquad\qquad - \nabla_\theta U(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n})\Bigg|^2 \Bigg| \mathcal{J}_{kT}\Bigg]\Bigg] \\
&\leq 4\mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg|\nabla_\theta U(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, X_{kT+n}) \\
&\qquad\qquad - \nabla_\theta U(\bar{\Phi}_t^{\lambda,k,\text{fSGLD}} + \epsilon_{kT+n}, \mathbb{E}\left[X_{kT+n} | \mathcal{J}_{kT}\right])\Bigg|^2 \Bigg| \mathcal{J}_{kT}\Bigg]\Bigg] \\
&\leq 8L_2^2\mathbb{E}\left[(\varphi(X_0) + \varphi(\mathbb{E}[X_0]))^2 |X_0 - \mathbb{E}[X_0]|^2\right]\left(\sigma^2 d + \mathbb{E}\left[\left(1 + \left|\bar{\Phi}_t^{\lambda,k,\text{fSGLD}}\right|^2\right)\right]\right) \\
&\leq 8L_2^2\mathbb{E}\left[(\varphi(X_0) + \varphi(\mathbb{E}[X_0]))^2 |X_0 - \mathbb{E}[X_0]|^2\right] \\
&\quad \times \left(e^{-\lambda t a/2}\mathbb{E}[V_2(\theta_0)] + c_1(\lambda_{\max} + a^{-1}) + 3\widetilde{v}_2(\bar{M}_2) + 1 + \sigma^2 d\right).
\end{aligned}
$$

□

In the next lemma, L^p denotes the usual space of p-integrable real-valued random variables for $1 \leq p < \infty$.

Lemma C.19. *Let $\mathcal{F}, \mathcal{X}, \mathcal{H} \subset \mathcal{M}$ be sigma-algebras. Let X, Y be $\mathbb{R}^d$-valued random vectors in L^p for any $p \geq 1$ such that Y is measurable with respect to $\mathcal{F} \vee \mathcal{X} \vee \mathcal{H}$. Then,*

$$
\mathbb{E}^{1/p}\left[|X - \mathbb{E}[X|\mathcal{F} \vee \mathcal{X} \vee \mathcal{H}]|^p | \mathcal{X} \vee \mathcal{H}\right] \leq 2\mathbb{E}^{1/p}\left[|X - Y|^p | \mathcal{X} \vee \mathcal{H}\right].
$$

Proof. This follows by applying Chau et al. (2019, Lemma 6.1) to $\mathcal{F} \vee \mathcal{N}$, where the sigma-algebra $\mathcal{N} := \mathcal{X} \vee \mathcal{H}$. □

D Experimental details

D.1 Details for Section 4.2

D.1.1 Software and hardware environments

We conduct all experiments with Python 3.10.9 and Pytorch 1.13.1, CUDA 11.6.2, NVIDIA Driver 510.10 on Ubuntu 22.04.1 LTS server which equipped with AMD Ryzen Threadripper PRO 5975WX, NVIDIA A100 GPUs.

D.1.2 Implementation details

We follow standard data preprocessing and augmentation strategies as adopted in prior work (Li et al., 2017; Wei et al., 2022) on noisy-label benchmarks. For CIFAR-10N and CIFAR-100N, we apply random cropping with padding, random horizontal flipping, and normalization using dataset-specific statistics. For WebVision, we follow the preprocessing protocol of Kodge (2024). We note that Noisy-label benchmarks and ViT fine-tuning are widely used in evaluating the optimizer's generalization ability (Luo et al., 2024; Baek et al., 2024; Tan et al., 2025).

Regarding model architectures, we employ the CIFAR-specific variants of ResNet-34 and ResNet-50 when training on CIFAR-10N and CIFAR-100N, where the first convolution layer is replaced by a 3×3 kernel with stride 1 (instead of the 7×7 stride-2 convolution and max pooling used in ImageNet models) to accommodate the smaller 32×32 resolution. For WebVision, we adopt the standard ResNet implementations as provided for ImageNet-scale data.

For both training-from-scratch and fine-tuning experiments, we use the same hyperparameter search spaces. Table 4 summarizes the ranges considered for each optimizer. We do not employ any early stopping or pruning strategy during the Optuna-based hyperparameter tuning, ensuring that each trial is fully evaluated to its final epoch. We performed the same number of hyper-parameter trials for all methods so that the search-space exploration budget (number of trials) was identical. Because each SAM update requires two gradient evaluations, this design implies that, for the same number of trials and training epochs, SAM consumed roughly twice the wall-clock compute time of the other baselines. Thus our tuning protocol is at least as favorable to SAM as to the proposed fSGLD, ensuring that our reported improvements are not due to weaker tuning of SAM.

For SGLD and fSGLD (β fixed), we set a large inverse temperature $\beta = 10^{14}$. This follows the common heuristic of using a near-zero temperature to minimize exploration when employing Langevin Dynamics as a optimizer for a given objective. For fSGLD (β-σ coupled), we leverage our theoretical analysis as a practical tuning strategy. We only search for the optimal perturbation scale σ and then deterministically set β via our theoretically-derived relationship, $\beta = \sigma^{-4/(1+\eta)}$ with $\eta = 0.01$. This is a practical choice, as a larger η would cause β to become too small, allowing the Langevin noise term to overwhelm the gradient term and turning the dynamics into a near-random exploration. A small η thus ensures stable optimization. This principled approach significantly simplifies the search space.

Table 4: Hyperparameter search spaces for different optimizers.

Optimizer	**Learning rate**	**Momentum**	**Weight decay**	**Other hyperparameters**
SGD	$10^{[-2,0]}$	$\{0.1, 0.9\}$	5×10^{-4}	–
AdamW	$10^{[-4,-2]}$	–	10^{-2}	$[\beta_1, \beta_2] \in \{[0.8, 0.95], [0.99, 0.999]\}$
SGLD	$10^{[-2,0]}$	–	5×10^{-4}	$\beta = 10^{14}$
SAM	$10^{[-2,0]}$	$\{0.1, 0.9\}$	5×10^{-4}	$\rho \in 10^{[-3,-1]}$
fSGLD (β fixed)	$10^{[-2,0]}$	–	5×10^{-4}	$\beta = 10^{14}$, $\sigma \in 10^{[-3,-2]}$
fSGLD (β-σ coupled)	$10^{[-2,0]}$	–	5×10^{-4}	$\beta = \sigma^{-4/1.01}$, $\sigma \in 10^{[-3,-2]}$

D.2 Details for Section 4.5

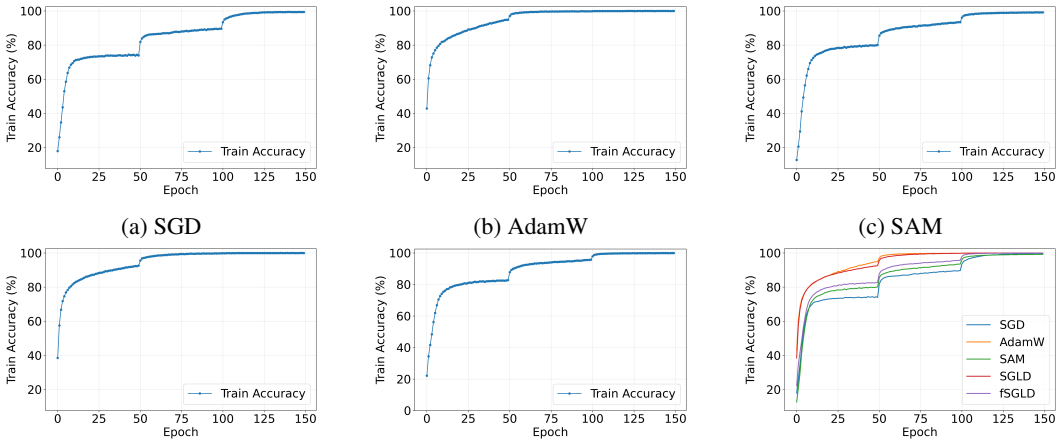
For the Hessian spectrum analysis, we use the best-performing ResNet-34 model trained on CIFAR-10N under each optimizer setting. Given a trained network f_θ and loss function L, we compute Hessian-vector products (HVPs) by applying automatic differentiation to the scalar product $\nabla_\theta L^\top v$ for a random vector v. For eigenvalue computation, we adopt the Lanczos algorithm (Lin et al., 2016) as implemented in `scipy.sparse.linalg.eigsh`, which allows us to approximate the top-k eigenvalues without explicitly forming the Hessian. In all reported results, we compute up to the top 50 eigenvalues. As a complementary measure of curvature, we estimate the trace of the Hessian using Hutchinson's stochastic estimator (Avron & Toledo, 2011) with Rademacher random vectors:

$$\mathrm{tr}(H(\theta)) \approx \frac{1}{m} \sum_{i=1}^{m} z_i^\top H(\theta) z_i, \quad z_i \sim \mathrm{Unif}\{\pm 1\}^d,$$

where $m = 1000$ in our experiments and d denotes the number of model parameters.

The analysis is conducted on the CIFAR-10N, where we randomly subsample at most 1,000 examples to reduce computational overhead. Eigenvalue computations are performed with a tolerance of 10^{-4} and a maximum of 500 iterations for the Lanczos solver.

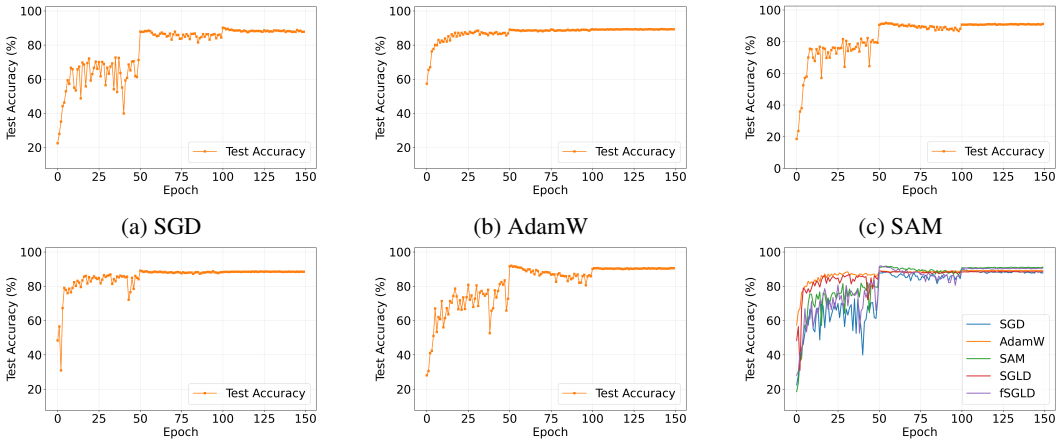
D.3 TRAINING CURVES



(a) SGD (b) AdamW (c) SAM

(d) SGLD (e) fSGLD (f) Comparision

Figure 4: Training accuracy trajectories of ResNet-34 on CIFAR-10N using the best hyperparameter settings for each optimizer.



(a) SGD (b) AdamW (c) SAM

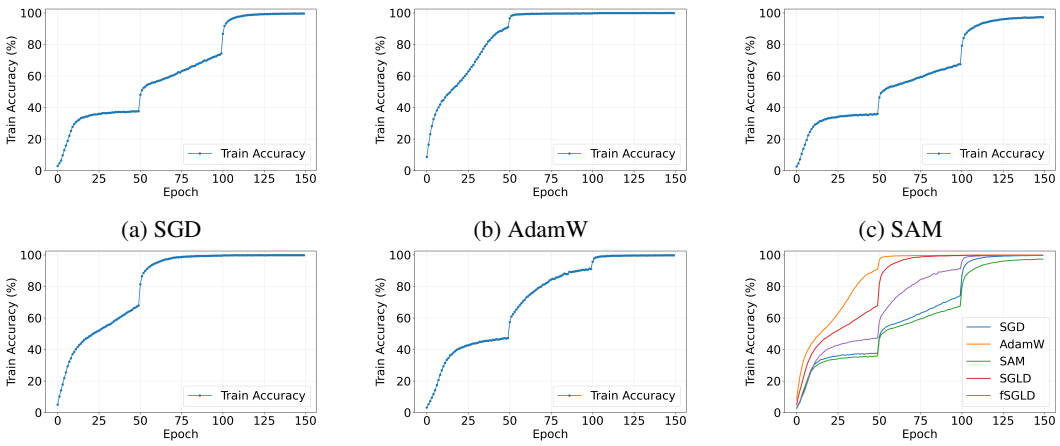
(d) SGLD (e) fSGLD (f) Comparision

Figure 5: Test accuracy trajectories of ResNet-34 on CIFAR-10N using the best hyperparameter settings for each optimizer.

Table 5: The best hyperparameter settings for ResNet-34 on CIFAR-10N.

Optimizer	**Learning rate**	**Momentum**	**Weight decay**	**Other hyperparameters**
SGD	8.69×10^{-1}	0.71	5×10^{-4}	–
AdamW	3.32×10^{-4}	–	10^{-2}	$[\beta_1, \beta_2] = [0.81, 0.99]$
SGLD	7.13×10^{-2}	–	5×10^{-4}	$\beta = 10^{14}$
SAM	5.51×10^{-1}	0.62	5×10^{-4}	$\rho = 2.69 \times 10^{-2}$
fSGLD	9.58×10^{-1}	–	5×10^{-4}	$\beta = \sigma^{-4/1.01},\ \sigma = 2.46 \times 10^{-3}$

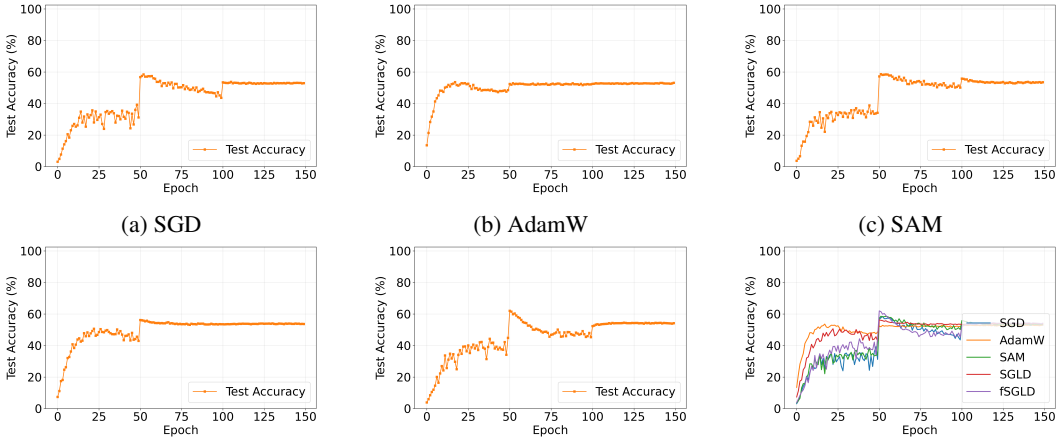
Figure 4 and Figure 5 present the training and test accuracy curves, respectively, for ResNet-34 trained on the CIFAR-10N dataset, utilizing the best hyperparameter configuration for each optimizer. The optimal hyperparameters are summarized in Table 5. Similarly, Figure 6 and Figure 7 illustrate the training and test accuracy trajectories for the CIFAR-100N dataset. The corresponding hyperparameter settings used for these experiments are detailed in Table 6.



(a) SGD (b) AdamW (c) SAM

(d) SGLD (e) fSGLD (f) Comparision

Figure 6: Training accuracy trajectories of ResNet-34 on CIFAR-100N using the best hyperparameter settings for each optimizer.



(a) SGD (b) AdamW (c) SAM

(d) SGLD (e) fSGLD (f) Comparision

Figure 7: Test accuracy trajectories of ResNet-34 on CIFAR-100N using the best hyperparameter settings for each optimizer.

Table 6: The best hyperparameter settings for ResNet-34 on CIFAR-100N.

Optimizer	**Learning rate**	**Momentum**	**Weight decay**	**Other hyperparameters**
SGD	7.76×10^{-1}	0.69	5×10^{-4}	–
AdamW	3.32×10^{-4}	–	10^{-2}	$[\beta_1, \beta_2] = [0.81, 0.99]$
SGLD	1.59×10^{-1}	–	5×10^{-4}	$\beta = 10^{14}$
SAM	7.77×10^{-1}	0.73	5×10^{-4}	$\rho = 8.84 \times 10^{-3}$
fSGLD	9.45×10^{-1}	–	5×10^{-4}	$\beta = \sigma^{-4/1.01}$, $\sigma = 2.47 \times 10^{-3}$

E USE OF LARGE LANGUAGE MODELS (LLMS).

In this manuscript, we used LLMs solely for writing assistance, such as grammar checking and minor language polishing. All sentences and substantive content of the paper are entirely our own. LLMs were not used for retrieval, discovery, research ideation, or any other purpose beyond basic language editing.