Title

On Large Batch Training and Sharp Minima: A Fokker—Planck Perspective

Journal

Journal of Statistical Theory and Practice, 14(3)

ISSN

1559-8608

Authors

Dai, Xiaowu

Zhu, Yuhua

Publication Date

2020-09-01

DOI

10.1007/s42519-020-00120-9

Journal of Statistical Theory and Practice (2020) 14:53
https://doi.org/10.1007/s42519-020-00120-9

ORIGINAL ARTICLE

On Large Batch Training and Sharp Minima: A Fokker–Planck Perspective

Xiaowu Dai[1] · **Yuhua Zhu**[2]

Published online: 26 July 2020

Abstract
We study the statistical properties of the dynamic trajectory of stochastic gradient descent (SGD). We approximate the mini-batch SGD and the momentum SGD as stochastic differential equations. We exploit the continuous formulation of SDE and the theory of Fokker–Planck equations to develop new results on the escaping phenomenon and the relationship with large batch and sharp minima. In particular, we find that the stochastic process solution tends to converge to flatter minima regardless of the batch size in the asymptotic regime. However, the convergence rate is rigorously proven to depend on the batch size. These results are validated empirically with various datasets and models.

Keywords Large batch training · Sharp minima · Fokker–Planck equation · Stochastic gradient algorithm · Deep neural network

Mathematics Subject Classification 90C15 · 35Q62 · 65K05

1 Introduction

We consider the following empirical risk minimization problem in statistical machine learning:

This article is part of the topical collection "Advances in Deep Learning" guest edited by David Banks, Ernest Fokoué, and Hailin Sang.

✉ Xiaowu Dai
xwdai@berkeley.edu

[1] CDAR and Simons Institute for the Theory of Computing, University of California, Berkeley, CA, USA

[2] Department of Mathematics, Stanford University, Stanford, CA, USA

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{n=1}^{N} L_n(\mathbf{w}),$$

where $\mathbf{w}$ represents the model parameters, $L_n(\mathbf{w})$ denotes the loss due to the n^{th} training sample, and N is the size of the training set. Since the training set for many application domains such as image (He et al.[12]) and speech recognition (Amodei et al.[1]) is large, the stochastic gradient descent (SGD) and its variants have become standard approaches of training complex model including deep neural networks (Bottou et al.[4]). The mini-batch SGD estimates the negative loss gradient based on a small subset of training examples. This approach incurs the computational complexity per iteration independent of N:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\gamma_k}{M_k}\sum_{n\in B_k} \nabla L_n(\mathbf{w}_k), \tag{1}$$

where $k \geq 0$, γ_k is the learning rate, and the mini-batch set B_k consists of M_k uniformly selected sample indices from $\{1, 2, \ldots, N\}$. A notable variant of mini-batch SGD is momentum SGD, which is a practical approach of speeding up the training (Nesterov[23]). For mini-batch SGD and its variant, we use the term *large batch training* to denote the use of a large mini-batch (Keskar et al.[17]).

Recently, several works have discussed the geometry of SGD (Keskar et al.[17]; Goyal et al.[11]; Hoffer et al.[14]). Specifically, Keskar et al.[17] find, based on empirical experiments, that the large batch training tends to converge to the sharp minima of the training function. In contrast, the small batch training is more likely to escape the sharp minima. In this work, we study theoretically and empirically the dynamic of the convergence and escaping phenomenon relating to the batch size for mini-batch SGD and momentum SGD.

We approximate SGD using continuous stochastic differential equation (SDE) (Chaudhari et al.[7]; Li et al.[21]; Mandt et al.[22]). Assuming isotropic gradient noise, we derive new results on the dynamic trajectory of the Fokker–Planck solution. In particular, the derived convergence rate in terms of the batch size provides new insights into the escaping phenomenon for mini-batch SGD and momentum SGD. Our main finding is that the stochastic process solution of SDE tends to converge to flatter minima regardless of the batch size in the asymptotic regime. However, the convergence rate depends on the batch size. Motivated by partial differential equation theory, we define the sharpness in terms of the Hessian's determinant. This result provides a new perspective into the ongoing discussion on the definition of the sharpness (e.g., Dinh et al.[9]). We verify our theoretical results experimentally on different datasets and deep neural network models. The proposed statistical view using tools from the Fokker–Planck equation can be used to analyze other stochastic algorithms for complex models.

The rest of the paper is organized as follows. We introduce the background in Sect. 2. We present our main result for mini-batch SGD in Sect. 3. We extend the result to momentum SGD in Sect. 4. We show numerical experiments in Sect. 5. Related works are provided in Sect. 6. We conclude the paper with discussions in Sect. 7. Proofs are given in "Appendix."

2 SDE Modeling for Large Batch Training

We study the dynamics of the mini-batch SGD in both the finite-time regime and the asymptotic regime, depending on whether the training time is finite or tends to infinity. It turns out that the dynamics of the finite-time regime given in this section are fundamentally different from those of the asymptotic regime to be discussed in Sects. 3–4.

The mini-batch SGD carries out the update at each step following (1), which can be rewritten as

$$\mathbf{w}_{k+1} - \mathbf{w}_k = -\gamma_k \nabla L(\mathbf{w}_k) + \frac{\gamma_k}{\sqrt{M_k}} \cdot \boldsymbol{\epsilon}_k, \tag{2}$$

where $L(\mathbf{w}) \equiv \mathbb{E}[L_n(\mathbf{w})]$ is the risk function and $\boldsymbol{\epsilon}_k = \frac{1}{\sqrt{M_k}} \sum_{n \in B_k} (\nabla L(\mathbf{w}_k) - \nabla L_n(\mathbf{w}_k))$ is a d-dimensional random vector. Assume that the covariance matrix $\mathrm{Var}[\nabla L_n(\mathbf{w})] \equiv \boldsymbol{\sigma}^2(\mathbf{w})$ is positive definite, which holds for typical loss functions, including the squared loss. By the dominated convergence theorem, $\boldsymbol{\epsilon}_k$ has mean 0 and covariance $\boldsymbol{\sigma}^2(\mathbf{w}_k)$ for any $k \geq 0$ (see, "Appendix A.1").

For the large batch training, the distribution of $\boldsymbol{\epsilon}_k$ is well approximated by the normal distribution due to the central limit theorem. Consider the following stochastic differential equation (SDE) model:

$$d\mathbf{W}(t) = -\nabla L(\mathbf{W}(t))\mathrm{d}t - \sqrt{\frac{\gamma(t)}{M(t)}}\boldsymbol{\sigma}(\mathbf{W}(t))d\mathbf{B}(t), \quad \mathbf{W}(0) = \mathbf{w}_0, \tag{3}$$

where the Brownian motion $\mathbf{B}(t)$ accounts for random fluctuations due to the use of mini-batches for gradient estimation in (2). The Euler discretization of SDE (3) resembles the mini-batch SGD (2), and the SDE solution approximates the mini-batch SGD in the weak sense (i.e., in distribution) under the finite-time setting $t \in [0, T]$ for any $T > 0$; see, e.g., Li et al.[21] and Mandt et al.[22] .

2.1 Escaping Phenomenon

Recently, Keskar et al.[17] note the escaping phenomenon of mini-batch SGD in training neural networks. Namely, the large batch training tends to converge to the sharp minima of the training function while the small batch training is more likely to *escape* the sharp minima. A conceptual sketch of "sharp" (and relatively, "flat") minima are shown in Fig. 1, where a mathematical definition of the sharpness is given in Sect. 2.3. Based on the numerical experiments, Keskar et al.[17] also find that a sharp minimum correlates with a worse generalization, which, however, will not be studied in the current paper.

The escaping phenomenon is important for understanding the algorithm design for complex statistics and machine learning models. The phenomenon has been validated in extensive numerical results; see, e.g., Goyal et al.[11] and Hoffer et al.[14].

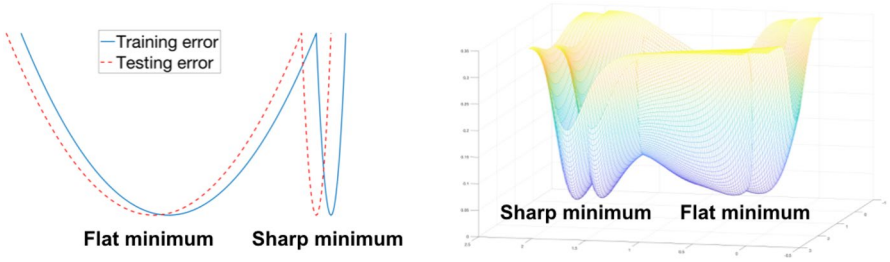

Fig. 1 Sketch of "flat" and "sharp" minima for one-dimensional case (left plot) and two-dimensional case (right plot). The vertical axis indicates the value of the loss function

However, the theoretical support for the phenomenon is limited in the literature. The current paper fills some gaps in this important direction. Our approach uses the SDE model (3) and studies the escaping phenomenon for the stochastic process solution to the SDE model.

2.2 Fokker–Planck Equation

We allow the learning rate γ_k and the batch size M_k in (2) to vary along with the step k, which is consistent with the practice. As a result, the functions $\gamma(t)$ and $M(t)$ in (3) are time-dependent. Consider the isotropic gradient covariance:

$$\boldsymbol{\sigma}^2(\mathbf{w}) = \beta(\mathbf{w}) \cdot \mathbf{I}, \tag{4}$$

where the scalar function $\beta(\mathbf{w})$ depends on $\mathbf{w}$. Similar assumptions as (4) have been made in the stochastic algorithm literature, for example, Chaudhari et al.[7] and Jastrzebski et al.[16], where $\beta(\mathbf{w}) \equiv \beta$ is restricted to a constant. Since our interest lies in the escaping phenomenon and the relationship with the scale of variance, the learning rate, and the batch size, we make the isotropic assumption (4) for simplicity and leave the anisotropic case for future study.

Denote by $p(\mathbf{w}, t)$ the probability density function of the stochastic process solution $\mathbf{W}(t)$. We can characterize $p(\mathbf{w}, t)$ in the following lemma, which is from the partial differential equations literature (e.g., Kolpas et al.[18]).

Lemma 1 *The probability density function $p(\mathbf{w}, t)$ satisfies the following Fokker–Planck equation*:

$$\partial_t p(\mathbf{w}, t) = \nabla \cdot \left(\left[\nabla \left(L(\mathbf{w}) + \frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} \right) \right] p(\mathbf{w}, t) + \frac{\gamma(t)\beta(\mathbf{w})}{2M(t)} \nabla p(\mathbf{w}, t) \right), \tag{5}$$

where $p(\mathbf{w}, 0) = \delta(\mathbf{w}_0)$, and $\delta(\cdot)$ denotes the Dirac's delta function.

We give a proof in "Appendix A.2." Note that the drift term in (5) $\nabla[L(\mathbf{w}) + \gamma(t)\beta(\mathbf{w})/2M(t)] \neq \nabla L(\mathbf{w})$, which implies that the stochastic process solution $\mathbf{W}(t)$ does not follow the mean drift direction $-\nabla L(\mathbf{w})$ as its update direction. A

smaller batch size $M(t)$ corresponds to a drift term deviates further from the mean drift direction.

2.3 Kramer's Formula

Based on the Fokker–Planck equation in Lemma 1, we can characterize the dynamics of the stochastic process solution in the finite-time regime. In particular, we have the escaping time of the stochastic process solution from one local minimizer $\check{\mathbf{w}}_1$ to its nearest local minimizer $\check{\mathbf{w}}_2$. Figure 2 gives an illustration, where $\mathbf{w}^*$ is the saddle point between $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$. There are possibly multiple saddle points between $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ in the multidimensional setting. The $\mathbf{w}^*$ should be defined as the saddle point with the minimal height among all saddle points in the following sense. Denote by $\mathbf{w}(t), 0 \le t \le 1$, be any continuous path from $\check{\mathbf{w}}_1$ to $\check{\mathbf{w}}_2$, and $\widehat{\mathbf{w}} = \arg\inf_{\mathbf{w}:\mathbf{w}(0)=\check{\mathbf{w}}_1,\mathbf{w}(1)=\check{\mathbf{w}}_2} \sup_{t\in[0,1]} L(\mathbf{w}(t))$ the path with the minimal saddle point height among all continuous path. Then, $\mathbf{w}^* \equiv \max_{t\in[0,1]} \widehat{\mathbf{w}}(t)$. It is known that the Hessian $\nabla^2 L(\mathbf{w}^*)$ has a single negative eigenvalue (e.g., Berglund[3]). Let $-\lambda^*$ be the negative eigenvalue of $\nabla^2 L(\mathbf{w}^*)$ and $H(\mathbf{w}^*, \check{\mathbf{w}}_1) \equiv L(\mathbf{w}^*) - L(\check{\mathbf{w}}_1)$ be the relative height of $\mathbf{w}^*$ to $\check{\mathbf{w}}_1$. We have the following lemma characterizing the escaping time of the stochastic process solution from $\check{\mathbf{w}}_1$ to $\check{\mathbf{w}}_2$.

Lemma 2 *Let $\tau_{\check{\mathbf{w}}_1\to\check{\mathbf{w}}_2}$ be the transition time for $\mathbf{W}(t)$ from a closed ball of radius $\epsilon > 0$ centered at $\check{\mathbf{w}}_1$ to a closed ball of radius $\epsilon > 0$ centered at $\check{\mathbf{w}}_2$. Then,*

$$\mathbb{E}[\tau_{\check{\mathbf{w}}_1\to\check{\mathbf{w}}_2}] = \frac{2\pi}{\lambda^*}\sqrt{\frac{|\nabla^2 L(\mathbf{w}^*)|}{|\nabla^2 L(\check{\mathbf{w}}_1)|}}\exp\left(\frac{H(\mathbf{w}^*,\check{\mathbf{w}}_1)\cdot 2M(\check{\mathbf{w}}_1)}{\gamma(\check{\mathbf{w}}_1)\beta(\check{\mathbf{w}}_1)}\right)\left[1+O\left(\sqrt{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)\right],$$

where $|\nabla^2 L(\cdot)|$ denotes the determinant of $\nabla^2 L(\cdot)$, $M(\check{\mathbf{w}}_1)$ is the batch size at $\check{\mathbf{w}}_1$, $\gamma(\check{\mathbf{w}}_1)$ is the learning rate at $\check{\mathbf{w}}_1$, and $\beta(\cdot)$ is defined in (4).

Similarly, we have the transition time from $\check{\mathbf{w}}_2$ to $\check{\mathbf{w}}_1$ (i.e., $\tau_{\check{\mathbf{w}}_2\to\check{\mathbf{w}}_1}$) with the only difference that the right side of the equation in Lemma 2 should be replaced by the geometry related to $\check{\mathbf{w}}_2$. This lemma is known in the diffusion process literature as the Eyring–Kramers formula; see, e.g., Berglund[3], Bovier et al.[5, 6]. If the radius of the ball centered at $\check{\mathbf{w}}_1$ is different from the radius of the ball centered at $\check{\mathbf{w}}_2$, Lemma 2 still holds by changing the radius ϵ to be the smaller one of the two radius (Bovier et al. [5]). Our observation is that the Eyring–Kramers formula can provide a quantitative

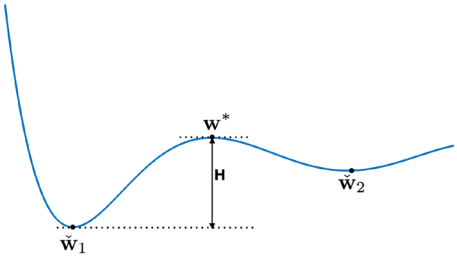

Fig. 2 Sketch of two local minimizer $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ of a risk function. The $\mathbf{w}^*$ is the saddle point between $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$

description of the escaping phenomenon in the *finite-time* regime. In particular, the time that $\mathbf{W}(t)$ escapes from one local minimum to its nearest local minimum depends on three factors. Namely, the diffusion factor $\gamma(\mathbf{w})\beta(\mathbf{w})/M(\mathbf{w})$, the potential barrier $H(\mathbf{w}^*, \check{\mathbf{w}}_1)$ that $\mathbf{W}(t)$ has to climb to escape $\check{\mathbf{w}}_1$, and the determinants of the Hessians of the risk function at $\check{\mathbf{w}}_1$ and $\mathbf{w}^*$. This fact suggests the following definition of the sharpness.

Definition 1 *(Sharpness)* The sharpness of a minimizer is defined as the determinant of the Hessian of the risk function at the minimizer, i.e., $|\nabla^2 L(\cdot)|$. A larger $|\nabla^2 L(\cdot)|$ corresponds to a sharper minimizer.

Lemma 2 shows that a larger batch size $M(\check{\mathbf{w}}_1)$ at a local minimizer $\check{\mathbf{w}}_1$ results in a longer time to escape from $\check{\mathbf{w}}_1$. If $\check{\mathbf{w}}_1$ corresponds to a sharp minimum with a large $|\nabla^2 L(\check{\mathbf{w}}_1)|$, the exponential term

$$\exp\left(\frac{H(\mathbf{w}^*, \check{\mathbf{w}}_1) \cdot 2M(\check{\mathbf{w}}_1)}{\gamma(\check{\mathbf{w}}_1)\beta(\check{\mathbf{w}}_1)}\right) \tag{6}$$

dominates the escaping time in the following sense. Compare the SGD training with two different batch sizes: $M_1(\check{\mathbf{w}}_1)$ and $M_2(\check{\mathbf{w}}_2)$ with $M_1(\check{\mathbf{w}}_1) > M_2(\check{\mathbf{w}}_2)$, and the same learning rate $\gamma(\check{\mathbf{w}}_1) = \gamma(\check{\mathbf{w}}_2)$. It takes a longer time for the large batch SGD to escape from $\check{\mathbf{w}}_1$ as compared with the small batch SGD to escape from $\check{\mathbf{w}}_2$, only if

$$\begin{aligned} &H(\mathbf{w}^*, \check{\mathbf{w}}_1)M(\check{\mathbf{w}}_1) - H(\mathbf{w}^*, \check{\mathbf{w}}_2)M(\check{\mathbf{w}}_2) \\ &> \frac{\gamma(\check{\mathbf{w}}_1)\beta(\check{\mathbf{w}}_1)}{4}\left[\log(|\nabla^2 L(\check{\mathbf{w}}_1)|) - \log(|\nabla^2 L(\check{\mathbf{w}}_2)|)\right]. \end{aligned}$$

That is, the exponential term (6) makes the effect of sharpness on the escaping time on a logarithm scale compared with the effect of batch size. A local minimizer of the risk function lies in a closed ball of a local minimizer of the training function. This result shows that *large batch training is more likely to be trapped at sharp minima of the training function in the finite-time regime than small batch training*. On the other hand, if the batch size $M(\check{\mathbf{w}}_1)$ decreases, the exponential term (6) decreases, and the stochastic process solution $\mathbf{W}(t)$ will be trapped at $\check{\mathbf{w}}_1$ only when the determinant $|\nabla^2 L(\check{\mathbf{w}}_1)|$ is small enough, as shown in Lemma 2. In words, it explains the escaping phenomenon that *small batch training tends to escape sharp minima and converge to flat minima*.

The escaping phenomenon in the asymptotic regime is different from that in the finite-time regime. However, the Eyring–Kramers formula fails when $t \to \infty$. We develop a new theory for the asymptotic regime in the following Sect. 3 and extend the result for momentum SGD-related SDE in Sect. 4.

3 Convergence Properties for Large Batch Training

We study the stochastic process solution $\mathbf{W}(t)$ of the SDE (3) in the asymptotic regime (i.e., $t \to \infty$).

3.1 Main Assumptions

The main assumptions are outlined as follows.

(A.1) The risk function $L(\mathbf{w})$ is confinement in the sense that

$$\lim_{\|\mathbf{w}\| \to +\infty} L(\mathbf{w}) = +\infty, \quad \int \mathrm{e}^{-L(\mathbf{w})} d\mathbf{w} < +\infty.$$

(A.2) Denote by $\mathrm{Tr}(\nabla^2 L)$ the trace of the Hessian of L. Assume

$$\lim_{\|\mathbf{w}\| \to +\infty} \left\{ \frac{1}{2} \|\nabla L(\mathbf{w})\|^2 - \mathrm{Tr}(\nabla^2 L(\mathbf{w})) \right\} = +\infty,$$
$$\lim_{\|\mathbf{w}\| \to +\infty} \left\{ \mathrm{Tr}(\nabla^2 L(\mathbf{w})) / \|\nabla L(\mathbf{w})\|^2 \right\} = 0.$$

(A.3) There exists a constant $M_{\mathbf{w}}$, such that

$$\left| \mathrm{e}^{-L(\mathbf{w})} \left(\|\nabla L(\mathbf{w})\|^2 - \mathrm{Tr}(\nabla^2 L(\mathbf{w})) \right) \right| \leq M_{\mathbf{w}}.$$

Assumptions (A.1)–(A.3) are common in the diffusion process literature, see, e.g., Pavliotis[24]. We show in "Appendix B.1" that (A.1)–(A.3) hold for typical loss functions, including the regularized mean cross-entropy and the squared loss. In particular, Assumption (A.1) ensures that the Gibbs density function $\mathrm{e}^{-L(\mathbf{w})}$ is well defined. Assumption (A.2) guarantees the measure $\mu(\mathbf{w}) = \int \mathrm{e}^{-L(\mathbf{w})} d\mathbf{w}$ satisfying the Poincaré inequality (see, Pavliotis[24]):

$$\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu(\mathbf{w}) \geq C_{\mathrm{P}} \int \left(f(\mathbf{w}) - \int f(\mathbf{w}) \mathrm{d}\mu(\mathbf{w}) \right)^2 \mathrm{d}\mu(\mathbf{w}) \tag{7}$$

with some $C_{\mathrm{P}} > 0$, where f is an integrable function satisfying $\int f^2(\mathbf{w}) \mathrm{d}\mathbf{w} < \infty$.

Lemma 3 *Under Assumption (A.1) and $\beta(\mathbf{w}) \equiv \beta$, the Fokker–Planck equation (5) has a stationary solution in the asymptotic regime (i.e., $t \to \infty$):*

$$p_\infty(\mathbf{w}) = \kappa \mathrm{e}^{-\frac{2M(\infty)L(\mathbf{w})}{\gamma(\infty)\beta}},$$

where κ is a normalization constant such that $\int p_\infty(\mathbf{w}) d\mathbf{w} = 1$, and the limiting batch size and learning rate are defined as $M(\infty) \equiv \lim_{t\to\infty} M(t)$ and $\gamma(\infty) \equiv \lim_{t\to\infty} \gamma(t)$, respectively.

A derivation of Lemma 3 is provided in "Appendix B.2." We remark that for a general $\beta(\mathbf{w})$, which depends on $\mathbf{w}$, the existence and an explicit form of the stationary solution to the Fokker–Planck equation (5) remains an open question. We focus on $\beta(\mathbf{w}) \equiv \beta$ in this section.

3.2 Escaping Phenomenon in the Asymptotic Regime

Related works on the analysis of stochastic algorithms have studied the stationary solution $p_\infty(\mathbf{w})$; see, e.g., Jastrzebski et al.[16]. However, it is unclear whether the density function $p(\mathbf{w}, t)$ converges to the stationary solution $p_\infty(\mathbf{w})$, not to mention the convergence rate. Theorem 1 gives an affirmative answer to this problem, and it also provides new insights into the escaping phenomenon and the relationship with large batch and sharp minima.

Theorem 1 *Under Assumptions (A.1)–(A.3), the density function $p(\mathbf{w}, t)$ of $\mathbf{W}(t)$ converges to the stationary solution $p_\infty(\mathbf{w})$. Moreover, there exists $T > 0$ such that for any $t > T$,*

$$\left\| \frac{p(\mathbf{w}, t) - p_\infty(\mathbf{w})}{\sqrt{p_\infty(\mathbf{w})}} \right\|^2_{L^2(\mathbb{R}^d)} \leq C(t, T) \mathrm{e}^{-\frac{C_P \cdot (t-T) \cdot \gamma(\infty)\beta}{2M(\infty)}},$$

where the constant C_P is defined in (7), and the function $C(t, T)$ is given by

$$C(t, T) \equiv \frac{C_P \cdot (t - T) \cdot \gamma(\infty)\beta}{2M(\infty)} + \left\| \frac{p(\mathbf{w}, T) - p_\infty(\mathbf{w})}{\sqrt{p_\infty(\mathbf{w})}} \right\|^2_{L^2(\mathbb{R}^d)}.$$

Theorem 1 is new in the literature, and its proof is given in "Appendix B.3." We also provide a quantification of the constant T in "Appendix B.4." We make three remarks for Theorem 1. First, the theorem verifies that $p(\mathbf{w}, t)$ converges to the stationary solution $p_\infty(\mathbf{w})$ with an exponential convergence rate regardless of the initial value. This result provides theoretical support for related works that analyze the density function $p(\mathbf{w}, t)$ based on analysis of the stationary distribution $p_\infty(\mathbf{w})$, for example, Jastrzebski et al.[16]. Second, large batch training with increasing batch size converges exponentially slower. Finally, there exists a trade-off in choosing the batch size and learning rate, since the convergence rate $\exp(-C_P \cdot (t - T) \cdot \gamma(\infty)\beta / 2M(\infty))$ depends on the batch size M and the learning rate γ.

From Theorem 1, we can also characterize the limiting behavior of $\mathbf{W}(t)$ in the asymptotic regime when $t \to \infty$.

Theorem 2 *Let $\check{\mathbf{w}}$ be a local minimizer. Then,*

$$\lim_{\epsilon\to 0}\mathbb{P}(|\mathbf{W}(\infty)-\check{\mathbf{w}}|\le\epsilon)$$
$$=\frac{\kappa e^{-4M(\infty)L(\check{\mathbf{w}})/[\gamma(\infty)\beta]}}{[2M(\infty)/\gamma(\infty)\beta]^{d/2}|\nabla^2 L(\check{\mathbf{w}})|}\lim_{\epsilon\to 0}\left[e^{\frac{2M(\infty)\epsilon^2}{\gamma(\infty)\beta}}\prod_{j=1}^{d}\sqrt{1-\exp\left(-\frac{2M(\infty)\epsilon^2\lambda_j}{\pi\gamma(\infty)\beta}\right)}\right],$$

where $\mathbf{w}\in\mathbb{R}^d$, λ_j*'s are eigenvalues of the Hessian* $\nabla^2 L(\check{\mathbf{w}})$, *and* $|\nabla^2 L(\check{\mathbf{w}})|$ *is the determinant of* $\nabla^2 L(\check{\mathbf{w}})$. *The constants* κ *and* $\gamma(\infty)$ *are defined in Lemma 3.*

The proof of Theorem 2 is given in "Appendix B.5." To better appreciate Theorem 2, we consider two local minimizers $\check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2$ which have the same value of $L(\check{\mathbf{w}}_1)=L(\check{\mathbf{w}}_2)$. Theorem 2 implies that

$$\lim_{\epsilon\to 0}\frac{\mathbb{P}(|\mathbf{W}(\infty)-\check{\mathbf{w}}_1|\le\epsilon)}{\mathbb{P}(|\mathbf{W}(\infty)-\check{\mathbf{w}}_2|\le\epsilon)}=\sqrt{\frac{|\nabla^2 L(\check{\mathbf{w}}_2)|}{|\nabla^2 L(\check{\mathbf{w}}_1)|}},\tag{8}$$

where the derivation is given in "Appendix B.6." Then, Eq. (8) suggests that in the asymptotic regime (i.e., $t\to\infty$), the probability of the stochastic process solution $\mathbf{W}(t)$ converging to a minimum with small determinant $|\nabla^2 L(\cdot)|$ is larger than that of converging to a minimum with large determinant $|\nabla^2 L(\cdot)|$. In words, by Definition 1, $\mathbf{W}(t)$ is more likely to converge to flatter minima. Moreover, the ratio in (8) does not depend on the batch size or learning rate. The ratio only depends on the determinant of the Hessian at the minimum.

Theorems 1 and 2 provide new insights into the escaping phenomenon in Sect. 2.1. Namely, the stochastic process solution $\mathbf{W}(t)$ tends to converge to flatter minima regardless of the batch size M in the asymptotic regime $t\to\infty$. However, the *convergence rate* depends on the batch size. We provide experiments in Sect. 5 to corroborate these findings for mini-batch SGD with various datasets and neural network models.

4 SDE Modeling for Momentum SGD

Momentum SGD (MSGD) is an effective approach of speeding up the mini-batch SGD; see, e.g., Qian[26], Nesterov[23], Sutskever et al.[29]. Instead of updating $\mathbf{w}_k$ directly in (1), MSGD adopts the following coupled updates:

$$\mathbf{z}_{k+1}=\xi\cdot\mathbf{z}_k-\frac{\gamma_k}{M_k}\sum_{n\in B_k}\nabla L_n(\mathbf{w}_k),$$
$$\mathbf{w}_{k+1}=\mathbf{w}_k+\mathbf{z}_{k+1}.$$

where ξ is the momentum parameter taking values in the range $0<\xi<1$. In this section, we focus on the constant learning rate and batch size: $\gamma_k\equiv\gamma, M_k\equiv M$ and leave the time-dependent case for future study. Let $\mathbf{v}_k=\mathbf{z}_k/\sqrt{\gamma}$. When the step size is small, $(\mathbf{v}_k,\mathbf{w}_k)$ can be approximated by the SDE (see, e.g., Li et al.[21], An et al. [2]),

$$\begin{cases} d\mathbf{V}(t) = -\nabla L(\mathbf{W}(t))\mathrm{d}t - \dfrac{1-\xi}{\sqrt{\gamma}}\mathbf{V}(t)\mathrm{d}t + \dfrac{\gamma^{1/4}}{\sqrt{M}}\sqrt{\beta(W(t))}d\mathbf{B}(t), \\ d\mathbf{W}(t) = \mathbf{W}(t)\mathrm{d}t. \end{cases}$$

where $\beta(\mathbf{w})$ is the scale of the covariance function defined in 4. The SDE modeling gives $\mathbf{V}(k\sqrt{\gamma}) \approx \mathbf{v}_k$, $\mathbf{W}(k\sqrt{\gamma}) \approx \mathbf{w}_k$, which is shown in "Appendix C.1".

4.1 Vlasov–Fokker–Planck Equation

Denote by $\psi(\mathbf{w}, \mathbf{v}, t)$ the joint probability density function of $(\mathbf{W}(t), \mathbf{V}(t))$. We have the following characterization of $\psi(\mathbf{w}, \mathbf{v}, t)$ from the partial differential equations literature (e.g., Pavliotis[24]) and also show the corresponding stationary solution.

Lemma 4 *The probability density function $\psi(\mathbf{w}, \mathbf{v}, t)$ satisfies the following Vlasov-Fokker–Planck equation*:

$$\begin{aligned} &\partial_t \psi(\mathbf{w}, \mathbf{v}, t) + \mathbf{v} \cdot \nabla_{\mathbf{w}} \psi(\mathbf{w}, \mathbf{v}, t) - \nabla L(\mathbf{w}) \cdot \nabla_{\mathbf{v}} \psi(\mathbf{w}, \mathbf{v}, t) \\ &\qquad = \nabla_{\mathbf{v}} \cdot \left(\frac{1-\xi}{\sqrt{\gamma}} \mathbf{v} \psi(\mathbf{w}, \mathbf{v}, t) + \frac{\sqrt{\gamma}\beta}{2M} \nabla_{\mathbf{v}} \psi(\mathbf{w}, \mathbf{v}, t) \right). \end{aligned} \tag{9}$$

Moreover, under Assumption (A.1) and $\beta(\mathbf{w}) \equiv \beta$, Eq. (9) has a stationary solution in the asymptotic regime (i.e., $t \to \infty$):

$$\psi_\infty(\mathbf{w}, \mathbf{v}) = \kappa' \mathrm{e}^{-\frac{2M}{\gamma\beta}(1-\xi)\left(L(\mathbf{w}) + \frac{|\mathbf{v}|^2}{2}\right)},$$

where κ' is a normalization constant such that $\int \psi_\infty d\mathbf{w} d\mathbf{v} = 1$.

We give a proof in "Appendix C.2." By integrating $\psi_\infty(\mathbf{w}, \mathbf{v})$ over $\mathbf{v}$, we obtain that $\int \psi_\infty(\mathbf{w}, \mathbf{v}) d\mathbf{v} = \kappa \mathrm{e}^{-\frac{2M}{\gamma\beta}(1-\xi)L(\mathbf{w})}$, which is similar to the stationary solution in Lemma 3 and implies Eq. (8) for MSGD. Hence, the stochastic process $\mathbf{W}(t)$ for MSGD-related SDE tends to converge to flatter minima regardless of the batch size in the asymptotic regime $t \to \infty$. However, we show in Sect. 4.2 that the convergence rate depends on the batch size.

4.2 Escaping Phenomenon of MSGD-Related SDE

In this section, we require an additional assumption.

(A.4) There exists a constant C_L such that the absolute values of eigenvalues of the matrix $\{\|(\nabla^2 \tilde{L})_{ij}\|_\infty\}_{1 \le i,j \le d}$ are bounded by a constant $b > 0$, where $\tilde{L}(\mathbf{w}) = L(\mathbf{w}) - \frac{1}{2} C_L^2 \|\mathbf{w}\|^2$ and $\{\|(\nabla^2 \tilde{L})_{ij}\|_\infty\}_{1 \le i,j \le d}$ consists of the (i, j)th entry $\|(\nabla^2 \tilde{L})_{ij}\|_\infty \equiv \sup_{\mathbf{w}} |(\nabla^2 \tilde{L})_{ij}|$.

We prove in "Appendix C.3" that Assumption (A.4) holds for typical loss functions including the regularized mean cross-entropy and the squared loss.

Theorem 3 *Under Assumption (A.1)–(A.4), the density function* $\psi(\mathbf{w},\mathbf{v},t)$ *of* $(\mathbf{W}(t),\mathbf{V}(t))$ *converges to the stationary solution* $\psi_\infty(\mathbf{w},\mathbf{v})$. *Moreover, there exists* $T>0$ *such that for any* $t>T$,

$$\left\|\frac{\psi(\mathbf{w},\mathbf{v},t)-\psi_\infty(\mathbf{w},\mathbf{v})}{\sqrt{\psi_\infty(\mathbf{w},\mathbf{v})}}\right\|^2_{L^2(\mathbb{R}^{2d})} \le \frac{\gamma\beta}{2M\min\{C_P,d\}(1-\xi)\lambda_{\min}}\mathrm{e}^{-2(\mu-\hat{\mu})t}H(0).$$

The parameters are specified as follows. First, C_P *is the Poincaré constant defined in* (7). *Define*

$$h(\mathbf{w},\mathbf{v},t)\equiv\frac{\psi(\mathbf{w},\mathbf{v},t)-\psi_\infty(\mathbf{w},\mathbf{v})}{\psi_\infty(\mathbf{w},\mathbf{v})} \quad \text{and matrix} \quad P\equiv\begin{bmatrix} I_d & \hat{C}I_d \\ \hat{C}I_d & CI_d\end{bmatrix},$$

where the constants C *and* $\hat{C}$ *together of the decay rate* μ *are determined by*

$$\begin{cases} \text{if } \dfrac{1-\xi}{\sqrt{\gamma}}<2C_L: \mu\equiv\dfrac{1-\xi}{\sqrt{\gamma}},\ C\equiv C_L^2,\ \hat{C}\equiv\dfrac{1-\xi}{2\sqrt{\gamma}}; \\ \text{if } \dfrac{1-\xi}{\sqrt{\gamma}}\ge 2C_L: \mu\equiv\dfrac{1-\xi}{\sqrt{\gamma}}-\sqrt{\dfrac{(1-\xi)^2}{\gamma}-4C_L^2},\ C\equiv\dfrac{(1-\xi)^2}{2\gamma}-C_L^2,\ \hat{C}\equiv\dfrac{1-\xi}{2\sqrt{\gamma}}. \end{cases}$$

Next, let $\lambda_{\min}$ *be the smallest eigenvalue of the matrix* P. *Finally, let*

$$H(0)=\int[\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{v},0),\nabla_{\mathbf{v}}h(\mathbf{w},\mathbf{v},0)]^\top P[\nabla_{\mathbf{w}}h(\mathbf{w},\mathbf{v},0),\nabla_{\mathbf{v}}h(\mathbf{w},\mathbf{v},0)]\psi_\infty d\mathbf{w}d\mathbf{v},$$

and $\hat{\mu}=\dfrac{(1+\sqrt{2})b}{2\lambda_{\min}}$, *where* b *the upper bound defined in Assumption (A.4),*

From Theorem 3, it is clear that the large batch training (i.e., as M increases) has a slower convergence as compared with the small batch training. Proof of this theorem is given in "Appendix C.4." Theorem 3 is new in the literature, and it builds on the result for quadratic function $L(\mathbf{w})=\frac{1}{2}C_L^2\|\mathbf{w}\|^2$ in the literature (e.g., Pavliotis[24]). Theorem 3 applies to general loss functions, including the regularized mean cross-entropy and the squared loss.

The main difficulty in the proof is that Eq. (9) is a degenerate diffusion PDE in the sense that it only has the diffusion on the $\mathbf{v}$ direction without the diffusion on the $\mathbf{w}$ direction; see an overview on the diffusion PDE in Evans[10]. We use the tools from *hypocoercivity* (Villani[30]), which links a degenerate diffusion operator and a conservative operator. The key idea in the proof is to construct a Lyapunov functional $H(t)$ (Villani[30]):

$$H(t)=\|\nabla_{\mathbf{w}}h\|_*^2+C\|\nabla_{\mathbf{v}}h\|_*^2+2\hat{C}\langle\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h\rangle_*,$$

where $\langle h, g\rangle_* \equiv \int hg\psi_\infty d\mathbf{w}d\mathbf{v}$ and $\|h\|_*$ is the corresponding norm. The above equation can be equivalently written as,

$$H(t) = \int [\nabla_{\mathbf{w}} h, \nabla_{\mathbf{v}} h]^\top P[\nabla_{\mathbf{w}} h, \nabla_{\mathbf{v}} h]\psi_\infty d\mathbf{w}d\mathbf{v}, \quad \text{with } P = \begin{bmatrix} I_d & \hat{C}I_d \\ \hat{C}I_d & CI_d \end{bmatrix}. \quad (10)$$

where $C, \hat{C}$ are constants to be determined. Note that

$$\frac{\mathrm{d}}{\mathrm{d}t}H(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\left\|\nabla_{\mathbf{w}} h\right\|_*^2 + C\left\|\nabla_{\mathbf{v}} h\right\|_*^2\right) + 2\hat{C}\frac{\mathrm{d}}{\mathrm{d}t}\langle \nabla_{\mathbf{w}} h, \nabla_{\mathbf{v}} h\rangle_*,$$

which implies following inequality with some constant $\tilde{C}$,

$$d_t H(t) + \tilde{C}H(t) \le 0,$$

and the exponential decay of $H(t)$. Finally, the relationship between $H(t)$ and $h(t)$ in (10) leads to the exponential decay for $\|h(t)\|_*^2$ as required for Theorem 3.

5 Numerical Experiments

We perform experiments using various datasets and deep learning models to corroborate theoretical findings in Sects. 2–4.

5.1 Escaping Phenomenon for Mini-Batch SGD

We consider three different neural network models: (1) a four-layer multilayer perception (MLP) with ReLU activation function and batch normalization (Ioffe and Szegedy[15]); (2) a shallow convolutional network N1; and (3) a deep convolutional network N2. The N1 network is a modified AlexNet configuration (Krizhevsky et al. [19]), and the N2 network is a modified VGG configuration (Simonyan and Zisserman[27]). We test and train the MLP with the MNIST dataset (LeCun et al.[20]), and N1 and N2 with the CIFAR-10 dataset, using the mean cross-entropy as the loss function. Details on the networks and dataset are given in "Appendix D." We study the escaping phenomenon of the mini-batch SGD with four pairs of learning rate and batch size: $(\gamma, M) = (0.1, 64)$, $(0.1, 128)$, $(0.2, 256)$, $(0.2, 512)$. A total of 100 epochs for each (γ, M) are trained, where the training loss stops decreasing. We repeat each experiment 100 times and average the results in Fig. 3. Due to the high computational cost for computing the determinant of Hessian, we use the Frobenius norm of Hessian as a substitute, similar to Wu et al.[31]. A smaller y-value in Fig. 3 indexes a flatter minimum. The x-axis denotes the number of steps, which equals the number of epoch $* N/M$, where N is the training sample size, and M is the batch size.

Figure 3 shows that under the same learning rate, the large batch training converges to sharper minima, for example, comparing the red solid curves with the blue dashed curves for all three plots, which agrees with Lemma 2. The mini-batch SGD

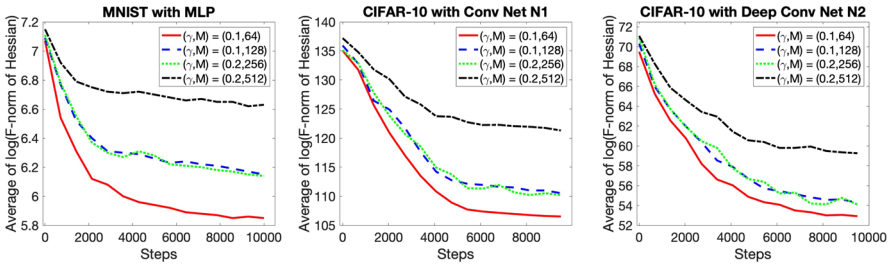

Fig. 3 Log of Frobenius norm of Hessian as a function of steps. The left plot is four-layer batch-normalized MLPs with MNIST dataset. The middle plot is convolutional network N1 with CIFAR-10 dataset. The right plot is deep convolutional network N2 with CIFAR-10 dataset. Four (γ, M) pairs are studied: (0.1, 64), (0.1, 128), (0.2, 256), and (0.2, 512), which are denoted in red, blue, green, and black, respectively. The plots show the averaged results of 100 experiments for each of the four (γ, M) pairs

with the same γ/M ratio follows a similar dynamic trajectory in terms of sharpness, consistent with the result of the SDE modeling in Lemma 3.

Theorem 1 shows that the large batch training converges to a flat minimum slower compared with the small batch training in the asymptotic regime. This is clear from Fig. 3. For example, the black dash–dot curve in the right plot takes 10,000 steps to converge at a minimum of $y = 58$, while the green dotted curve only takes 4000 steps to achieve it. On the other hand, for any batch size, SGD is more likely to saturate with a flatter minimum. For example, in the average case, the black curve in the right plot explores minima with a y-values range from 58 to 71 while it ends up with a minimum of $y = 58$, which corroborates Theorem 2.

5.2 Escaping Phenomenon for Momentum SGD

We empirically study the escaping phenomenon for momentum SGD. We use the three neural network models as Sect. 5.1: MLP, convolutional network N1, and deep convolutional network N2, which are trained with MNIST, CIFAR-10, and CIFAR-10, respectively. We consider four pairs of momentum parameter and batch size: $(\xi, M) = (0.9, 64), (0.9, 128), (0.99, 64), (0.99, 128)$, while the learning rate is $\gamma = 0.1$. A total of 100 epochs for each (ξ, M) are trained, where the training loss stops decreasing near the ending of the training. We repeat each experiment 100 times and average the results in Fig. 4.

Figure 4 shows that under the same momentum parameter, the large batch training converges to sharp minima. In the asymptotic regime, Theorem 3 shows that the large batch training converges to a flat minimum slower compared to the small batch training, which is clear from Fig. 4. For example, the blue dashed curve in the middle plot takes 8000 steps to converge at a minimum of $y = 110$, while the red solid curve only takes 2000 steps to achieve it. This phenomenon is robust to the momentum parameter (e.g., $\xi = 0.9$ or 0.99). On the other hand, Theorem 3 suggests there is no monotonic rule for tuning the momentum parameter ξ since both $\lambda_{\min}$ and μ depend on ξ. We observe a similar pattern in Fig. 4. While $\xi = 0.99$ leads the

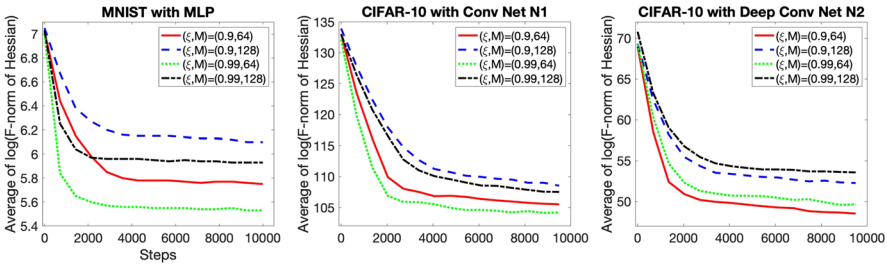

Fig. 4 Log of Frobenius norm of Hessian as a function of steps. The left plot is four-layer batch-normalized MLPs with MNIST dataset. The middle plot is convolutional network N1 with the CIFAR-10 dataset. The right plot is deep convolutional network N2 with the CIFAR-10 dataset. Four (ξ, M) pairs are studied: (0.9, 64), (0.9, 128), (0.99, 64), and (0.99, 128), which are denoted in red, blue, green, and black, respectively. The plots show the averaged results of 100 experiments for each of the four (ξ, M) pairs

momentum SGD to converge to flatter minima for MLP and N1 networks, $\xi = 0.99$ ends up with sharper minima for N2.

6 Related Work

Our work continues the line of research on the geometry of SGD, see, for example, Bottou et al.[4] for a comprehensive review. In particular, our interest lies in the role of large batch size and the sharpness of minima found in terms of generalization; see, e.g., Goyal et al.[11], Hoffer et al.[14] and Keskar et al.[17]. In particular, Keskar et al.[17] find, based on empirical experiments, that the large batch training tends to converge to a sharp minimum. Goyal et al.[11] and Hoffer et al.[14] observed through experiments that training for more epochs and scaling up the learning rate give good generalization when using large batch size. This paper is complementary to the existing works in this direction. Motivated by partial differential equation theory, we define the sharpness in terms of the determinant of the Hessian, which provides a new perspective into the discussion on the definition of the sharpness (e.g., Dinh et al.[9]). We explain the dynamic of the convergence and escaping phenomenon theoretically and empirically relating to the batch size for mini-batch SGD and momentum SGD.

Several authors have developed the relationship between SGD and sampling a posterior distribution via stochastic Langevin methods; see, e.g., Chaudhari et al.[7], Mandt et al.[22]. In particular, Mandt et al.[22] study SGD using an approximate Bayesian inference method in a locally convex setting. The modeling of SGD as a continuous-time stochastic process can also be achieved using SDE; see, e.g., Chaudhari and Soatto[8], Li et al.[21] and Smith and Le[28]. In particular, Li et al.[21] rigorously derive an approximation error of SDE solution to SGD in the finite-time regime. Smith and Le[28] use Bayesian principles to relate the generalization error with the batch size. Chaudhari and Soatto[8] discuss the stationary non-equilibrium solution for the stochastic differential equation. They allow the gradient

noise to be non-isotropic but require additional conditions to enforce the stationary distribution to be path-independent. Instead, we strictly focus on the convergence rate of the SDE solution to the stationary distribution with isotropic noise. This approach allows us to explore the dynamics of the convergence relating to the batch size and sharp minima. The results are verified empirically with various datasets and deep neural network models.

We discuss the Fokker–Planck equation and its variant, which modelings have appeared in the machine learning literature. Heskes and Kappen[13] derive a Gibbs distribution in the online setting. Jastrzebski et al.[16] discuss how the width and height of minima correlate with the learning rate to batch size ratio, but they focus on the stationary equilibrium distribution. Our result also shows that the ratio of learning rate to batch size correlates with the sharpness of minima (e.g., Lemma 3) in the stationary solution. In contrast to other work, we derive new results on the dynamic trajectory of the Fokker–Planck solution, including the convergence rate in terms of the batch size, which provides new insights into the escaping phenomenon for mini-batch SGD and momentum SGD.

7 Conclusion

We study the convergence rate of the SDE solution to the stationary distribution, which is new in the literature. It allows us to explore the dynamics of the escaping phenomenon and the relationship with the batch size and sharp minima. The perspective from the Fokker–Planck equation and its variant provide novel insights into the escaping phenomenon for mini-batch SGD and momentum SGD. Namely, the stochastic process solution tends to converge to flatter minima regardless of the batch size in the asymptotic regime. However, the convergence rate depends on the batch size. These results are validated theoretically and empirically with various datasets and deep neural network models.

We made the isotropic assumption on the covariance of the gradients, which is to derive a closed-form for the convergence rate of the SDE solution to the stationary distribution. It is of interest to study whether the practical techniques such as batch normalization would give a covariance of the gradients close to the isotropy. We also leave the study of extending this paper to anisotropic covariance structure for future work. Finally, the derived asymptotic dynamic reflects the transition dynamics of the SDE, which is an idealization of SGD. For the asymptotic regime to directly represent the SGD escape dynamics, one requires the additional uniform-in-time approximation of SGD by SDE, which remains an open question for non-convex loss functions.

Appendix A: Proofs for Section 2

Appendix A.1: Mean and Variance for Random Error Vector

By the mean value theorem with some $\tau(h) \in (0, h)$,

$$\begin{aligned}\nabla L(\mathbf{w}) &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\mathbb{E}[L_n(\mathbf{w})] \\ &= \lim_{h\to 0}\frac{1}{h}\big\{\mathbb{E}[L_n(\mathbf{w}+h)] - \mathbb{E}[L_n(\mathbf{w})]\big\} \\ &= \lim_{h\to 0}\mathbb{E}\left\{\frac{L_n(\mathbf{w}+h)-L_n(\mathbf{w})}{h}\right\} = \lim_{h\to 0}\mathbb{E}\big\{\nabla L_n(\mathbf{w}+\tau(h))\big\}.\end{aligned}$$

By the continuity of ∇L_n and the dominated convergence theorem,

$$\lim_{h\to 0}\mathbb{E}\big\{\nabla L_n(\mathbf{w}+\tau(h))\big\} = \mathbb{E}\left\{\lim_{h\to 0}\nabla L_n(\mathbf{w}+\tau(h))\right\} = \mathbb{E}\big\{\nabla L_n(\mathbf{w})\big\}.$$

Hence, $\boldsymbol{\epsilon}_k$ has mean 0. Since the independent and uniform sampling for the mini-batch B_k, we have $\mathrm{Var}[\boldsymbol{\epsilon}_k] = \boldsymbol{\sigma}^2(\mathbf{w})$ as desired.

We remark that a different view of sampling distribution has been adopted in the literature, for example Jastrzebski et al.[16] and Li et al.[21], where the expectation and variance are taken with respect to the sampling distribution of drawing the mini-batch B_k from $\{1,\ldots,N\}$. On the contrary, we use the sampling distribution with respect to the joint distribution of the underlying population, since our interest is the risk function $L(\cdot)$ instead of the sample average loss

$$\frac{1}{N}[L_1(\cdot)+\cdots+L_N(\cdot)],$$

and we regard the training data only a subset of the underlying population.

Appendix A.2: Proof of Lemma 1

We first consider a special case that $\beta(\mathbf{w}) \equiv \beta$ is a constant and derive the Fokker–Planck equation by following Kolpas et al.[18]. If $\mathbf{W}(t) = W(t) \in \mathbb{R}$, $W(t)$ is a Markov process and the Chapman–Kolmogorov equation gives the conditional probability density function for any $t_1 \le t_2 \le t_3$,

$$p\big(W(t_3)|W(t_1)\big) = \int_{-\infty}^{+\infty} p\big(W(t_3)|W(t_2)=w\big)p\big(W(t_2)=w|W(t_1)\big)\mathrm{d}w.$$

Denote the integral

$$I(h) = \int_{-\infty}^{+\infty} h(w)\partial_t p(w,t|W)\mathrm{d}w, \tag{11}$$

where $h(w)$ is a smooth function with compact support. Observe that

$$\int_{-\infty}^{+\infty} h(w)\partial_t p(w,t|W)\mathrm{d}w = \lim_{\Delta t\to 0}\int_{-\infty}^{+\infty} h(w)\left(\frac{p(w,t+\Delta t|W)-p(w,t|W)}{\Delta t}\right)\mathrm{d}w.$$

Letting Z be an intermediate point between w and W. Applying the Chapman–Kolmogorov identity on the right hand side yields

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\int_{-\infty}^{+\infty}h(w)\int_{-\infty}^{+\infty}p(w,\Delta t|Z)p(Z,t|W)\mathrm{d}Z\mathrm{d}w-\int_{-\infty}^{+\infty}h(w)p(w,t|W)\mathrm{d}w\right).$$

By changing the order of integrations in the first term and letting w approach Z in the second term, we obtain that

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\int_{-\infty}^{+\infty}p(Z,t|W)\int_{-\infty}^{+\infty}p(w,\Delta t|Z)(h(w)-h(Z))\mathrm{d}w\mathrm{d}Z\right).$$

Expand $h(w)$ as a Taylor series about Z, we can write the above integral as

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left(\int_{-\infty}^{+\infty}p(Z,t|W)\int_{-\infty}^{+\infty}p(w,\Delta t|Z)\sum_{n=1}^{\infty}h^{(n)}(Z)\frac{(w-Z)^n}{n!}\right)\mathrm{d}w\mathrm{d}Z.$$

Now we define the function

$$D^{(n)}(Z)=\frac{1}{n!}\frac{1}{\Delta t}\int_{-\infty}^{+\infty}p(w,\Delta t|Z)(w-Z)^n\mathrm{d}w.$$

We can write the integral $I(h)$ defined in (11) as

$$\int_{-\infty}^{+\infty}h(w)\partial_t p(w,t|W)\mathrm{d}w=\int_{-\infty}^{+\infty}p(Z,t|W)\sum_{n=1}^{\infty}D^{(n)}(Z)h^{(n)}(Z)\mathrm{d}Z.$$

Taking the integration by parts n times gives

$$\partial_t p(w,t)=\sum_{n=1}^{\infty}-\frac{\partial^n}{\partial Z^n}\left[D^{(n)}(Z)p(Z,t|W)\right].$$

Let $D^{(1)}(w)=-L(w)$, $D^{(2)}(w)=-\gamma(t)\beta/[2M(t)]$ and $D^{(n)}(w)=0$ for all $n\geq 3$. Then, the above equation yields

$$\partial_t p(w,t)=\frac{\partial}{\partial w}\left[\nabla L(w)p(w,t)\right]+\frac{\partial}{\partial w^2}\left[\frac{\gamma(t)\beta}{2M(t)}p(w,t)\right],$$

which is the Fokker–Planck equation in one variable. For the multidimensional case that $\mathbf{W}=(W_1,W_2,\ldots,W_p)\in\mathbb{R}^p$, we similarly generalize the above procedure to get

$$\begin{aligned}\partial_t p(\mathbf{w},t)&=\sum_{i=1}^{p}\frac{\partial}{\partial w_i}\left[\nabla L(\mathbf{w})p(\mathbf{w},t)\right]+\sum_{i=1}^{p}\frac{\partial^2}{\partial w_i^2}\left[\frac{\gamma(t)\beta}{2M(t)}p(\mathbf{w},t)\right]\\&=\nabla\cdot\left(\nabla L(\mathbf{w})p+\frac{\gamma(t)\beta}{2M(t)}\nabla p\right).\end{aligned}\tag{12}$$

Since $\mathbf{W}(0)=\mathbf{w}_0$, $p(\mathbf{w},0)=\delta(\mathbf{w}_0)$. This completes the derivation of the Fokker–Planck equation for constant $\beta(\mathbf{w})=\beta$.

For deriving (5) with general $\beta(\mathbf{w})$, we can simply apply (12) together with the fact that

$$\nabla\left[\frac{\gamma(t)\beta(\mathbf{w})}{2M(t)}p\right]=\nabla\left[\frac{\gamma(t)\beta(\mathbf{w})}{2M(t)}\right]p+\frac{\gamma(t)\beta(\mathbf{w})}{2M(t)}\nabla p.$$

This completes the proof.

Appendix B: Proofs for Section 3

Appendix B.1: Discussion on Main Assumptions (A.1)–(A.3)

We show that Assumptions (A.1)–(A.3) hold for the squared loss and the regularized mean cross-entropy loss. Denote by $\{(\mathbf{x}_n, y_n), 1 \le n \le N\}$ the set of training data. Without loss of generality, let $\mathrm{Var}[y_n|\mathbf{x}_n] = 1$. First, we consider the squared loss with the corresponding risk function

$$L(\mathbf{w})=\left(\mathbf{w}-\mathbf{w}^0\right)^\top\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^\top\right]\left(\mathbf{w}-\mathbf{w}^0\right)+1,$$

where $\mathbf{w}^0$ is the true parameter vector. Since $\mathrm{Var}[\nabla L_n(\mathbf{w})] \equiv \boldsymbol{\sigma}^2(\mathbf{w})$ is positive definite, we have

$$\begin{aligned}\lim_{\|\mathbf{w}\|\to+\infty}L(\mathbf{w}) &\ge \lim_{\|\mathbf{w}\|\to+\infty}\lambda_{\min}\left\{\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^\top\right]\right\}\|\mathbf{w}-\mathbf{w}^0\|^2+1\\ &\ge \lim_{\|\mathbf{w}\|\to+\infty}\lambda_{\min}\left\{\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^\top\right]\right\}\left[\|\mathbf{w}\|^2/2-\|\mathbf{w}^0\|^2/2\right]+1=+\infty,\end{aligned}\tag{13}$$

where $\lambda_{\min}\{\cdot\}$ denotes the minimal eigenvalue. Note that

$$\begin{aligned}\int \mathrm{e}^{-L(\mathbf{w})}d\mathbf{w} &= \int \exp\left(-(\mathbf{w}-\mathbf{w}^0)^\top\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top](\mathbf{w}-\mathbf{w}^0)-1\right)d\mathbf{w}\\ &\le \int \exp\left(-\lambda_{\min}\{\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top]\}[\|\mathbf{w}\|^2/2-\|\mathbf{w}^0\|^2/2]-1\right)d\mathbf{w}<+\infty.\end{aligned}$$

Hence, Assumption (A.1) holds. To prove (A.2), note that

$$\|\nabla L(\mathbf{w})\|^2/2=2\left(\mathbf{w}-\mathbf{w}^0\right)^\top\left\{\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^\top\right]\right\}^2\left(\mathbf{w}-\mathbf{w}^0\right),\quad \mathrm{Tr}\left(\nabla^2L(\mathbf{w})\right)=\mathrm{Tr}\left\{\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^\top\right]\right\}.$$

Similar to (13), we can prove that

$$\lim_{\|\mathbf{w}\|\to+\infty}\left\{\|\nabla L(\mathbf{w})\|^2/2-\mathrm{Tr}(\nabla^2L(\mathbf{w}))\right\}=+\infty,\quad \lim_{\|\mathbf{w}\|\to+\infty}\left\{\mathrm{Tr}(\nabla^2L(\mathbf{w}))/\|\nabla L(\mathbf{w})\|^2\right\}=0.$$

This finishes the proof for Assumption (A.2). Finally, (A.3) can be shown similarly by following the proof for (A.2) and we omit the details.

Next, we consider the mean cross-entropy loss with the l_2-penalty for the logistic regression. Without loss of generality, we consider the binary classification:

$$L(\mathbf{w})=\mathbb{E}[-y_n\log\widehat{y}_n-(1-y_n)\log(1-\widehat{y}_n)]+\lambda\|\mathbf{w}\|^2$$

with $\widehat{y}_n=(1+\mathrm{e}^{-\mathbf{w}\cdot\mathbf{x}_n})^{-1}$. Note that

$$\lim_{\|\mathbf{w}\|\to+\infty} L(\mathbf{w}) \geq \lambda\|\mathbf{w}\|^2 = +\infty, \quad \int \mathrm{e}^{-L(\mathbf{w})}\mathrm{d}\mathbf{w} \leq \int \mathrm{e}^{-\lambda\|\mathbf{w}\|^2}\mathrm{d}\mathbf{w} < +\infty$$

which proves (A.1). For (A.2), since

$$\nabla L(\mathbf{w}) = \mathbb{E}[-\mathbf{x}_n y_n + \mathbf{x}_n/(1+\mathrm{e}^{-\mathbf{w}\cdot\mathbf{x}_n})] + 2\lambda\mathbf{w},$$

and

$$\mathrm{Tr}(\nabla^2 L(\mathbf{w})) = \mathbb{E}\left[\frac{\mathrm{e}^{-\mathbf{w}\cdot\mathbf{x}_n}}{(1+\mathrm{e}^{-\mathbf{w}\cdot\mathbf{x}_n})^2}\mathrm{Tr}(\mathbf{x}_n\mathbf{x}_n^\top)\right] + 2\lambda d,$$

we have

$$\|\nabla L(\mathbf{w})\|^2/2 - \mathrm{Tr}(\nabla^2 L(\mathbf{w})) \to \infty, \quad \mathrm{Tr}(\nabla^2 L(\mathbf{w}))/\|\nabla L(\mathbf{w})\|^2 \to 0, \text{ as } \|\mathbf{w}\| \to \infty.$$

Similarly, Assumption (A.3) can be verified as by following the proof for (A.2).

Appendix B.2: Proof of Lemma 3

By Assumption (A.1), the density function $p_\infty(\mathbf{w}) \equiv \kappa \mathrm{e}^{-2M(\infty)L(\mathbf{w})/[\gamma(\infty)\beta]}$ is well-defined. Moreover, $p_\infty(\mathbf{w})$ satisfies

$$\nabla\cdot\left[\nabla\left(L(\mathbf{w}) + \frac{\gamma(\infty)\beta}{2M(\infty)}\right)p_\infty(\mathbf{w}) + \frac{\gamma(\infty)\beta}{2M(\infty)}\nabla p_\infty(\mathbf{w})\right] = 0.$$

Hence, $p_\infty(\mathbf{w})$ is a stationary solution to Fokker–Planck equation (5) by letting $\partial_t p(\mathbf{w}, t) = 0$.

Appendix B.3: Proof of Theorem 1

Parallel to the notation $p_\infty(\mathbf{w}) = \kappa\exp(-\frac{2M(\infty)L(\mathbf{w})}{\gamma(\infty)\beta})$ in Lemma 3, we define

$$\hat{p}(\mathbf{w}, t) \equiv \kappa(t)\exp\left(-\eta(t)L(\mathbf{w})\right),$$

where

$$\eta(t) \equiv 2M(t)/[\gamma(t)\beta], \tag{14}$$

and $\kappa(t)$ is a time-dependent normalization factor such that

$$\int \hat{p}(\mathbf{w}, t)\mathrm{d}\mathbf{w} = 1.$$

We can rewrite (5) as

$$\partial_t p = \frac{1}{\eta}\nabla_{\mathbf{w}}\cdot\left(\hat{p}\nabla_{\mathbf{w}}\left(\frac{p}{\hat{p}}\right)\right). \tag{15}$$

Let

$$\delta(t,\mathbf{w}) \equiv \frac{\kappa(t)}{\kappa}\exp\left(L(\mathbf{w})(\eta(\infty)-\eta(t))\right).$$

Then

$$\hat{p}(t,\mathbf{w}) = p_\infty(\mathbf{w})\delta(t,\mathbf{w}).$$

Denote by $h(\mathbf{w},t)$ the scaled distance between $p(\mathbf{w},t)$ and $p_\infty(\mathbf{w})$:

$$h(\mathbf{w},t) \equiv \frac{p(\mathbf{w},t)-p_\infty(\mathbf{w})}{\sqrt{p_\infty(\mathbf{w})}},$$

which satisfies the following equation:

$$\begin{aligned}\partial_t h &= \frac{1}{\eta\sqrt{p_\infty}}\nabla_{\mathbf{w}}\cdot\left[\hat{p}\,\nabla_{\mathbf{w}}\left(\frac{1}{\delta}+\frac{h}{\sqrt{p_\infty}\delta}\right)\right]\\ &= \frac{1}{\eta\sqrt{p_\infty}}\nabla_{\mathbf{w}}\cdot\left[p_\infty\left(\nabla_{\mathbf{w}}L\hat{\delta}+\nabla_{\mathbf{w}}L\hat{\delta}\left(\frac{h}{\sqrt{p_\infty}}\right)+\nabla_{\mathbf{w}}\left(\frac{h}{\sqrt{p_\infty}}\right)\right)\right].\end{aligned} \tag{16}$$

Here, $\hat{\delta}$ is defined as $\hat{\delta}(t)=\eta(t)-\eta(\infty)$, where $\eta(\infty)=\lim_{t\to\infty}\eta(t)$. We multiply h to the both sides of (16) and integrate them over $\mathbf{w}$. Using the integration by parts, we can obtain

$$\begin{aligned}\frac{1}{2}\partial_t\|h\|^2 = &\underbrace{\frac{\hat{\delta}}{\eta}\int\frac{h}{\sqrt{p_\infty}}\nabla_{\mathbf{w}}\cdot\left(p_\infty\nabla_{\mathbf{w}}L\right)\mathrm{d}\mathbf{w}}_{I} + \underbrace{\frac{\hat{\delta}}{\eta}\int\frac{1}{2}\left\|\frac{h}{\sqrt{p_\infty}}\right\|^2\nabla_{\mathbf{w}}\cdot\left(p_\infty\nabla_{\mathbf{w}}L\right)\mathrm{d}\mathbf{w}}_{II}\\ &\underbrace{-\frac{1}{\eta}\int p_\infty\left\|\nabla_{\mathbf{w}}\left(\frac{h}{\sqrt{p_\infty}}\right)\right\|^2\mathrm{d}\mathbf{w}}_{III}.\end{aligned} \tag{17}$$

We study the parts *I, II, III* in the right-hand side of above equation separately.

For the part *I*, note that

$$\nabla_{\mathbf{w}}\cdot\left(p_\infty\nabla_{\mathbf{w}}L\right) = p_\infty\left(\nabla_{\mathbf{w}}\cdot\nabla_{\mathbf{w}}L-\eta(\infty)\|\nabla_{\mathbf{w}}L\|^2\right).$$

Hence, Assumption (A.3) yields that

$$\left|\nabla_{\mathbf{w}}\cdot\left(p_\infty\nabla_{\mathbf{w}}L\right)\right| \le p_\infty^{2/3}\max\{1,\eta(\infty)\}M(\infty),$$

which implies that an upper bound of part *I* in (17):

$$I \leq \frac{\max\{1, \eta(\infty)\}M(\infty)}{2}\left(\|h\|^2 + \int p_\infty^{1/3} \mathrm{d}\mathbf{w}\right).$$

For the part *II*, note that Assumption (A.3) gives

$$\lim_{\|\mathbf{w}\|\to\infty} \frac{\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L}{2\eta(\infty)\left\|\nabla_{\mathbf{w}} L\right\|^2} = 0,$$

which together with Assumption (A.2) implies that

$$\lim_{\|\mathbf{w}\|\to\infty} \left\|\nabla_{\mathbf{w}} L\right\|^2 \to +\infty.$$

Thus, there exists a constant R, such that

$$\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - 2\eta(\infty)\left\|\nabla_{\mathbf{w}} L\right\|^2 \leq \eta(\infty), \quad \eta(\infty)\left\|\nabla_{\mathbf{w}} L\right\|^2 \geq \eta(\infty), \quad \text{for } \forall \|\mathbf{w}\| > R.$$

Hence,

$$\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta(\infty)\left\|\nabla_{\mathbf{w}} L\right\|^2 \leq 0, \quad \text{for } \forall \|\mathbf{w}\| > R.$$

By the continuity of $L(\mathbf{w})$, there exists a constant C_2 such that

$$\left|\nabla_{\mathbf{w}} \cdot \nabla_{\mathbf{w}} L - \eta(\infty)\left\|\nabla_{\mathbf{w}} L\right\|^2\right| \leq C_2, \quad \text{for } \forall \|\mathbf{w}\| < R.$$

Therefore, we have the following upper bound for the part *II* in (17):

$$|II| \leq \frac{C_2}{2}\|h\|^2.$$

By combining the estimates for the parts *I* and *II*, we have

$$I + II \leq C_1\|h\|^2 + C_1,$$

where $C_1 = \frac{1}{2}\max\{1, \eta(\infty)\}\max\left\{\int p_\infty^{1/3}\mathrm{d}\mathbf{w}, 1 + C_2/2\right\}M(\infty)$.

For the part *III*, note that Assumption (A.2) implies the following Poincaré inequality (see, e.g.,[24]),

$$\int \left\|\nabla_{\mathbf{w}}\left(\frac{h}{\sqrt{p_\infty}}\right)\right\|^2 p_\infty \,\mathrm{d}\mathbf{w} \geq C_P \int \left(\frac{h}{\sqrt{p_\infty}} - \int h\sqrt{p_\infty}\mathrm{d}\mathbf{w}\right)^2 p_\infty \,\mathrm{d}\mathbf{w}. \quad (18)$$

We need to show that

$$\int h\sqrt{p_\infty}\,\mathrm{d}\mathbf{w} = 0. \quad (19)$$

The (19) can be proven using the conservation of mass. In particular, if we integrate (15) over $\mathbf{w}$ and use the integration by parts,

$$\partial_t \left(\int p(\mathbf{w}, t)\, \mathrm{d}\mathbf{w} \right) = 0,$$

which implies $\int h\sqrt{p_\infty}\, \mathrm{d}\mathbf{w} = \int p\, \mathrm{d}\mathbf{w} - \int p_\infty\, \mathrm{d}\mathbf{w} = 0$. Combining (18) with (19) gives a lower bound for the part *III*:

$$III \geq C_P \|h\|^2.$$

Combining () and () gives

$$\frac{1}{2}\partial_t \|h\|^2 + \frac{C_P}{\eta}\|h\|^2 \leq \frac{C_1 \hat{\delta}}{\eta}\left(\|h\|^2 + 1\right) \tag{20}$$

Since $\eta(t) \to \eta(\infty) > 0$ as $t \to \infty$, there exists some T large enough and for $\forall t > T$,

$$\hat{\delta} = |\eta(t) - \eta(\infty)| \leq \min\left\{ \frac{\eta(\infty)}{3}, \frac{C_P}{3C_1} \right\}. \tag{21}$$

Plugging $\hat{\delta} \leq C_P/3C_1$ into (20), we have

$$\frac{1}{2}\partial_t \|h\|^2 + \frac{2C_P}{3\eta}\|h\|^2 \leq \frac{C_P}{3\eta}, \quad \text{for } \forall t > T. \tag{22}$$

Note that (21) also implies that $2\eta(\infty)/3 \leq \eta(t) \leq 4\eta(\infty)/3$. Thus,

$$\frac{2C_P}{3\eta} \geq \frac{C_P}{2\eta(\infty)}, \quad \frac{C_P}{3\eta} \leq \frac{C_P}{2\eta(\infty)}.$$

Plugging back to (22), we arrive at

$$\frac{1}{2}\partial_t \|h\|^2 + \frac{C_P}{2\eta(\infty)}\|h\|^2 \leq \frac{C_P}{2\eta(\infty)}, \quad \text{for } \forall t > T.$$

Integrating the above equation from T to $t > T$, we have

$$\|h(t)\|^2 \leq \left(\|h(T)\|^2 + \frac{C_P}{\eta(\infty)}(t - T) \right) - \frac{C_P}{\eta(\infty)} \int_T^t \|h(s)\|^2 \mathrm{d}s.$$

By Gronwall's Inequality, we finally get

$$\|h(t)\|^2 \leq \left(\frac{C_P}{\eta(\infty)}(t - T) + \|h(T)\|^2 \right) \exp\left(-\frac{C_P}{\eta(\infty)}(t - T) \right).$$

This completes the proof.

Appendix B.4: Quantification of *T* in Theorem 1

We quantify T by giving a condition that a minimum T should satisfy. From the proof in Sect. 3, it is clear that T should be large enough such that for all $t > T$,

$$|\eta(t) - \eta(\infty)| \le \min\left\{\frac{\eta(\infty)}{3}, \frac{C_P}{3C_1}\right\},$$

where $\eta(t)$ is defined in (14) and $\eta(\infty) = \lim_{t\to\infty}\eta(t)$, and

$$C_1 = \frac{M}{2}\max\{1, \eta(\infty)\}\max\left\{\int p_\infty^{1/3}\mathrm{d}\mathbf{w}, 1 + \frac{C_2}{2}\right\},$$

and $C_2 > 0$ is an upper bound for $\left|\nabla_{\mathbf{w}}\cdot\nabla_{\mathbf{w}}L - \eta(\infty)\left\|\nabla_{\mathbf{w}}L\right\|^2\right|$ in the bounded domain $\{\|\mathbf{w}\| < R\}$ such that

$$\nabla_{\mathbf{w}}\cdot\nabla_{\mathbf{w}}L - \eta(\infty)\left\|\nabla_{\mathbf{w}}L\right\|^2 \le \begin{cases} 0, & \text{for } \forall\|\mathbf{w}\| > R,\\ C_2, & \text{for } \forall\|\mathbf{w}\| < R.\end{cases}$$

Appendix B.5: Proof of Theorem 2

Denote by $P_\epsilon(\check{\mathbf{w}}) = \mathbb{P}(\|\mathbf{W}(\infty) - \check{\mathbf{w}}\| \le \epsilon)$ the probability that $\mathbf{W}(\infty)$ is trapped in an ϵ-neighborhood of the minimum $\check{\mathbf{w}}$. Recall the probability density function of $\mathbf{W}(\infty)$ is $p_\infty(\mathbf{w})$. Then,

$$\begin{aligned} P_\epsilon(\check{\mathbf{w}}) &= \int_{\|\mathbf{w}-\check{\mathbf{w}}\|^2\le\epsilon^2} \kappa \mathrm{e}^{-\eta(\infty)L(\mathbf{w})}\mathrm{d}\mathbf{w}\\ &= \int_{\|\mathbf{w}-\check{\mathbf{w}}\|^2\le\epsilon^2} \kappa\exp\left(-\eta(\infty)[L(\check{\mathbf{w}}) + (\mathbf{w}-\check{\mathbf{w}})'\nabla^2 L(\check{\mathbf{w}})(\mathbf{w}-\check{\mathbf{w}}) + o\{(\mathbf{w}-\check{\mathbf{w}})^2\}]\right)\mathrm{d}\mathbf{w}, \end{aligned}$$

where $\eta(t)$ is defined in (14) and $\eta(\infty) = \lim_{t\to\infty}\eta(t)$. Since $\check{\mathbf{w}}$ is a local minimum of $L(\mathbf{w})$, $\nabla^2 L(\check{\mathbf{w}})$ is positive definite. There exists an orthogonal matrix O and diagonal matrix F such that $\nabla^2 L = O'FO$. For simplicity, we assume that $\nabla^2 L = F = \mathrm{diag}(\lambda_{\min}, \cdots, \lambda_d)$. Then,

$$\begin{aligned} &\lim_{\epsilon\to 0} P_\epsilon(\check{\mathbf{w}})\\ &= \lim_{\epsilon\to 0}\left[\kappa \mathrm{e}^{-\eta(\infty)L(\check{\mathbf{w}})}\int_{\|\mathbf{w}\|^2\le\epsilon^2}\prod_{j=1}^d \mathrm{e}^{-\eta(\infty)\lambda_j w_j}\mathrm{d}\mathbf{w}\right]\mathrm{e}^{\eta(\infty)\epsilon^2}\\ &= \lim_{\epsilon\to 0}\left[\kappa \mathrm{e}^{-\eta(\infty)L(\check{\mathbf{w}})}\prod_{j=1}^d \frac{1}{\sqrt{\eta(\infty)\lambda_j}}\int_{-\epsilon\sqrt{\eta(\infty)\lambda_j}}^{\epsilon\sqrt{\eta(\infty)\lambda_j}} \mathrm{e}^{-w^2}\mathrm{d}w\right]\mathrm{e}^{\eta(\infty)\epsilon^2}\\ &= \lim_{\epsilon\to 0}\left[\kappa\eta(\infty)^{-d/2}\mathrm{e}^{-\eta(\infty)L(\check{\mathbf{w}})}\prod_{j=1}^d \frac{1}{\sqrt{\lambda_j}}\left(\Phi\left(\epsilon\sqrt{\eta(\infty)\lambda_j}\right) - \Phi\left(-\epsilon\sqrt{\eta(\infty)\lambda_j}\right)\right)\right]\mathrm{e}^{\eta(\infty)\epsilon^2}, \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative density function for standard normal distribution. The first equality is from the change of variable by writing $\mathbf{w} - \check{\mathbf{w}}$ as $\mathbf{w}$. The second equality is from changing $\eta(\infty)\lambda_j\mathbf{w}_j$ to $\mathbf{w}_j$. Using the approximation of the cumulative density function in Pólya[25], we can simplify the above equation as

$$\lim_{\epsilon\to 0} P_\epsilon(\check{\mathbf{w}}) = \lim_{\epsilon\to 0}\left[\frac{\kappa \mathrm{e}^{-2\eta(\infty)L(\check{\mathbf{w}})}}{\eta(\infty)^{d/2}}\prod_{j=1}^{d}\sqrt{\frac{1-\mathrm{e}^{-\epsilon^2\eta(\infty)\lambda_j/\pi}}{\lambda_j}}\right]\mathrm{e}^{\eta(\infty)\epsilon^2}$$
$$= \frac{\kappa \mathrm{e}^{-2\eta(\infty)L(\check{\mathbf{w}})}}{\eta(\infty)^{d/2}|\nabla^2 L(\check{\mathbf{w}})|}\lim_{\epsilon\to 0}\left[\mathrm{e}^{\eta(\infty)\epsilon^2}\prod_{j=1}^{d}\sqrt{1-\mathrm{e}^{-\epsilon^2\eta(\infty)\lambda_j/\pi}}\right].$$

We complete the proof.

Appendix B.6: Proof of Equation 8

Denote by λ_j^k's are eigenvalues of the Hessian $\nabla^2 L(\check{\mathbf{w}}_k)$, $k = 1, 2$ and $j \geq 1$. By Theorem 2 and $L(\check{\mathbf{w}}_1) = L(\check{\mathbf{w}}_2)$, we have that

$$\lim_{\epsilon\to 0}\frac{\mathbb{P}(|\mathbf{W}(\infty)-\check{\mathbf{w}}_1|\leq\epsilon)}{\mathbb{P}(|\mathbf{W}(\infty)-\check{\mathbf{w}}_2|\leq\epsilon)} = \frac{|\nabla^2 L(\check{\mathbf{w}}_2)|}{|\nabla^2 L(\check{\mathbf{w}}_1)|}\sqrt{\lim_{\epsilon\to 0}\prod_{j=1}^{d}\frac{1-\exp\left(-\frac{\epsilon^2\eta(\infty)\lambda_j^1}{\pi}\right)}{1-\exp\left(-\frac{\epsilon^2\eta(\infty)\lambda_j^2}{\pi}\right)}}$$
$$= \frac{|\nabla^2 L(\check{\mathbf{w}}_2)|}{|\nabla^2 L(\check{\mathbf{w}}_1)|}\sqrt{\lim_{\epsilon\to 0}\prod_{j=1}^{d}\frac{\lambda_j^1\exp\left(-\frac{\epsilon^2\eta(\infty)\lambda_j^2}{\pi}\right)}{\lambda_j^2\exp\left(-\frac{\epsilon^2\eta(\infty)\lambda_j^2}{\pi}\right)}} = \frac{|\nabla^2 L(\check{\mathbf{w}}_2)|}{|\nabla^2 L(\check{\mathbf{w}}_1)|}\sqrt{\prod_{j=1}^{d}\frac{\lambda_j^1}{\lambda_j^2}} = \sqrt{\frac{|\nabla^2 L(\check{\mathbf{w}}_2)|}{|\nabla^2 L(\check{\mathbf{w}}_1)|}},$$

where $\eta(t)$ is defined in (14) and $\eta(\infty) = \lim_{t\to\infty}\eta(t)$.

Appendix C: Proofs for Section 4

Appendix C.1: Derivation of SDE for MSGD

For constant learning rate and batch size: $\gamma_k \equiv \gamma, M_k \equiv M$, we rewrite the MSGD as

$$\frac{\mathbf{z}_{k+1}}{\sqrt{\gamma}} = \frac{\mathbf{z}_k}{\sqrt{\gamma}} + \sqrt{\gamma}\left(-\frac{1-\xi}{\gamma}\mathbf{z}_k - \nabla L(\mathbf{w}_k)\right) + \sqrt{\gamma}\left(\nabla L(\mathbf{w}_k) - \left(\frac{1}{M}\sum_{n\in B_k}\nabla L_n(\mathbf{w}_k)\right)\right)$$
$$\mathbf{w}_{k+1} = \mathbf{w}_k + \frac{\mathbf{z}_{k+1}}{\sqrt{\gamma}}\sqrt{\gamma}.$$

Let $\mathbf{v}_k = \mathbf{z}_k/\sqrt{\gamma}$. We have the approximation for MSGD

$$\mathbf{v}_{k+1} - \mathbf{v}_k = -\frac{1-\xi}{\sqrt{\gamma}}\mathbf{v}_k\sqrt{\gamma} - \nabla L(\mathbf{w}_k)\sqrt{\gamma} + \frac{\gamma^{1/4}}{\sqrt{M}}\sqrt{\beta}\nabla^2 B_t,$$
$$\mathbf{w}_{k+1} - \mathbf{w}_k = \mathbf{v}_{k+1}\sqrt{\gamma},$$

where $\beta(\mathbf{w})$ is the covariance function defined in (4). Hence, MSGD is approximated as the Euler–Maruyama discretization for the following SDE,

$$\begin{cases} \mathrm{d}\mathbf{V}(t) = -\nabla L(\mathbf{W}(t))\mathrm{d}t - \dfrac{1-\xi}{\sqrt{\gamma}}\mathbf{V}(t)\mathrm{d}t + \dfrac{\gamma^{1/4}}{\sqrt{M}}\sqrt{\beta(\mathbf{W}(t))}\mathrm{d}\mathbf{B}(t), \\ \mathrm{d}\mathbf{W}(t) = \mathbf{V}(t)\mathrm{d}t, \end{cases}$$

where $\mathbf{v}_k \approx \mathbf{V}(k\sqrt{\gamma})$, $\mathbf{w}_k \approx \mathbf{W}(k\sqrt{\gamma})$.

Appendix C.2: Proof of Lemma 4

We give a formal derivation, which is similar to the procedure in Pavliotis[24]. Let $\phi(\cdot,\cdot)$ be any bivariate function in C^∞ with a compact support. Using the Itô's formula,

$$\begin{aligned} \mathrm{d}\phi(\mathbf{W}(t),\mathbf{V}(t)) &= \frac{\gamma^{1/4}}{\sqrt{M}}\sqrt{\beta}\nabla_{\mathbf{w}}\cdot\nabla_{\mathbf{w}}\phi\mathrm{d}\mathbf{B}(t) \\ &\quad + \left(\mathbf{V}(t)\cdot\nabla_{\mathbf{w}}\phi + \left(-\nabla L(\mathbf{W}(t)) - \frac{1-\xi}{\sqrt{\gamma}}\mathbf{V}(t)\right)\cdot\nabla_{\mathbf{v}}\phi + \frac{\gamma^{1/2}}{2M}\beta(\mathbf{W}(t))\nabla_{\mathbf{v}}\cdot\nabla_{\mathbf{v}}\phi\right)\mathrm{d}t. \end{aligned}$$

By taking the expectation of the above equation and integrating it over the range $[t, t+h]$, we obtain that

$$\begin{aligned} &\frac{1}{h}\mathbb{E}(\phi(\mathbf{W}(t+h),\mathbf{V}(t+h)) - \phi(\mathbf{W}(t),\mathbf{V}(t))) \\ &= \frac{1}{h}\int_t^{t+h}\mathbb{E}\left(\mathbf{V}(s)\cdot\nabla_{\mathbf{w}}\phi + \left(-\nabla L(\mathbf{W}(s)) - \frac{1-\xi}{\sqrt{\gamma}}\mathbf{V}(s)\right)\cdot\nabla_{\mathbf{v}}\phi + \frac{\gamma^{1/2}\beta(\mathbf{W}(s))}{2M}\nabla_{\mathbf{v}}\cdot\nabla_{\mathbf{v}}\phi\right)\mathrm{d}s. \end{aligned}$$

Let $\psi(\mathbf{w},\mathbf{v},t)$ be the joint probability density function of $(\mathbf{W}(t),\mathbf{V}(t))$. The above equation can also be written as

$$\begin{aligned} &\frac{1}{h}\int \phi(\mathbf{w},\mathbf{v})(\psi(\mathbf{w},\mathbf{v},t+h) - \psi(\mathbf{w},\mathbf{v},t))\,\mathrm{d}\mathbf{w}\,\mathrm{d}\mathbf{v} \\ &= \frac{1}{h}\int_t^{t+h}\int\left(\mathbf{v}\cdot\nabla_{\mathbf{w}}\phi + \left(-\nabla L(\mathbf{w}) - \frac{1-\xi}{\sqrt{\gamma}}\mathbf{v}\right)\cdot\nabla_{\mathbf{v}}\phi + \frac{\gamma^{1/2}\beta(\mathbf{w})}{2M}\nabla_{\mathbf{v}}\cdot\nabla_{\mathbf{v}}\phi\right)\psi(\mathbf{w},\mathbf{v},s)\,\mathrm{d}\mathbf{w}\,\mathrm{d}\mathbf{v}\,\mathrm{d}s. \end{aligned}$$

Then, using the integration by parts and letting $h \to 0$ gives

$$\begin{aligned} &\int \phi(\mathbf{w},\mathbf{v})\partial_t\psi\,\mathrm{d}\mathbf{w}\,\mathrm{d}\mathbf{v} \\ &= \int \phi\left(-\mathbf{v}\cdot\nabla_{\mathbf{w}}\psi + \nabla L(\mathbf{w})\cdot\nabla_{\mathbf{v}}\psi + \nabla_{\mathbf{v}}\cdot\left(\frac{1-\xi}{\sqrt{\gamma}}\mathbf{v}\psi\right) + \frac{\gamma^{1/2}\beta(\mathbf{w})}{2M}\nabla_{\mathbf{v}}\cdot\nabla_{\mathbf{v}}\psi\right)\mathrm{d}\mathbf{w}\,\mathrm{d}\mathbf{v}, \end{aligned}$$

which is satisfied for any test functions. Therefore, the density function $\psi(\mathbf{w},\mathbf{v},t)$ satisfies

$$\partial_t \psi + \mathbf{v} \cdot \nabla_{\mathbf{w}} \psi - \nabla L(\mathbf{w}) \cdot \nabla_{\mathbf{v}} \psi = \nabla_{\mathbf{v}} \cdot \left(\frac{1-\xi}{\sqrt{\gamma}} \mathbf{v} \psi + \frac{\gamma^{1/2} \beta(\mathbf{w})}{2M} \nabla_{\mathbf{v}} \psi \right),$$

which agrees with (9).

Next, we can verify that $\psi_\infty(\mathbf{w}, \mathbf{v})$ is a stationary solution of the Vlasov-Fokker–Planck equation (9) by a direct calculation as "Appendix B.2".

Appendix C.3: Discussion on Assumption (A.4)

We show that Assumption (A.4) holds for the squared loss and the regularized mean cross-entropy loss. Denote by $\{(\mathbf{x}_n, y_n), 1 \le n \le N\}$ the set of training data. Without loss of generality, let $\mathrm{Var}[y_n|\mathbf{x}_n] = 1$. For the squared loss,

$$\tilde{L}(\mathbf{w}) = \left(\mathbf{w} - \mathbf{w}^0\right)^\top \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top] \left(\mathbf{w} - \mathbf{w}^0\right) + 1 - \frac{1}{2} C_L^2 \|\mathbf{w}\|^2,$$

where $\mathbf{w}^0$ is the true parameter vector. By a direct calculation,

$$\nabla^2 \tilde{L}(\mathbf{w}) = 2\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top] - C_L^2.$$

Since the eigenvalues of the design matrix $\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]$ are bounded, the eigenvalues of $\nabla^2 \tilde{L}(\mathbf{w})$ are bounded for any C_L. Hence, Assumption (A.4) holds for the squared loss.

Nest, we consider the regularized mean cross-entropy loss for the logistic regression. Similar to "Appendix B.1," letting $C_L = \sqrt{2\lambda}$ yields that

$$\tilde{L}(\mathbf{w}) = \mathbb{E}[-y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)].$$

The (i, j)th entry of the Hessian $\nabla^2 L(\mathbf{w})$ is

$$(\nabla^2 L(\mathbf{w}))_{ij} = \mathbb{E}\left[x_{ni} x_{nj} \frac{\mathrm{e}^{-\mathbf{w} \cdot \mathbf{x}_n}}{(1 + \mathrm{e}^{-\mathbf{w} \cdot \mathbf{x}_n})^2} \right],$$

where x_{ni} is the ith element of $\mathbf{x}_n$. Then,

$$(\nabla^2 L(\mathbf{w}))_{ij} \to 0 \quad \text{as } \|\mathbf{w}\| \to \infty,$$

which implies that there exists finite constant $b_{ij} > 0$ such that $\|(\nabla^2 L(\mathbf{w}))_{ij}\|_\infty \le b_{ij}$ and the largest row sum of the matrix $\{\|(\nabla^2 L)_{ij}\|_\infty\}_{1 \le i,j \le d}$ is upper bounded by $b \equiv \max_i(\sum_j b_{ij})$. Since the largest eigenvalue of a non-negative matrix is upper bounded by its largest row sum, the eigenvalues of $\{\|(\nabla^2 L)_{ij}\|_\infty\}_{1 \le i,j \le d}$ are bounded by b. Hence, Assumption (A.4) also holds for the regularized mean cross-entropy loss.

Appendix C.4: Proof of Theorem 3

Recall the function defined in Theorem 3:

$$h(\mathbf{w},\mathbf{v},t) \equiv \frac{\psi(t,\mathbf{w},\mathbf{v}) - \psi_\infty(\mathbf{w},\mathbf{v})}{\psi_\infty(\mathbf{w},\mathbf{v})},$$

which is the weighted fluctuation function around the stationary solution $\psi_\infty(\mathbf{w},\mathbf{v})$. Then, $h(\mathbf{w},\mathbf{v},t)$ satisfies the following partial differential equation,

$$\partial_t h + Th = Fh, \tag{23}$$

where

$$T = \mathbf{v}\cdot\nabla_{\mathbf{w}} - \nabla L(\mathbf{w})\cdot\nabla_{\mathbf{v}} \quad \text{is the transport operator;}$$
$$F = \frac{\gamma^{1/2}\beta}{2M}\frac{1}{\psi_\infty}\nabla_{\mathbf{v}}\cdot\left(\psi_\infty\nabla_{\mathbf{v}}\right) \quad \text{is the Fokker Planck operator.}$$

Also recall the norm $\|\cdot\|_*$ defined in Theorem 3:

$$\text{For any } h(\mathbf{w},\mathbf{v},t), g(\mathbf{w},\mathbf{v},t): \quad \langle h,g\rangle_* = \int hg\psi_\infty \,\mathrm{d}\mathbf{w}d\mathbf{v}, \quad \|h\|_*^2 = \int |h|^2\psi_\infty \,\mathrm{d}\mathbf{w}d\mathbf{v},$$

Lemma 5 *One have the following properties for the operator T, F:*

(1) $\langle Tf, g\rangle_* = -\langle f, Tg\rangle_*$,
(2) $\langle Tf, f\rangle_* = 0$,
(3) $\langle Ff, g\rangle_* = -\frac{\gamma^{1/2}\beta}{2M}\langle\nabla_{\mathbf{v}}f, \nabla_{\mathbf{v}}g\rangle_*$.

This lemma can be verified by direct calculations, and we omit the details. These properties of operators F, T will be frequently used later.

Lemma 6 *For the positive definite matrix P defined in (10), the function $h(t,\mathbf{w},\mathbf{v})$ satisfies*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}H(t) + \frac{1}{2}\int [\nabla_{\mathbf{w}}h, \nabla_{\mathbf{v}}h]K[\nabla_{\mathbf{w}}h, \nabla_{\mathbf{v}}h]^\top\psi_\infty \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{v}$$
$$\le \left\langle\nabla^2\tilde{L}\nabla_{\mathbf{v}}h, \nabla_{\mathbf{w}}h\right\rangle_* + \left\langle\nabla^2\tilde{L}\nabla_{\mathbf{v}}h, \nabla_{\mathbf{v}}h\right\rangle_*$$

where the modified risk function $\tilde{L}$ is defined in Assumption (A.4), and

$$K \equiv \begin{bmatrix} 2\hat{C}I_d & (C - C_L^2 + \gamma\hat{C})I_d \\ (C - C_L^2 + \gamma\hat{C})I_d & (2\gamma C - 2C_L^2\hat{C})I_d \end{bmatrix}. \tag{24}$$

Proof Taking the gradient $\nabla_{\mathbf{w}}$ to (23) and multiplying it by $\nabla_{\mathbf{w}}h\psi_\infty$ gives

$$\frac{1}{2}\partial_t\left\|\nabla_{\mathbf{w}}h\right\|_*^2 - \langle T\nabla_{\mathbf{w}}h, \nabla_{\mathbf{w}}h\rangle_* - \left\langle\nabla^2 L\nabla_{\mathbf{v}}h, \nabla_{\mathbf{w}}h\right\rangle_* = \langle F\nabla_{\mathbf{w}}h, \nabla_{\mathbf{w}}h\rangle_*$$

Them, applying Lemma 5 yields,

$$\frac{1}{2}\partial_t\|\nabla_{\mathbf{w}}h\|_*^2-\left\langle\nabla^2L\nabla_{\mathbf{v}}h,\nabla_{\mathbf{w}}h\right\rangle_*=-\frac{\gamma^{1/2}\beta}{2M}\sum_{i=1}^{d}\left\|\partial_{v_i}\nabla_{\mathbf{w}}h\right\|_*^2.$$

By Assumption (A.4), we have

$$\frac{1}{2}\partial_t\|\nabla_{\mathbf{w}}h\|_*^2-C_L^2\langle\nabla_{\mathbf{v}}h,\nabla_{\mathbf{w}}h\rangle_*=-\frac{\gamma^{1/2}\beta}{2M}\sum_{i=1}^{d}\left\|\partial_{v_i}\nabla_{\mathbf{w}}h\right\|_*^2+\left\langle\nabla^2\tilde{L}\nabla_{\mathbf{v}}h,\nabla_{\mathbf{w}}h\right\rangle_*. \tag{25}$$

Similarly, taking the gradient $\nabla_{\mathbf{v}}$ to (23), multiplying it by $\nabla_{\mathbf{v}}h\psi_\infty$ and applying Lemma 5 gives,

$$\frac{1}{2}\partial_t\|\nabla_{\mathbf{v}}h\|_*^2+\langle\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h\rangle_*=-\frac{\gamma^{1/2}\beta}{2M}\sum_{i=1}^{d}\left\|\partial_{v_i}\nabla_{\mathbf{v}}h\right\|_*^2-\frac{1-\xi}{\sqrt{\gamma}}\|\nabla_{\mathbf{v}}h\|_*^2. \tag{26}$$

Taking the gradient $\nabla_{\mathbf{v}}$ to (23) and multiply it by $\nabla_{\mathbf{w}}h\psi_\infty$, then taking the gradient $\nabla_{\mathbf{w}}$ to (23) and multiply it by $\nabla_{\mathbf{v}}h\psi_\infty$, and combine the results gives,

$$\begin{aligned}&\partial_t\langle\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h\rangle_*-C_L^2\langle\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h\rangle_*+\|\nabla_{\mathbf{w}}h\|_*^2\\&\quad=-\frac{\gamma^{1/4}\sqrt{\beta}}{\sqrt{M}}\sum_{i=1}^{d}\left\langle\partial_{v_i}\nabla_{\mathbf{v}}h,\partial_{v_i}\nabla_{\mathbf{w}}h\right\rangle_*-\frac{1-\xi}{\sqrt{\gamma}}\langle\nabla_{\mathbf{v}}h,\nabla_{\mathbf{w}}h\rangle_*+\left\langle\nabla^2\tilde{L}\nabla_{\mathbf{v}}h,\nabla_{\mathbf{v}}h\right\rangle_*.\end{aligned} \tag{27}$$

Finally, (25) + C·(26) + $2\hat{C}$· (25) yields

$$\begin{aligned}&\frac{1}{2}\partial_tH(t)+\frac{\gamma^{1/2}\beta}{2M}\sum_{i=1}^{d}\int[\partial_{v_i}\nabla_{\mathbf{w}}h,\partial_{v_i}\nabla_{\mathbf{v}}h]^\top P[\partial_{v_i}\nabla_{\mathbf{w}}h,\partial_{v_i}\nabla_{\mathbf{v}}h]\mathrm{d}\mathbf{w}d\mathbf{v}\\&+\frac{1}{2}\int[\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h]^\top K[\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h]\mathrm{d}\mathbf{w}d\mathbf{v}=\left\langle\nabla^2\tilde{L}\nabla_{\mathbf{w}}h,\nabla_{\mathbf{v}}h\right\rangle_*+\left\langle\nabla^2\tilde{L}\nabla_{\mathbf{v}}h,\nabla_{\mathbf{v}}h\right\rangle_*,\end{aligned} \tag{28}$$

where function $H(t)$ and the positive definite matrix P are defined in (10). The positive definite property of P implies that

$$\frac{\gamma^{1/2}\beta}{2M}\sum_{i=1}^{d}\int[\partial_{v_i}\nabla_{\mathbf{w}}h,\partial_{v_i}\nabla_{\mathbf{v}}h]^\top P[\partial_{v_i}\nabla_{\mathbf{w}}h,\partial_{v_i}\nabla_{\mathbf{v}}h]\mathrm{d}\mathbf{w}d\mathbf{v}\geq 0,$$

which together with (28) complete the proof. □

Lemma 7 *For P, K defined in (10) and (24), respectively, there exists μ, C, and $\hat{C}$ such that*

$$K\geq 2\mu P\geq 0,$$

where value of μ, C, $\hat{C}$ can be quantifies as follows:

$$\begin{cases} \text{when } \dfrac{1-\xi}{\sqrt{\gamma}} < 2C_L : \mu \equiv \dfrac{1-\xi}{\sqrt{\gamma}},\ C \equiv C_L^2,\ \hat{C} \equiv \dfrac{1-\xi}{2\sqrt{\gamma}}; \\ \text{when } \dfrac{1-\xi}{\sqrt{\gamma}} \geq 2C_L : \mu \equiv \dfrac{1-\xi}{\sqrt{\gamma}} - \sqrt{\dfrac{(1-\xi)^2}{\gamma} - 4C_L^2},\ C \equiv \dfrac{(1-\xi)^2}{2\gamma} - C_L^2,\ \hat{C} \equiv \dfrac{1-\xi}{2\sqrt{\gamma}}. \end{cases}$$

This lemma can be verified by direct calculations and we omit the details. We now go back to the proof of Theorem 3.

Proof of Theorem 3 By Lemmas 6, 7, and Assumption (A.4), we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}H(t) + \mu H(t) \leq \frac{1+\sqrt{2}}{2} b(\|\nabla_{\mathbf{w}} h\|_*^2 + \|\nabla_{\mathbf{v}} h\|_*^2)$$

Let $\lambda_{\min}$ be the smallest eigenvalue of the positive definite matrix P, we have

$$\lambda_{\min}(\|\nabla_{\mathbf{w}} h\|_*^2 + \|\nabla_{\mathbf{v}} h\|_*^2) \leq H(t), \tag{29}$$

which implies

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}H(t) + (\mu - \hat{\mu})H(t) \leq 0,$$

where $\hat{\mu} = \dfrac{1+\sqrt{2}}{2}\dfrac{b}{\lambda_{\min}}$. Solving the above inequality yields,

$$H(t) \leq \mathrm{e}^{-2(\mu-\hat{\mu})t} H(0).$$

Inserting this inequality to (29) gives

$$\|\nabla_{\mathbf{w}} h\|_*^2 + \|\nabla_{\mathbf{v}} h\|_*^2 \leq \frac{1}{\lambda_{\min}} \mathrm{e}^{-2(\mu-\hat{\mu})t} H(0). \tag{30}$$

Besides, the Poincaré inequality w.r.t. the measure $\psi_\infty(\mathbf{w}, \mathbf{v})$ is

$$\|\nabla_{\mathbf{w}} h\|_*^2 + \|\nabla_{\mathbf{v}} h\|_*^2 \geq \frac{2M(1-\xi)}{\gamma\beta} \min\{C_P, d\} \|h\|_*^2.$$

Inserting it back to (30) leads to,

$$\|h\|_*^2 \leq \frac{\gamma\beta}{2M(1-\xi)\min\{C_P, d\}} \frac{1}{\lambda_{\min}} \mathrm{e}^{-2(\mu-\hat{\mu})t} H(0)$$

□

Appendix D: Networks and Dataset Used in Sect. 5.1

The N1 network is a *shallow convolutional* network, which is a modified AlexNet configuration (Krizhevsky et al.[19]). Let $n \times [a, b, c, d]$ denote a stack of n convolution layers of a filters and a Kernel size of $b \times c$ with stride length of d. Then, N1 network uses two sets of [65, 5, 5, 2]–MaxPool(3) and two dense layers of sizes (384, 192), and finally an output layer of size 10. We use ReLU activations.

The N2 network is a *deep convolutional* network, which is a modified VGG configuration (Simonyan and Zisserman[27]). The N2 network uses the configuration: $2 \times [64, 3, 3, 1]$, $2 \times [128, 3, 3, 1]$, $3 \times [256, 3, 3, 1]$, $3 \times [512, 3, 3, 1]$, $3 \times [512, 3, 3, 1]$ and a MaxPool(2) after each stack. This stack is followed by a 512-dimensional dense layer and finally, a ten-dimensional output layer. We use ReLU activations.

The MNIST dataset (LeCun et al.[20]) contains 60,000 training images and 10,000 testing images, where each image is black and white and normalized to fit into a 28×28 pixel bounding box and it belongs to one of total ten classes of handwritten digits (i.e., $0, 1, 2, \ldots, 10$).

The CIFAR-10 dataset consists of 50,000 training data and 10,000 testing data, where each data is a color image with 32×32 features and it belongs to one of total ten classes representing airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

References

1. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, Chen J (2016) Deep speech 2: end-to-end speech recognition in English and Mandarin. In: International conference on machine learning (ICML), pp 173–182
2. An J, Lu J, Ying L (2019) Stochastic modified equations for the asynchronous stochastic gradient descent. Inf Inference. https://doi.org/10.1093/imaiai/iaz030
3. Berglund N (2013) Kramers' law: validity, derivations and generalisations. Markov Process Relat Fields 19(3):459–490
4. Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60(2):223–311
5. Bovier A, Eckhoff M, Gayrard V, Klein M (2004) Metastability in reversible diffusion processes I: sharp asymptotics for capacities and exit times. J Eur Math Soc 6(4):399–424
6. Bovier A, Gayrard V, Klein M (2004) Metastability in reversible diffusion processes II: precise asymptotics for small eigenvalues. J Eur Math Soc 7(1):69–99
7. Chaudhari P, Oberman A, Osher S, Soatto S, Carlier G (2017) Deep relaxation: partial differential equations for optimizing deep neural networks. In: International conference on learning representations (ICLR)
8. Chaudhari P, Soatto S (2018) Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In: International conference on learning representations (ICLR)
9. Dinh L, Pascanu R, Bengio S, Bengio Y (2017) Sharp minima can generalize for deep nets. In: International conference on machine learning (ICML)
10. Evans LC (2010) Partial differential equations, vol 19. American Mathematical Society, Providence
11. Goyal P, Dollar P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y, He K (2017) Accurate, large minibatch SGD: training ImageNet in 1 hour. arXiv:1706.02677
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778

13. Heskes TM, Kappen B (1993) On-line learning processes in artificial neural networks. Mathematical foundations of neural networks. Elsevier, Amsterdam, pp 199–233
14. Hoffer E, Hubara I, Soudry D (2017) Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In: Advances in neural information processing systems (NIPS), pp 1729–1739
15. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning (ICML)
16. Jastrzebski S, Kenton Z, Arpit D, Ballas N, Fischer A, Bengio Y, Storkey A (2017) Three factors influencing minima in SGD. arXiv:1711.04623
17. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP (2017) On large-batch training for deep learning: Generalization gap and sharp minima. In: International conference on learning representations (ICLR)
18. Kolpas A, Moehlis J, Kevrekidis IG (2007) Coarse-grained analysis of stochasticity-induced switching between collective motion states. Proc Natl Acad Sci 104(14):5931–5935
19. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NIPS), pp 1091–1105
20. LeCun Y, Cortes C, Christopher JC (1998) The MNIST dataset of handwritten digit. http://yann.lecun.com/exdb/mnist
21. Li Q, Tai C, Weinan E (2017) Stochastic modified equations and adaptive stochastic gradient algorithms. In: International conference on machine learning (ICML)
22. Mandt S, Hoffman MD, Blei DM (2017) Stochastic gradient descent as approximate bayesian inference. J Mach Learn Res 18:1–35
23. Nesterov Y (2013) Introductory lectures on convex optimization: a basic course, vol 87. Springer, Berlin
24. Pavliotis GA (2014) Stochastic processes and applications: diffusion processes, the Fokker–Planck and Langevin equations. Springer, Berlin
25. Pólya G (1945) Remarks on computing the probability integral in one and two dimensions. In: Proceedings of the 1st Berkeley symposium on mathematical statistics and probability
26. Qian N (1999) On the momentum term in gradient descent learning algorithm. Neural Netw 12(1):145–151
27. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR)
28. Smith SL, Le QV (2018) A Bayesian perspective on generalization and stochastic gradient descent. In: International conference on learning representations (ICLR)
29. Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. In: International conference on machine learning (ICML), pp 1139–1147
30. Villani C (2009) Hypocoercivity. Memoirs of the American Mathematical Society 202 (950)
31. Wu L, Zhu Z (2017) Towards understanding generalization of deep learning: perspective of loss landscapes. In: International conference on machine learning (ICML) workshop on principled approaches to deep learning