Why is parameter averaging beneficial in SGD?
An objective smoothing perspective

Atsushi Nitanda[†‡]   **Ryuhei Kikuchi**[*]   **Shugo Maeda**[*]   **Denny Wu**[⋆∘]

atsushi_nitanda@cfar.a-star.edu.sg, kikuchi.ryuhei189@mail.kyutech.jp,
maeda.shugo401@mail.kyutech.jp, dennywu@nyu.edu

† IHPC, Agency for Science, Technology and Research, Singapore
‡ CFAR, Agency for Science, Technology and Research, Singapore
∗ Kyushu Institute of Technology, Japan
⋆ New York University, United States ∘ Flatiron Institute, United States

Abstract

It is often observed that stochastic gradient descent (SGD) and its variants implicitly select a solution with good generalization performance; such implicit bias is often characterized in terms of the *sharpness* of the minima. Kleinberg et al. (2018) connected this bias with the smoothing effect of SGD which eliminates sharp local minima by the convolution using the stochastic gradient noise. We follow this line of research and study the commonly-used averaged SGD algorithm, which has been empirically observed in Izmailov et al. (2018) to prefer a flat minimum and therefore achieves better generalization. We prove that in certain problem settings, averaged SGD can efficiently optimize the smoothed objective which avoids sharp local minima. In experiments, we verify our theory and show that parameter averaging with an appropriate step size indeed leads to significant improvement in the performance of SGD.

1 INTRODUCTION

Stochastic gradient descent (SGD) (Robbins and Monro, 1951) is a powerful learning method for training deep neural networks. SGD often exhibits higher generalization performance than many of its variants, even when they achieve faster convergence with respect to the training loss (Keskar and Socher, 2017; Wilson

et al., 2017; Luo et al., 2019). Therefore, the study of *implicit bias* to characterize the parameters obtained by SGD has become an active research topic.

Among such studies, *flat minima* (Hinton and Van Camp, 1993; Hochreiter and Schmidhuber, 1997) has been recognized as an important notion relevant to the generalization performance of deep neural networks. Hochreiter and Schmidhuber (1997); Keskar et al. (2017) suggested that the flat minima generalize well compared to sharp minima, and Neyshabur et al. (2017) rigorously supported this correlation under ℓ_2-regularization by using the PAC-Bayesian framework (McAllester, 1998, 1999). Jiang et al. (2020) verified that the flatness measures reliably capture the generalization performance compared to many other complexities. Furthermore, Keskar et al. (2017) empirically demonstrated that SGD prefers a flat minimum due to its own stochastic gradient noise and Kleinberg et al. (2018) proved this implicit bias via the smoothing effect brought by the noise.

Along this line of research, there have been several attempts to develop optimization methods with stronger bias than SGD. Especially, stochastic weight averaging (SWA) (Izmailov et al., 2018) and sharpness aware minimization (SAM) (Foret et al., 2020) achieved significant improvement in generalization performance over SGD. SWA is a cyclic averaging scheme for SGD, which includes the averaged SGD (Ruppert, 1988; Polyak and Juditsky, 1992) as a special case. It is well known that averaged SGD with an appropriately small step size or diminishing step size is a statistically optimal method for the convex optimization problems (Bach and Moulines, 2011; Lacoste-Julien et al., 2012; Rakhlin et al., 2012). On the other hand, Izmailov et al. (2018) found that an appropriately large step size is rather preferable to a small step size when training deep neural

networks with the averaged SGD.

The success of using a large step size can be attributed to the strong bias towards a flat minimum as discussed in Izmailov et al. (2018). SGD with a large step size cannot stay in sharp regions because of the amplified stochastic gradient noise, and thus it moves to a flatter region. After a long run, SGD will finally oscillate according to an invariant distribution. Then, by taking the average, we can get the mean of this distribution, which is located inside a flat region. Although this provides a good insight into how the averaged SGD with a large step size behaves, the theoretical understanding remains elusive. Hence, the research problem we aim to address is

When and why does the averaged SGD converge to a flat region more stably than SGD?

In our work, we address this question by analyzing the averaged SGD based on the smoothing effect of stochastic gradient noise.

1.1 Contributions

We employ the *alternative view* of SGD (Kleinberg et al., 2018) that connects SGD for the objective $f(w)$ with the optimization of the smoothed objective $F(v) = \mathbb{E}[f(v - \eta\epsilon')]$ through the change of variable $w \mapsto v = w - \eta\nabla f(w)$, where ϵ' is a stochastic gradient noise. In fact, F is a smoothed function because it is basically the convolution using the stochastic gradient noise, and the strength of the smoothness is controlled by the step size η as seen in the left figures of Figure 1.

Therefore, an optimization method that can minimize F with a certain accuracy is expected to converge to a flat region, and our aim is to clarify *when the averaged SGD can get closer to the minimizer of F than SGD under the same step size.* Figures 1 and 2 illustrate toy examples, respectively, where the averaged SGD works better in the above sense especially when using a large step size and where the averaged SGD can approximately optimize F with η varying.

Our contributions are summarized below:

- We derive the upper and lower bound conditions in Theorem 3.2 and Eq. (5) for the SGD error: $D_\infty = \lim_{T\to\infty} \sqrt{\frac{1}{T+1}\sum_{t=0}^{T}\|v_t - v_*\|^2}$ such that averaged SGD gets closer to the minimizer $v_* = \arg\min F(v)$ than SGD.

- We estimate the SGD error D_∞ in Proposition 3.4 and 3.5 under additional lower bound on the stochastic gradient noise and sort of one-point strong convexity for F, with a specific Example 3.6 satisfying the required condition.

- We empirically observe that averaged SGD achieves high test accuracy on image classification tasks when using a relatively large step size that makes SGD itself unstable.

2 PRELIMINARY

In this section, we introduce the stochastic gradient descent (SGD) for general problems including the risk minimization problems appearing in machine learning, and introduce the alternative view of SGD developed by Kleinberg et al. (2018).

2.1 Stochastic gradient descent

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth nonconvex objective function to be minimized. For simplicity, we assume f is nonnegative. A stochastic gradient descent, randomly initialized at w_0, for optimizing f is described as follows: for $t = 0, 1, 2, \ldots$

$$w_{t+1} = w_t - \eta\left(\nabla f(w_t) + \epsilon_{t+1}(w_t)\right), \qquad (1)$$

where $\eta > 0$ is the step size and $\epsilon_{t+1} : \mathbb{R}^d \to \mathbb{R}^d$ is a random field corresponding to the stochastic gradient noise i.e., for any $w \in \mathbb{R}^d$, $\{\epsilon_{t+1}(w)\}_{t=0}^{\infty}$ is a sequence of zero-mean random variables taking values in $\mathbb{R}^d$. A typical setup of the above is an empirical risk minimization in machine learning.

Example 2.1. Let $\ell(w, z)$ be a loss function where $w \in \mathbb{R}^d$ and $z \in \mathbb{R}^p$ represent the parameter and data example, respectively, and let $r(w)$ be a regularization function. Let $\{z_i\}_{i=1}^{n}$ be training examples. Then, the regularized empirical risk is defined as follows:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}\ell(w, z_i) + \lambda r(w),$$

where $\lambda > 0$ is a regularization coefficient. A standard stochastic gradient at t-th iterate w_t is defined as $\nabla_w\ell(w_t, z_{i_{t+1}}) + \lambda\nabla r(w)$ where $\{i_{t+1}\}_{t=0}^{\infty}$ are i.i.d. random variables following the uniform distribution on $\{1, \ldots, n\}$. Hence, the stochastic noise is $\epsilon_{t+1}(w) = \nabla_w\ell(w, z_{i_{t+1}}) - \frac{1}{n}\sum_{i=1}^{n}\nabla_w\ell(w, z_i)$. Furthermore, we note that the above setup can extend to the case involving the data augmentation by replacing the loss $\ell(w, z_i)$ with $\mathbb{E}[\ell(w, Z_i)]$, where Z_i is a random variable of an augmented example, and replacing the stochastic gradient of the loss $\nabla_w\ell(w_t, z_{i_{t+1}})$ with $\nabla_w\ell(w_t, Z_{i_{t+1}})$.

2.2 Alternative view of SGD

An alternative view (Kleinberg et al., 2018) of SGD is key in our analysis connecting the stochastic gradient noise with the smoothing for the objective function.

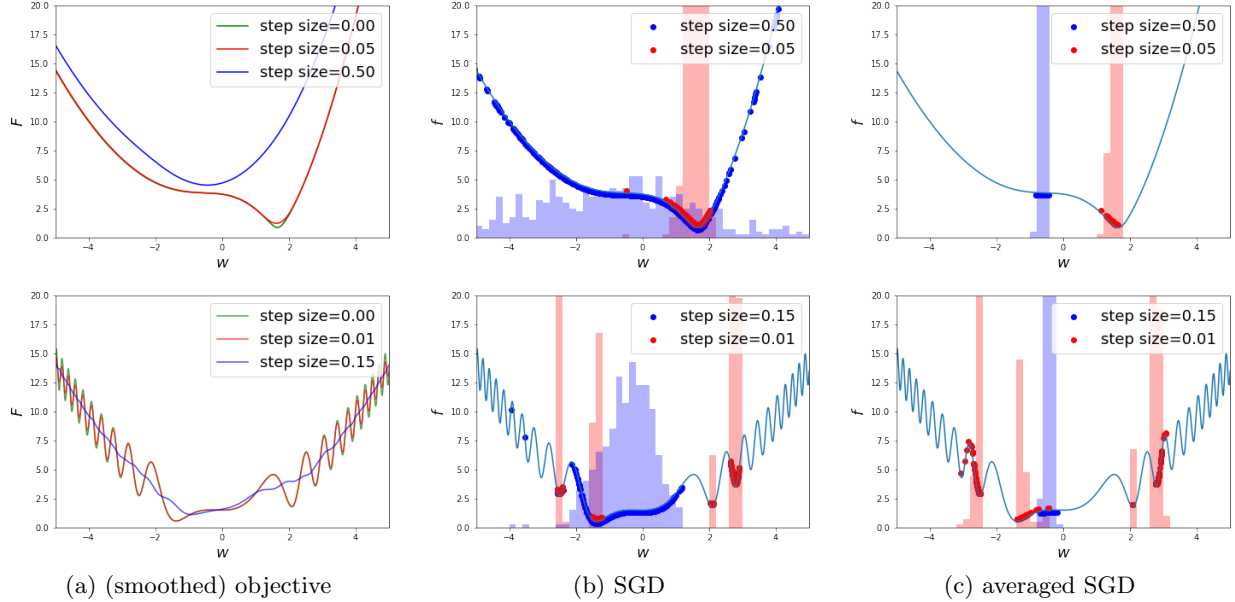| (a) (smoothed) objective | (b) SGD | (c) averaged SGD |

Figure 1: We run SGD and averaged SGD 500 times with the uniform stochastic gradient noise for two objective functions (top and bottom). Figure (a) depicts the objective function f (green, $\eta = 0$) and smoothed objectives F (red and blue, $\eta > 0$). Figures (b) and (c) plot convergent points by SGD and averaged SGD with histograms, respectively.
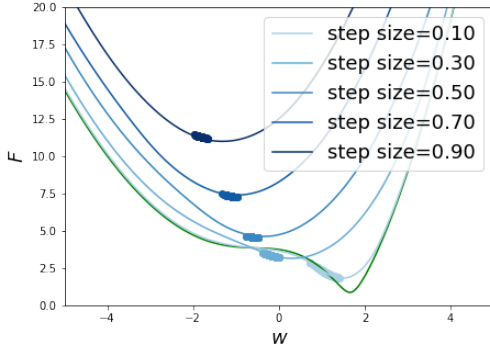

Figure 2: The figure plots the original objective (green), smoothed objectives (blue, darker is smoother), and convergent points obtained by the averaged SGD which is run 500 times for each step size $\eta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

This view of SGD translates $\{w_t\}_{t=0}^{\infty}$ into associated iterations $\{v_t\}_{t=0}^{\infty}$ and we analyze the update of v_t instead of w_t. We define v_t as a parameter obtained by the exact gradient descent from w_t, that is, $v_t = w_t - \eta \nabla f(w_t)$. Since $w_{t+1} = v_t - \eta \epsilon_{t+1}(w_t)$, we get $v_{t+1} = v_t - \eta \epsilon_{t+1}(w_t) - \eta \nabla f(v_t - \eta \epsilon_{t+1}(w_t))$, where the notation $\nabla f(v_t - \eta \epsilon_{t+1}(w_t))$ represents $\nabla f(w)|_{w=v_t - \eta \epsilon_{t+1}(w_t)}$. As shown in Appendix A.1, under a specific setting given later, the translation $w \mapsto v = w - \eta \nabla f(w)$ is a smooth invertible injection and its inverse is differentiable. Thus, we can identify $\epsilon'_{t+1}(v)$ with $\epsilon_{t+1}(w)$ through the map $w \mapsto v$. Then,

we get an update rule of v_t:

$$v_{t+1} = v_t - \eta \epsilon'_{t+1}(v_t) - \eta \nabla f(v_t - \eta \epsilon'_{t+1}(v_t)). \quad (2)$$

For convenience, we refer to the rule (2) as an *alternative stochastic gradient descent* in this paper. Since, the conditional expectation of $\epsilon'_{t+1}(v_t)$ at v_t is zero, we expect that the alternative SGD (2) minimizes the following smoothed objective function:

$$F(v) = \mathbb{E}[f(v - \eta \epsilon'(v))], \quad (3)$$

where ϵ' is an independent copy of $\epsilon'_1, \epsilon'_2, \ldots$. However, we note that the alternative SGD is a biased SGD because $\nabla F(v_t) \neq \mathbb{E}[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))]$ in general [1], unless $\epsilon'_{t+1}(v)$ is free from v.

The function (3) is indeed a smoothed function of f by the convolution using the stochastic gradient noise $\eta \epsilon'$, where the step size η controls the strength of smoothness. Figure 1 depicts how a nonconvex function f is smoothened. In this figure, we observe that by using an appropriately large step size, sharp local minima of f vanishes and the solution of F emerges at the flat area of f. Moreover, by taking Taylor expansion of f, we see that $F(v)$ is a regularized objective that penalizes the high (positive) curvature of f along the

---

[1]We see by the chain rule $\nabla F(v) = \mathbb{E}[(I - \eta J_{\epsilon'(v)}^{\top}(v)) \nabla f(v - \eta \epsilon'(v))]$. Here, the notation $\nabla f(v - \eta \epsilon'(v))$ represents $\nabla f(w)|_{w=v-\eta\epsilon'(v)}$. Hence, $\nabla F(v) \neq \mathbb{E}[\nabla f(v - \eta \epsilon'(v))]$ unless $J_{\epsilon'(v)}^{\top}(v) = 0$.

noise direction in expectation:

$$F(v) = f(v) + \frac{\eta^2}{2} \text{Tr} \left(\nabla^2 f(v) \mathbb{E}[\epsilon'(v)\epsilon'(v)^\top] \right) + O(\eta^3).$$

Therefore, we can expect that stochastic gradient descent avoids sharp minima and converges to a flat region.

Kleinberg et al. (2018) showed the convergence to a point $v_\circ \in \mathbb{R}^d$ under the regularity condition: $\mathbb{E}[\nabla f(v - \eta\epsilon'(v))]^\top (v - v_\circ) \geq \exists c \|v - v_\circ\|^2$ via the alternative view of SGD. Inspired by their work, we analyze the optimization capability of the averaged SGD for approximating the minimizer $v_* = \arg\min_{v \in \mathbb{R}^d} F(v)$.

3 CONVERGENCE ANALYSIS

Izmailov et al. (2018) empirically demonstrated that averaged SGD converges to a flat region and achieves superior generalization when using a relatively large step size such that SGD oscillates. We theoretically explain this phenomenon by showing that the averaged SGD can get closer to v_* than SGD using the same step size under certain settings. In the averaged SGD, we run SGD (1) and take the average as follows: $\overline{w}_{T+1} = \frac{1}{T+1} \sum_{t=1}^{T+1} w_t$.

3.1 Analysis of averaged SGD

Our aim is to show $\lim_{T\to\infty} \overline{w}_T$ can be closer to v_* than $\{w_t\}_{t=0}^\infty$ and $\{v_t\}_{t=0}^\infty$. Preferably, the alternative view of SGD (2) is more useful in analyzing the averaged SGD because the average $\overline{v}_T = \frac{1}{T+1} \sum_{t=0}^T v_t$ is consistent with $\overline{w}_T$ as confirmed below. By definition,

$$\overline{w}_{T+1} = \overline{v}_T + \frac{\eta}{T+1} \sum_{t=0}^T \epsilon_{t+1}(w_t),$$

where the noise term $\sum_{t=0}^T \epsilon_{t+1}(w_t)/(T+1)$ is zero in expectation and its variance is upper bounded by $\sigma_1^2/(T+1)$ under Assumption **(A2)**. Hence, $\overline{w}_{T+1} - \overline{v}_T$ converges to zero in probability by Chebyshev's inequality; for any $r > 0$, $\mathbb{P}[\|\overline{w}_{T+1} - \overline{v}_T\| > r] \leq \sigma_1^2/(T+1)r^2 \to 0$ as $T \to \infty$, and the analysis of $\lim_{T\to\infty} \overline{w}_T$ reduces to that of $\lim_{T\to\infty} \overline{v}_T$.

We can immediately see that the averaged SGD is always closer to the solution than SGD in average because of Jensen's inequality: $\mathbb{E}[\|\overline{v}_T - v_*\|^2] \leq \frac{1}{T+1} \sum_{t=0}^T \mathbb{E}[\|v_t - v_*\|^2]$. We next present a nontrivial bound on the left-hand side of this inequality (i.e., the error achieved by the averaged SGD). To do so, we make the following assumptions on the objective function and stochastic gradient noise.

Assumption 3.1.

(A1) $f : \mathbb{R}^d \to \mathbb{R}$ is nonnegative, twice continuously differentiable, and its Hessian is bounded, i.e., there is a constant $L > 0$ such that for any $w \in \mathbb{R}^d$, $-LI \preceq \nabla^2 f(w) \preceq LI$.

(A2) Random fields $\{\epsilon_{t+1}\}_{t=0}^\infty$ are independent copies each other[2] and $\epsilon_{t+1}(w)$ is differentiable in w. Moreover, for any $w \in \mathbb{R}^d$ $\mathbb{E}[\epsilon_{t+1}(w)] = 0$ and there are $\sigma_1, \sigma_2 > 0$ such that for any $w \in \mathbb{R}^d$, $\mathbb{E}[\|\epsilon_{t+1}(w)\|^2] \leq \sigma_1^2$ and $\mathbb{E}[\|J_{\epsilon_{t+1}}^\top(w)\|_2] \leq \sigma_2$, where $J_{\epsilon_{t+1}}$ is Jacobian of ϵ_{t+1}.

(A3) Hessian of F at v_* is positive definite, i.e., there is a positive constant $\mu > 0$ such that $\nabla^2 F(v_*) \succeq \mu I$.

(A4) There exist positive constants γ and M such that for any $v \in \mathbb{R}^d$, $\|\nabla F(v) - \nabla^2 F(v_*)(v - v_*)\| \leq \gamma + M\|v - v_*\|^2$.

The smoothness and boundedness conditions **(A1)** on the objective function and the zero-mean and the bounded variance conditions **(A2)** on stochastic gradient noise are commonly assumed in the convergence analysis for the stochastic optimization methods. Standard stochastic gradient noises $\{\epsilon_{t+1}\}_{t=0}^\infty$ defined in Example 2.1 are independent copies. Moreover, if Hessian matrix satisfies $-LI \preceq \nabla_w^2 \ell(w, z) \preceq LI$ in Example 2.1, then the last condition on $J_{\epsilon_{t+1}}$ also holds with at least $\sigma_2 = 2L$ because $J_{\epsilon_{t+1}}(w) = \nabla_w^2 \ell(w, z_{i_{t+1}}) - \nabla^2 f(w)$.

(A3) and **(A4)** are assumptions made on the smoothed objective function and used to connect the error of the averaged SGD with that of SGD. **(A3)** ensures the strict positivity of Hessian $\nabla^2 F(v_*)$. **(A4)** measures the linear approximation error of ∇F at v_*. For instance, functions that have bounded third-order derivatives satisfy **(A4)** with $\gamma = 0$, which is also assumed in Dieuleveut et al. (2020) to show the superiority of the averaging scheme for strongly convex optimization problems.

The following theorem describes the relationship between the errors achieved by SGD and averaged SGD.

Theorem 3.2. *Under Assumptions **(A1)**–**(A4)**, run the averaged SGD for T-iterations with the step size $\eta \leq \frac{1}{2L}$, then $\overline{v}_T$ satisfies the following inequality:*

$$\mathbb{E}[\|\overline{v}_T - v_*\|] \leq \min \left\{ D_T, \ O\left(T^{-\frac{1}{2}}\right) \right.$$

$$\left. + \frac{4\sigma_1\sigma_2\eta^{\frac{3}{2}}L^{\frac{1}{2}}}{\sqrt{3}\mu} + \frac{2D_{T+1}}{\eta\mu\sqrt{T+1}} + \frac{\gamma + MD_T^2}{\mu} \right\},$$

where we set $D_T := \sqrt{\frac{1}{T+1}\mathbb{E}\left[\sum_{t=0}^T \|v_t - v_*\|^2\right]}$.

---

[2]We suppose $\{\epsilon_{t+1}\}_{t=0}^\infty$ are independent copies each other. That is, there is a measurable map from a probability space: $\Omega \ni z \mapsto \epsilon(w, z) \in \mathbb{R}^d$, and then ϵ_{t+1} can be written as a measurable map from a product probability space: $\Omega^{\mathbb{Z}_{\geq 0}} \ni \{z_{s+1}\}_{s=0}^\infty \mapsto \epsilon(w, z_{t+1}) \in \mathbb{R}^d$ when explicitly representing them.

Proof sketch. We set $R(v) = \nabla F(v) - \nabla^2 F(v_*)(v - v_*)$. Taking average of $R(v_t)$ over $t \in \{0, 1, \ldots, T\}$, we get $\nabla^2 F(v_*)(\overline{v}_T - v_*) = \frac{1}{T+1} \sum_{t=0}^{T} \nabla F(v_t) - \frac{1}{T+1} \sum_{t=0}^{T} R(v_t)$. Therefore, by taking the expectation of the norm of both sides, we obtain

$$\mu \mathbb{E}\left[\|\overline{v}_T - v_*\|\right] \leq \frac{1}{T+1} \mathbb{E}\left[\left\|\sum_{t=0}^{T} \nabla F(v_t)\right\|\right] + \gamma + M D_T^2,$$

where we applied **(A3)** and **(A4)**. Since the first term of the RHS is the norm of the average of gradients $\nabla F(v_t)$ across all iterations, this term will decrease up to a small error as the optimization proceeds:

$$O(T^{-1/2}) + \frac{4}{\sqrt{3}} \sigma_1 \sigma_2 \eta^{\frac{3}{2}} L^{\frac{1}{2}} + \frac{2 D_{T+1}}{\eta \sqrt{T+1}}.$$

See Lemma A.5 in Appendix for the detailed derivation of this error. Moreover, Jensen's inequality yields $\mathbb{E}[\|\overline{v}_T - v_*\|] \leq \sqrt{\mathbb{E}[\|\overline{v}_T - v_*\|^2]} \leq D_T$. □

The term $D_T = \sqrt{\frac{1}{T+1} \mathbb{E}\left[\sum_{t=0}^{T} \|v_t - v_*\|^2\right]}$ in Theorem 3.2 is the root mean square of the distances between v_* and SGD iterations. Hence D_T converges up to an error depending on η under certain assumptions as shown in Proposition 3.4 later or at least it is uniformly bounded as long as SGD performs in a compact domain. Hence, we suppose $D_\infty = \lim_{T \to \infty} D_T < \infty$ in the following argument.

Let us see the relationship between errors finally achieved by SGD and averaged SGD. Taking the limit $T \to \infty$, we get the upper bound on $\lim_{T \to \infty} \mathbb{E}[\|\overline{v}_T - v_*\|]$ as follows:

$$\min\left\{D_\infty, \frac{4 \sigma_1 \sigma_2 \eta^{\frac{3}{2}} L^{\frac{1}{2}}}{\sqrt{3}\mu} + \frac{\gamma + M D_\infty^2}{\mu}\right\}. \quad (4)$$

The bound (4) shows nontrivial improvement by averaged SGD over SGD in the sense: $\lim_{T \to \infty} \mathbb{E}[\|\overline{v}_T - v_*\|] \ll D_\infty$ when SGD error D_∞ satisfies the following condition:

$$\frac{4 \sigma_1 \sigma_2 \eta^{\frac{3}{2}} L^{\frac{1}{2}}}{\sqrt{3}\mu} + \frac{\gamma}{\mu} \ll D_\infty \ll \frac{\mu}{M}. \quad (5)$$

Here, the informal notation $a \ll b$ (for $a, b > 0$) is used for describing the situation where b is sufficiently large such that a/b approaches 0. To discuss the condition (5), consider the case where μ, $M > 0$ satisfying Assumptions **(A3)** and **(A4)** can be uniformly chosen with respect to $\eta \to 0$ and $\gamma = 0$. Then, the upper bound $D_\infty \ll \mu/M = O(1)$ means the convergence of SGD to some extent. And the lower bound $\Omega(\eta^{3/2}) \ll D_\infty$ is also typically satisfied for mildly small η because SGD oscillates at least of the order of η if the stochastic gradient does not vanish (see Proposition 3.5).

3.2 Evaluation of SGD error D_∞

In this section, we estimate the value D_∞ under reasonable problem setups. Inspired by Kleinberg et al. (2018)[3], we make the following regularity condition on the smoothed objective F at v_*.

Assumption 3.3. (A5) There exists $c > 0$ such that for any $v \in \mathbb{R}^d$, $\nabla F(v)^\top (v - v_*) \geq c\|v - v_*\|^2$.

(A5) is a sort of one-point strong convexity at the solution v_*, but we note that it allows for the nonconvexity of F. Under this assumption, we obtain the upper bound on SGD error D_T.

Proposition 3.4. *Under Assumptions* **(A1)**, **(A2)**, *and* **(A5)**, *run SGD for T-iterations with the step size $\eta \leq \frac{1}{2L}$, then we get*

$$D_T^2 \leq O(T^{-1}) + \frac{2\eta\sigma_1^2}{c} + \frac{8\eta^2\sigma_1^2 L}{3c}\left(1 + \frac{2\eta\sigma_2^2}{c}\right).$$

Our result allows for a larger step size than that in Kleinberg et al. (2018) because we do not stick to the convergence of the last iterate. Proposition 3.4 implies $D_\infty \leq \sqrt{\frac{2\eta\sigma_1^2}{c} + \frac{8\eta^2\sigma_1^2 L}{3c}\left(1 + \frac{2\eta\sigma_2^2}{c}\right)}$, meaning convergence of SGD to a point at least at a distance of the RHS from the minimizer v_* of F.

The next proposition provides the lower bound on SGD error D_T under the assumption that the variance of stochastic gradient noise is uniformly lower bounded.

Proposition 3.5. *Suppose that Assumptions* **(A1)** *and* **(A2)** *hold, and that there exists $\sigma_3 > 0$ such that for any $w \in \mathbb{R}^d$, $\mathbb{E}[\|\epsilon_{t+1}(w)\|^2] \geq \sigma_3^2$. Running SGD with $\eta \leq \frac{3\sigma_3^2}{32\sigma_1^2 L}$, we get*

$$\frac{\eta^2\sigma_3^2}{8} \leq D_T^2 + O(T^{-1}).$$

Propositions 3.4 and 3.5 with $T \to \infty$ yield the estimation of D_∞. That is, we get that under the assumptions made in these two propositions,

$$\frac{\eta\sigma_3}{2\sqrt{2}} \leq D_\infty \leq \sqrt{\frac{2\eta\sigma_1^2}{c} + \frac{8\eta^2\sigma_1^2 L}{3c}\left(1 + \frac{2\eta\sigma_2^2}{c}\right)}. \quad (6)$$

Summarizing the discussion so far, the averaged SGD achieves a nontrivial improvement if the above estimation Eq. (6) satisfies the condition Eq. (5).

---

[3]**(A5)** is slightly different from that in Kleinberg et al. (2018). They assumed $\mathbb{E}[\nabla f(v - \eta\epsilon'(v))]^\top (v - v_\circ) \geq \exists c\|v - v_\circ\|^2$ with a point $v_\circ$ which does not necessarily coincides with v_*. If $J_{\epsilon'} \equiv 0$, then $v_\circ$ equals v_* and both assumptions coincide because $\nabla F(v) = \mathbb{E}[\nabla f(v - \eta\epsilon'(v))]$ in this case.

3.3 Example

As a sanity check, we present examples that satisfy the improvement condition Eq. (5) by instantiating the estimation Eq. (6) of SGD error D_∞ and we confirm the certain improvement by the averaged SGD. For a detailed explanation, see Appendix B and C.
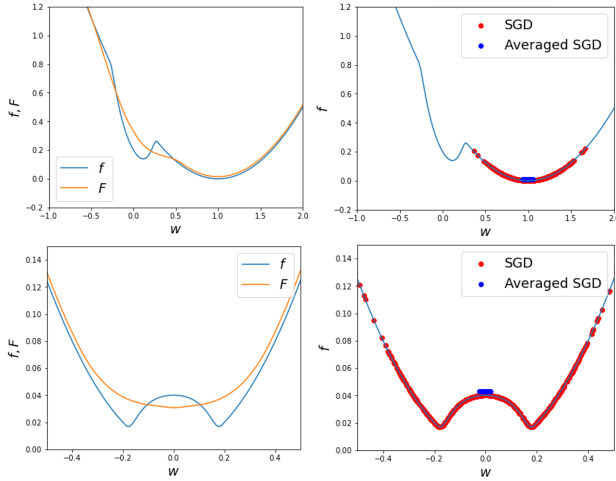

Figure 3: We run SGD and averaged SGD for two problems corresponding top and bottom figures. For each case, the left figure depicts the original objective (blue) f and smoothed objective (orange) F and the right figure depicts convergent points of SGD (red) and averaged SGD (blue).

Example 3.6. We consider two examples of functions f on $\mathbb{R}$ which fall into the setting of $\gamma = 0$ or $M = 0$. Blue curves in the top and bottom of Figure 3 depict these functions, respectively. As the stochastic gradient noises, we take i.i.d. uniform noise on an interval $[-r, r]$ in both cases. Hence, $\sigma_1^2 = \sigma_3^2 = r^2/3$ and $\sigma_2 = 0$. Then, the smoothed objectives F (orange curves in Figure 3) with an appropriate step size η look like the quadratic functions that possess the unique global minimizers v_*. In both cases, we observe the improved performance of the averaged SGD regarding the convergence to v_*. We verify the condition Eq. (5) below. For the details, see Appendix.

1. **First case (top of Figure 3):** An objective is $f(w) = 0.5(w - 1)^2 + g_\delta(w)$, where g_δ is a downward bump of width δ and height δ. This bump makes a local minimum of f around $v = 0$ and f has the optimal solution $v = 1$. Lipschitz constant is $L = 1 + O(1/\delta)$. In this setting, the smoothing only affects the term g_δ, i.e.,

$$F(v) = \frac{1}{2}(v - 1)^2 + \mathbb{E}[g_\delta(v - \eta\epsilon')] + (const).$$

Then, the local minimum of f around $v = 0$ is

eliminated by the smoothing with an appropriate η, and the global solution $v_* = 1$ of F only remains. In fact, if we take $\eta = O(\delta/r)$, we can verify all assumptions with constants: $\mu = 1$, $c = 1/3$, $\gamma = 0$, $M = 8/9$. By instantiating Eq. (6), we see

$$\delta \lesssim D_\infty \lesssim \delta + \sqrt{\delta(1 + r)}.$$

Hence, the condition Eq. (5): $0 \ll D_\infty \ll 9/8$ is satisfied under the setting with mildly small δ. Specifically, we have $\lim_{T \to \infty} \mathbb{E}[\|\bar{v}_T - v_*\|] \leq \frac{8}{9}D_\infty^2$.

2. **Second case (bottom of Figure 3):** An objective is $f(w) = 0.5w^2 + g_\delta(w)$, where g_δ is a upward bump of width δ and height δ^2. f has two global minima beside the bump. Lipschitz constant is $O(1)$ regardless of δ. As in the previous example, we see

$$F(v) = \frac{1}{2}v^2 + \mathbb{E}[g_\delta(v - \eta\epsilon')] + (const).$$

Then, two global minima of f are eliminated by the smoothing with an appropriate η and a unique global minimum $v_* = 0$ of F emerges. In this case, we do not need to make η small depending on δ. In fact, as long as ηr greater than δ by a certain amount, we can verify all assumptions with constants: $\mu = 1$, $c = 1/2$, $\gamma = \exists C\delta^2/(\eta r)$, $M = 0$. By instantiating Eq. (6), we obtain the estimation of D_∞ as follows.

$$\eta r \lesssim D_\infty \lesssim \sqrt{\eta}r + \eta r.$$

Hence, the condition Eq. (5): $C\delta^2/(\eta r) \ll D_\infty \ll \infty$ is satisfied under the setting $\delta^2 \ll \eta^2 r^2$. Specifically, we have $\lim_{T \to \infty} \mathbb{E}[\|\bar{v}_T - v_*\|] \leq C\delta^2/(\eta r)$.

4 EXPERIMENTS

We evaluate the empirical performance of SGD and averaged SGD on image classification tasks using CIFAR10 and CIFAR100 datasets. To evaluate the usefulness of the parameter averaging for the other methods, we also compare SAM (Foret et al., 2020) with its averaging variant. For the parameter averaging, we employ the tail-averaging scheme where the average is taken over the last phase of training.

We use the CNN architectures: ResNet (He et al., 2016) with 50-layers (ResNet-50), WideResNet (Zagoruyko and Komodakis, 2016) with 28 layers and width 10 (WRN-28-10), and Pyramid Network (Han et al., 2017) with 272 layers and widening factor 200. In all settings, we use the standard data augmentations:

Table 1: Decay schedules for (averaged) SGD.

η	milestones
s	{80,160,240}
m	{80,160}
l	{300}

horizontal flip, normalization, padding by four pixels, random crop, and cutout (DeVries and Taylor, 2017), and we employ the weight decay with the coefficient 0.005. Moreover, we use the multi-step strategy for the step size, which decays the step size by a factor once the number of epochs reaches one of the given milestones. To see the dependence on the step size, we use two decay schedules for the parameter averaging.

Table 1 summarizes milestones labeled by the symbols: 's', 'm', and 'l'. The initial step size and a decay factor of the step size are set to 0.1 and 0.2 in all cases. The averages are taken from 300 epochs for the schedules 's' and 'l', and from 160 epochs for the schedule 'm'. These hyperparameters were tuned based on the validation performance.

For a fair comparison, we run (averaged) SGD with 400 epochs and (averaged) SAM with 200 epochs because SAM requires two gradients per iteration, and thus the milestones and starting epoch of taking averages are also halved for (averaged) SAM. We evaluate each method 5 times for ResNet-50 and WRN-28-10, and 3 times for Pyramid network. The averages of classification accuracies are listed in Table 2 with the standard deviations in brackets. We observe from the table that the parameter averaging for SGD improves the classification accuracies in all cases, especially on CIFAR100 dataset. Eventually, the averaged SGD achieves comparable performance with SAM. Moreover, we also observe improvement by parameter averaging for SAM in most cases, which is consistent with the observations in Kaddour et al. (2022). A similar improvement is also observed for the cosine annealing learning rate (see Appendix D).
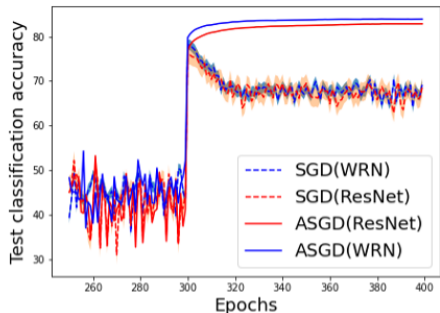

Figure 4: Test accuracies achieved by SGD and averaged SGD on CIFAR100 with ResNet-50 and WRN-28-10.

Comparing results on CIFAR100 and CIFAR10, the large step size is better, and the small step size is relatively poor on CIFAR100 dataset, whereas the small step size generally works on CIFAR10 dataset. If we use the step-size strategy 'l' for CIFAR10, then the improvement becomes small (see Appendix D for this

result). This is because the strong smoothing brought by a large step size can be harmful to simple datasets such that the normal SGD already achieves high accuracies. Moreover, we note that the averaged SGD on CIFAR100 quite works well with the large step-size schedule 'l', even though SGD itself is unstable and poorly performs under this schedule as seen in Figure 4. Indeed, the accuracy of SGD temporarily increases at the 300 epochs because of the decay of the step size, and it decreases thereafter. However, the parameter averaging brings significant improvement in accuracy even under such a situation.

Next, we observe in Figure 5 that the loss landscape around the convergent point is in better shape and forms an asymmetric valley as observed by He et al. (2019). Specifically, Figure 5 depicts the section of train and test loss functions across parameters obtained by the averaged SGD and SGD. The middle figure is the close-up view at the edge and plots each parameter. The right figure depicts the train and test losses, and the smoothed train losses with Gaussian noises in log-scale. We observe in Figure 5 the phenomenon that SGD converges to an edge and averaged SGD converges to a flat side. Our theory helps explain this phenomenon because the minimizer of the asymptotic valley can be shifted to a flat side by the smoothing as confirmed in the right of Figure 5 as well as a synthetic setting (Figure 2). Moreover, the right figure indicates the possibility that the smoothed objective with appropriate stochastic gradient noise well approximates test loss. In fact, we observe that averaged SGD achieves a lower test loss which makes about 2% improvement in the classification error on CIFAR100 dataset. These observations are consistent with the experiments conducted in He et al. (2019).

Finally, to observe the non-trivial bias of the averaged SGD, we run SGD with the smaller step size from the parameter obtained by the averaged SGD with 600 epochs. Then, we observe in Figure 6 that the test loss is gradually getting worse, whereas the training loss is getting better. This result proves that the averaged SGD with a large step size does not necessarily converge to a local minimum of the training loss, but converges to a parameter achieving a smaller test loss. Moreover, we confirm that the test accuracy also decreases from 83.35 attained by the averaged SGD to 82.42.

CONCLUSION

Izmailov et al. (2018) observed the averaged SGD with a large step size finds a flatter solution than SGD and gave an intuitive explanation of this phenomenon. To support their observation, we analyzed the convergence capability of the averaged SGD for the smoothed ob-

Table 2: Comparison of test classification accuracies on CIFAR100 and CIFAR10 datasets.

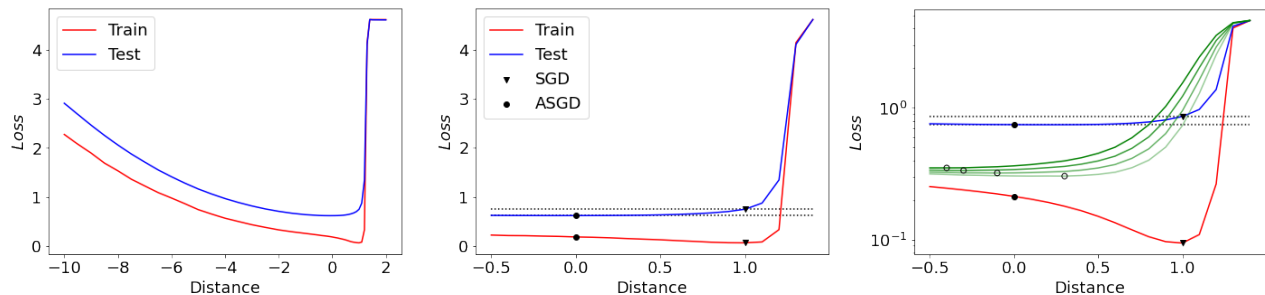|		CIFAR100				CIFAR10		
	η	ResNet-50	WRN-28-10	Pyramid	η	ResNet-50	WRN-28-10	Pyramid
SGD	s	80.83 (0.21)	81.81 (0.29)	81.43 (0.32)	s	95.95 (0.11)	96.85 (0.16)	96.41 (0.22)
Averaged SGD	s	82.13 (0.22)	83.13 (0.13)	84.23 (0.03)	s	96.58 (0.14)	97.24 (0.07)	97.07 (0.08)
	l	**82.87 (0.13)**	**84.23 (0.10)**	**85.12 (0.20)**	m	**96.89 (0.05)**	**97.44 (0.04)**	**97.28 (0.13)**
SAM	s	82.56 (0.14)	83.80 (0.27)	84.59 (0.24)	s	**96.34 (0.12)**	97.14 (0.05)	97.34 (0.03)
Averaged SAM	s	82.64 (0.12)	84.09 (0.30)	85.40 (0.12)	s	96.33 (0.10)	**97.21 (0.05)**	97.34 (0.03)
	l	**82.73 (0.28)**	**84.55 (0.17)**	**86.00 (0.04)**	m	96.31 (0.11)	97.20 (0.06)	**97.35 (0.06)**

Figure 5: Sections of the train (red) and test (blue) loss landscapes across the parameters obtained by averaged SGD (distance=0) and SGD (distance=1) for ResNet-50 with CIFAR100 dataset. SGD is run with a small step size after running averaged SGD with a large step size. The middle figure is the close-up view at the edge. The triangle and circle markers represent convergent parameters by SGD and averaged SGD, respectively. The right figure plots smoothed train loss functions (green, darker is smoother) with Gaussian noises in addition to train and test losses. The blank circles are the minimizers of smoothed objectives.
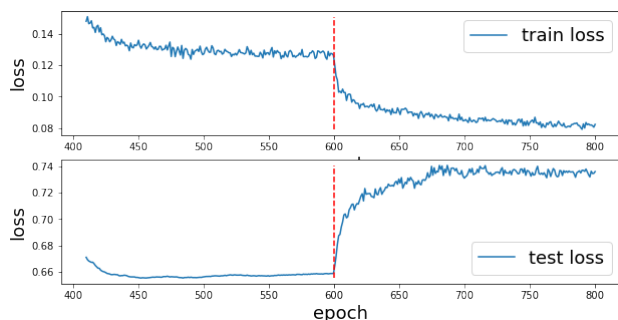

Figure 6: For ResNet-50 with CIFAR100 dataset, we run SGD for 200 epochs from a parameter obtained by the averaged SGD with 600 epochs. The figure depicts training loss (top) and test loss (bottom). The red line is a change point of the methods. The learning rate for SGD is eventually annealed to 0 from 0.02 used for the final phase of averaged SGD.

jective F via the alternative view of SGD (Kleinberg et al., 2018). Specifically, we derived the upper and lower bound conditions (Theorem 3.2 and Eq. (5)) for

SGD error D_∞ so that the averaged SGD converges closer to $v_* = \arg\min F(v)$ than SGD. Furthermore, to verify these conditions, we estimated D_∞ under additional lower bound on the stochastic gradient noise and sort of one-point strong convexity for F. Finally, we empirically observed that the averaged SGD with a large step size achieved superior performance on the image classification task and confirmed that the obtained parameter differs from that of SGD.

Based on our findings, we suggest using large step sizes for difficult datasets to learn so that SGD itself oscillates and performs somewhat poorly. Then, by averaging the parameters, we can expect a significant improvement in generalization.

One limitation is that to verify the improvement condition: Eq. (5), we imposed a sort of one-point strong convexity on F in Proposition 3.4, which is difficult to satisfy for neural networks perfectly. We believe Eq. (5) is met in a broader problem setting and will help us understand the actual behavior of averaged SGD in deep learning. Further exploration of this condition is left to future work.

References

Ahn, K., Zhang, J., and Sra, S. (2022). Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 247–257.

Arora, S., Li, Z., and Panigrahi, A. (2022). Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 948–1024.

Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 451–459.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. (2019). Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2021). Gradient descent on neural networks typically occurs at the edge of stability. In *Proceedings of the 9th International Conference on Learning Representations*.

Damian, A., Ma, T., and Lee, J. D. (2021). Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461.

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Dieuleveut, A., Durmus, A., and Bach, F. (2020). Bridging the gap between constant step size stochastic gradient descent and markov chains. *Annals of Statistics*, 48(3):1348–1382.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1019–1028.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2020). Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the 8th International Conference on Learning Representations*.

Han, D., Kim, J., and Kim, J. (2017). Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5927–5935.

Haruki, K., Suzuki, T., Hamakawa, Y., Toda, T., Sakai, R., Ozawa, M., and Kimura, M. (2019). Gradient noise convolution (gnc): Smoothing loss function for distributed large-batch sgd. *arXiv preprint arXiv:1906.10822*.

He, H., Huang, G., and Yuan, Y. (2019). Asymmetric valleys: beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems*, volume 32, pages 2553–2564.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13.

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

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence*, pages 876–885.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2020). Fantastic generalization measures and where to find them. In *Proceedings of the 8th International Conference on Learning Representations*.

Kaddour, J., Liu, L., Silva, R., and Kusner, M. (2022). When do flat minima optimizers work? *Advances in Neural Information Processing Systems*.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., and Smelyanskiy, M. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the 5th International Conference on Learning Representations*.

Keskar, N. S. and Socher, R. (2017). Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*.

Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does sgd escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2698–2707.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an o $(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. (2020). The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*.

Lin, T., Kong, L., Stich, S., and Jaggi, M. (2020). Extrapolation for large-batch training in deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6094–6104.

Luo, L., Xiong, Y., Liu, Y., and Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. In *Proceedings of the 7th International Conference on Learning Representations*.

McAllester, D. A. (1998). Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234.

McAllester, D. A. (1999). Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170.

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

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring generalization in deep learning. *Advances in neural information processing systems*, 30.

Orvieto, A., Raj, A., Kersting, H., and Bach, F. (2022). Explicit regularization in overparametrized models via noise injection. *arXiv preprint arXiv:2206.04613*.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of International Conference on Machine Learning 29*, pages 1571–1578.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Wen, W., Wang, Y., Yan, F., Xu, C., Wu, C., Chen, Y., and Li, H. (2018). Smoothout: Smoothing out sharp minima to improve generalization in deep learning. *arXiv preprint arXiv:1805.07898*.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30.

Yu, L., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. (2020). An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. *arXiv preprint arXiv:2006.07904*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. (2020). Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A   Proofs

A.1   Alternative Stochastic Gradient Descent

Let denote by $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ a change of variables from w to v introduced in Section 2.2, i.e., $v = \varphi(w) = w - \eta \nabla f(w)$.

Lemma A.1. *Under Assumption* **(A1)** *and* $\eta \leq \frac{1}{2L}$, *the function* φ *is injective and invertible, and its inverse* φ^{-1} *defined on on* $\mathrm{Im}\varphi$ *is differentiable.*

Proof. For $w, w' \in \mathbb{R}^d$, we suppose $\varphi(w) = \varphi(w')$. Then, it holds that

$$\|w - w'\| = \eta \|\nabla f(w) - \nabla f(w')\| \leq \eta L \|w - w'\| \leq \frac{1}{2} \|w - w'\|,$$

where we used L-Lipschitz continuity of ∇f due to **(A1)**. Therefore, we see $w = w'$ and φ is an injection. Moreover, since $J_\varphi(w) = I - \eta \nabla^2 f(w) \succeq (1 - \eta L)I \succeq \frac{1}{2}I$. Thus, φ is invertible and φ^{-1}, which is defined on $\mathrm{Im}\varphi$, is differentiable because of the injectivity and the inverse map theorem. $\qquad\square$

Using φ, we see $\epsilon'(v) = \epsilon(\varphi^{-1}(v))$ for $v \in \mathrm{Im}\varphi$. Let $(\Omega, \mathcal{F}, P)$ be a probability space such that $\epsilon'(v)$ can be represented as a measurable map $z \in \Omega \mapsto \epsilon'(v, z)$. Note that we use $\epsilon'(v)$ and $\epsilon'(v, z)$ depending on the situation. For a function $g : \mathbb{R}^d \to \mathbb{R}^d$, we denote by $J_g(w)$ Jacobian of g, i.e., $J_g(w) = (\partial g_i(w)/\partial w_j)_{i,j=1}^d$.

Lemma A.2. *Under Assumptions* **(A1)** *and* **(A2)**, *we get for any* $v \in \mathrm{Im}\varphi \subset \mathbb{R}^d$,

$$\nabla F(v) = \mathbb{E}[\nabla f(v - \eta \epsilon'(v))] - \eta \int J_{\epsilon'(\cdot,z)}^\top(v) \nabla f(v - \eta \epsilon'(v, z)) \mathrm{d}P(z).$$

Moreover, if $\eta \leq \frac{1}{2L}$, *then*

$$\|\nabla F(v) - \mathbb{E}[\nabla f(v - \eta \epsilon'(v))]\| \leq 2\eta \sigma_2 \sqrt{\mathbb{E}\left[\|\nabla f(v - \eta \epsilon'(v))\|^2\right]}.$$

Proof. The first equality of the statement can be confirmed by the direct calculation as follows:

$$
\begin{aligned}
\nabla F(v) &= \nabla \mathbb{E}[f(v - \eta \epsilon'(v))] \\
&= \int \nabla(f(v - \eta \epsilon'(v, z))) \mathrm{d}P(z) \\
&= \int (I - \eta J_{\epsilon'(\cdot,z)}^\top(v)) \nabla f(v - \eta \epsilon'(v, z)) \mathrm{d}P(z) \\
&= \mathbb{E}[\nabla f(v - \eta \epsilon'(v))] - \eta \int J_{\epsilon'(\cdot,z)}^\top(v) \nabla f(v - \eta \epsilon'(v, z)) \mathrm{d}P(z).
\end{aligned}
$$

Next, we evaluate the last term below. By the chain rule and inverse map theory,

$$J_{\epsilon'(\cdot,z)}(v) = J_{\epsilon(\phi^{-1}(\cdot),z)}(v) = J_{\epsilon(\cdot,z)}(\phi^{-1}(v)) J_{\phi^{-1}}(v) = J_{\epsilon(\cdot,z)}(\varphi^{-1}(v)) J_\varphi^{-1}(\varphi^{-1}(v)).$$

Note that from assumption for any $w \in \mathbb{R}^d$, $J_\varphi(w) = I - \eta \nabla^2 f(w) \succeq (1 - \eta L)I \succeq \frac{1}{2}I$. Hence,

$$\|J_{\epsilon'(\cdot,z)}^\top(v)\|_2 \leq \|J_\varphi^{-1}(\varphi^{-1}(v))\|_2 \|J_{\epsilon(\cdot,z)}^\top(\varphi^{-1}(v))\|_2 \leq 2\sigma_2.$$

Finally, we get

$$
\begin{aligned}
\left\| \int J_{\epsilon'(\cdot, z)}^\top(v) \nabla f(v - \eta \epsilon'(v, z)) \mathrm{d}P(z) \right\| & \\
\leq \sqrt{\int \|J_{\epsilon'(\cdot, z)}^\top(v)\|_2^2 \|\nabla f(v - \eta\epsilon'(v, z))\|^2 \mathrm{d}P(z)} & \\
\leq 2\sigma_2 \sqrt{\int \|\nabla f(v - \eta\epsilon'(v, z))\|^2 \mathrm{d}P(z)} & \\
\leq 2\sigma_2 \sqrt{\mathbb{E}\left[\|\nabla f(v - \eta\epsilon'(v))\|^2\right]}. &
\end{aligned}
$$

This finishes the proof. $\qquad \square$

The following proposition is the restatement of the well-known convergence result to a stationary point using the coordinate v.

Proposition A.3. *Under Assumptions* **(A1)**, **(A2)**, *and* $\eta \leq \frac{1}{2L}$, *we get*

$$
\sum_{t=0}^T \mathbb{E}\left[\|\nabla f(v_t - \eta\epsilon'_{t+1}(v_t))\|^2\right] \leq \frac{4}{3\eta}\mathbb{E}[f(w_0)] + \frac{2}{3}\eta\sigma_1^2 L(T+2). \tag{7}
$$

Proof. It is known that **(A1)** derives the following (Nesterov, 2004): for any $w, w' \in \mathbb{R}^d$,

$$
f(w') \leq f(w) + \nabla f(w)^\top (w' - w) + \frac{L}{2}\|w' - w\|^2. \tag{8}
$$

Substituting the update Eq. (1) into this inequality with $w' = w_{t+1}$ and $w = w_t$, and taking the conditional expectation $\mathbb{E}[\cdot | \mathcal{F}_t]$, we get

$$
\begin{aligned}
\mathbb{E}[f(w_{t+1}) | \mathcal{F}_t] &\leq f(w_t) - \eta\|\nabla f(w_t)\|^2 + \frac{\eta^2 L}{2}\mathbb{E}\left[\|\nabla f(w_t) + \epsilon_{t+1}(w_t)\|^2 | \mathcal{F}_t\right] \\
&= f(w_t) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(w_t)\|^2 + \frac{\eta^2 L}{2}\mathbb{E}\left[\|\epsilon_{t+1}(w_t)\|^2 | \mathcal{F}_t\right] \\
&\leq f(w_t) - \frac{3\eta}{4}\|\nabla f(w_t)\|^2 + \frac{\eta^2 \sigma_1^2 L}{2}.
\end{aligned}
$$

Thus, we have $\mathbb{E}[f(w_{t+1})] \leq \mathbb{E}[f(w_t)] - \frac{3\eta}{4}\mathbb{E}[\|\nabla f(w_t)\|^2] + \frac{\eta^2\sigma_1^2 L}{2}$. By summing up this inequality, we get

$$
\sum_{t=0}^{T+1} \mathbb{E}[\|\nabla f(w_t)\|^2] \leq \frac{4}{3\eta}\mathbb{E}[f(w_0)] + \frac{2}{3}\eta\sigma_1^2 L(T+2),
$$

where we used the nonnegativity of f. By dropping the term with $t = 0$ of the sum in the left hand side and using $w_{t+1} = v_t - \eta\epsilon'_{t+1}(v_t)$, we finally get

$$
\sum_{t=0}^T \mathbb{E}[\|\nabla f(v_t - \eta\epsilon'_{t+1}(v_t))\|^2] \leq \frac{4}{3\eta}\mathbb{E}[f(w_0)] + \frac{2}{3}\eta\sigma_1^2 L(T+2).
$$

$\qquad \square$

A.2 Proof of Theorem 3.2

We give technical lemmas for proving Theorem 3.2.

Lemma A.4. *Under Assumptions* **(A1)** *and* **(A2)**, *run the stochastic gradient descent with T-iterations with the step size $\eta \leq \frac{1}{2L}$, then the alternative SGD satisfies the following inequality:*

$$\mathbb{E}\left[\left\|\sum_{t=0}^{T} \nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\right\|\right] \leq \sigma_1 \sqrt{T} + \frac{1}{\eta}\sqrt{2\sum_{t=0}^{T+1} \mathbb{E}\left[\|v_t - v_*\|^2\right]},$$

$$\mathbb{E}\left[\left\|\sum_{t=0}^{T} \left(\nabla F(v_t) - \mathbb{E}\left[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t\right]\right)\right\|\right] \leq 2\sigma_2 \eta^{\frac{1}{2}} O(T^{\frac{1}{2}}) + 2\sigma_1 \sigma_2 \eta^{\frac{3}{2}}\sqrt{\frac{2}{3}L(T+1)(T+2)}.$$

Proof of Lemma A.4. By the simple calculation, we get

$$\mathbb{E}\left[\left\|\sum_{t=0}^{T} \nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{t=0}^{T} \left(\nabla f(v_t - \eta \epsilon'_{t+1}(v_t)) + \epsilon'_{t+1}(v_t)\right)\right\|\right] + \mathbb{E}\left[\left\|\sum_{t=0}^{T} \epsilon'_{t+1}(v_t)\right\|\right]$$

$$\leq \frac{1}{\eta}\mathbb{E}\left[\|v_0 - v_{T+1}\|\right] + \mathbb{E}\left[\left\|\sum_{t=0}^{T} \epsilon'_{t+1}(v_t)\right\|\right].$$

Each term in the right-hand side can be evaluated as follows:

$$\frac{1}{\eta}\mathbb{E}\left[\|v_0 - v_{T+1}\|\right] \leq \frac{1}{\eta}\sqrt{\mathbb{E}\left[\|v_0 - v_{T+1}\|^2\right]}$$

$$\leq \frac{1}{\eta}\sqrt{2\mathbb{E}\left[\|v_0 - v_*\|^2 + \|v_{T+1} - v_*\|^2\right]}$$

$$\leq \frac{1}{\eta}\sqrt{2\sum_{t=0}^{T+1} \mathbb{E}\left[\|v_t - v_*\|^2\right]}.$$

$$\mathbb{E}\left[\left\|\sum_{t=0}^{T} \epsilon'_{t+1}(v_t)\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|\sum_{t=0}^{T} \epsilon'_{t+1}(v_t)\right\|^2\right]}$$

$$= \sqrt{\mathbb{E}\left[\sum_{t=0}^{T} \|\epsilon'_{t+1}(v_t)\|^2\right]}$$

$$\leq \sigma_1 \sqrt{T},$$

where we used that for $s < t$, $\mathbb{E}[\epsilon'_{s+1}(v_s)^\top \epsilon'_{t+1}(v_t)] = \mathbb{E}[\epsilon'_{s+1}(v_s)^\top \mathbb{E}[\epsilon'_{t+1}(v_t)|\mathcal{F}_t]] = 0$.

Next, we show the second statement by using Lemma A.2 and Proposition A.3 as follows:

$$\mathbb{E}\left[\left\|\sum_{t=0}^{T} \left(\nabla F(v_t) - \mathbb{E}\left[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t\right]\right)\right\|\right]$$

$$\leq \mathbb{E}\left[\sum_{t=0}^{T} \left\|\nabla F(v_t) - \mathbb{E}\left[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t\right]\right\|\right]$$

$$\leq \mathbb{E}\left[\sum_{t=0}^{T} 2\eta \sigma_2 \sqrt{\mathbb{E}\left[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t\right]}\right]$$

$$\leq 2\eta \sigma_2 \sum_{t=0}^{T} \sqrt{\mathbb{E}\left[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2\right]}$$

$$\leq 2\eta \sigma_2 \sqrt{(T+1)\sum_{t=0}^{T} \mathbb{E}\left[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2\right]}$$

$$\leq 2\eta\sigma_2\sqrt{\frac{1}{\eta}O(T) + \frac{2}{3}\eta\sigma_1^2 L(T+1)(T+2)}$$

$$\leq 2\sigma_2\eta^{\frac{1}{2}}O(T^{\frac{1}{2}}) + 2\sigma_1\sigma_2\eta^{\frac{3}{2}}\sqrt{\frac{2}{3}L(T+1)(T+2)}.$$

$\square$

Lemma A.5. *Under Assumptions* **(A1)** *and* **(A2)***, run the stochastic gradient descent with T-iterations with the step size $\eta \leq \frac{1}{2L}$, then the alternative SGD satisfies the following inequality:*

$$\frac{1}{T+1}\mathbb{E}\left[\left\|\sum_{t=0}^{T}\nabla F(v_t)\right\|\right] \leq O(T^{-\frac{1}{2}}) + \frac{1}{\eta(T+1)}\sqrt{2\sum_{t=0}^{T+1}\mathbb{E}\left[\|v_t - v_*\|^2\right]} + \frac{4}{\sqrt{3}}\sigma_1\sigma_2\eta^{\frac{3}{2}}L^{\frac{1}{2}}.$$

Proof of Lemma A.5. We consider the following decomposition:

$$\frac{1}{T+1}\mathbb{E}\left[\left\|\sum_{t=0}^{T}\nabla F(v_t)\right\|\right] \leq \frac{1}{T+1}\mathbb{E}\left[\left\|\sum_{t=0}^{T}\left(\nabla F(v_t) - \mathbb{E}\left[\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))|\mathcal{F}_t\right]\right)\right\|\right]$$

$$+ \frac{1}{T+1}\mathbb{E}\left[\left\|\sum_{t=0}^{T}\left(\mathbb{E}\left[\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))|\mathcal{F}_t\right] - \nabla f(v_t - \eta\epsilon_{t+1}'(v_t))\right)\right\|\right]$$

$$+ \frac{1}{T+1}\mathbb{E}\left[\left\|\sum_{t=0}^{T}\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))\right\|\right].$$

The first and last terms of the right-hand side can be upper bounded by Lemma A.4. The second term can be evaluated as follows:

$$\mathbb{E}\left[\left\|\sum_{t=0}^{T}\left(\mathbb{E}\left[\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))|\mathcal{F}_t\right] - \nabla f(v_t - \eta\epsilon_{t+1}'(v_t))\right)\right\|\right]$$

$$\leq \sqrt{\mathbb{E}\left[\left\|\sum_{t=0}^{T}\left(\mathbb{E}\left[\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))|\mathcal{F}_t\right] - \nabla f(v_t - \eta\epsilon_{t+1}'(v_t))\right)\right\|^2\right]}$$

$$= \sqrt{\mathbb{E}\left[\sum_{t=0}^{T}\left\|\left(\mathbb{E}\left[\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))|\mathcal{F}_t\right] - \nabla f(v_t - \eta\epsilon_{t+1}'(v_t))\right)\right\|^2\right]}$$

$$\leq \sqrt{\mathbb{E}\left[4\eta^2 L^2\sum_{t=0}^{T}\mathbb{E}\left[\left\|\epsilon_{t+1}'(v_t))\right\|^2|\mathcal{F}_t\right]\right]}$$

$$\leq 2\eta L\sigma_1\sqrt{T},$$

where we used that for $X_t := \mathbb{E}\left[\nabla f(v_t - \eta\epsilon_{t+1}'(v_t))|\mathcal{F}_t\right] - \nabla f(v_t - \eta\epsilon_{t+1}'(v_t))$ and $s < t$, $\mathbb{E}[X_s X_t] = \mathbb{E}[X_s\mathbb{E}[X_t|\mathcal{F}_t]] = 0$. Therefore, we eventually obtain the statement of the proposition. $\square$

We here prove Theorem 3.2 which is restated below.

Theorem A.6. *Under Assumptions* **(A1)**–**(A4)***, run the averaged SGD for T-iterations with the step size $\eta \leq \frac{1}{2L}$, then the average $\bar{v}_T$ satisfies the following inequality:*

$$\mathbb{E}[\|\bar{v}_T - v_*\|] \leq O\left(T^{-\frac{1}{2}}\right) + \frac{4\sigma_1\sigma_2\eta^{\frac{3}{2}}L^{\frac{1}{2}}}{\sqrt{3}\mu} + \frac{D_T}{\eta\mu}\sqrt{\frac{2}{T+1}} + \frac{\gamma + MD_T^2}{\mu},$$

where we set $D_T := \sqrt{\frac{1}{T+1}\mathbb{E}\left[\sum_{t=0}^{T}\|v_t - v_*\|^2\right]}$.

Proof. We define $R(v) = \nabla F(v) - \nabla^2 F(v_*)(v - v_*)$. Then, by **(A4)**, we see $\|R(v)\| \le \gamma + M\|v - v_*\|^2$. By taking average of $R(v_t)$ over $t \in \{0, 1, \ldots, T\}$ and rearranging terms, we get

$$\nabla^2 F(v_*)(\overline{v}_T - v_*) = \frac{1}{T+1} \sum_{t=0}^{T} \nabla F(v_t) - \frac{1}{T+1} \sum_{t=0}^{T} R(v_t).$$

Using the positivity of $\nabla^2 F(v_*)$ and taking the expectation, we get

$$\mu \mathbb{E}[\|\overline{v}_T - v_*\|] \le \mathbb{E}[\|\nabla^2 F(v_*)(\overline{v}_T - v_*)\|]$$

$$\le \frac{1}{T+1} \mathbb{E}\left[\left\|\sum_{t=0}^{T} \nabla F(v_t)\right\|\right] + \frac{1}{T+1} \mathbb{E}\left[\left\|\sum_{t=0}^{T} R(v_t)\right\|\right]$$

$$\le \frac{1}{T+1} \mathbb{E}\left[\left\|\sum_{t=0}^{T} \nabla F(v_t)\right\|\right] + \gamma + \frac{M}{T+1} \mathbb{E}\left[\sum_{t=0}^{T} \|v_t - v_*\|^2\right].$$

The first term of the right-hand side can be bounded by Lemma A.5. Thus, we finally get

$$\mu \mathbb{E}[\|\overline{v}_T - v_*\|] \le O\left(T^{-\frac{1}{2}}\right) + \frac{4\sigma_1 \sigma_2 \eta^{\frac{3}{2}} L^{\frac{1}{2}}}{\sqrt{3}} + \frac{1}{\eta(T+1)} \sqrt{2 \sum_{t=0}^{T+1} \mathbb{E}\left[\|v_t - v_*\|^2\right]} + \gamma + \frac{M}{T+1} \mathbb{E}\left[\sum_{t=0}^{T} \|v_t - v_*\|^2\right].$$

Moreover, Jensen's inequality yields $\mathbb{E}[\|\overline{v}_T - v_*\|] \le \sqrt{\mathbb{E}[\|\overline{v}_T - v_*\|^2]} \le D_T$. This concludes the proof. $\qquad\square$

A.3 Proof of Propositions 3.4 and 3.5

We prove Proposition 3.4, which is restated below.

Proposition A.7. *Under Assumptions* **(A1)**, **(A2)**, *and* **(A5)**, *run SGD for T-iterations with the step size* $\eta \le \frac{1}{2L}$, *then we get*

$$D_T^2 \le O\left(T^{-1}\right) + \frac{2\eta \sigma_1^2}{c} + \frac{8\eta^2 \sigma_1^2 L}{3c}\left(1 + \frac{2\eta \sigma_2^2}{c}\right).$$

Proof. To evaluate $\|v_{t+1} - v_*\|^2$ for the alternative SGD (2), we first give several bounds as follows. By Assumption **(A3)**, Young's inequality, and Lemma A.2, we get

$$-2(v_t - v_*)^\top \mathbb{E}[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t]$$

$$= -2(v_t - v_*)^\top \nabla F(v_t) + 2(v_t - v_*)^\top (\nabla F(v_t) - \mathbb{E}[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t])$$

$$\le -2c\|v_t - v_*\|^2 + c\|v_t - v_*\|^2 + \frac{1}{c}\|\nabla F(v_t) - \mathbb{E}[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t]\|^2$$

$$\le -c\|v_t - v_*\|^2 + \frac{4\eta^2 \sigma_2^2}{c} \mathbb{E}\left[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t\right].$$

By Assumption **(A2)** and Young's inequality again, we get

$$\mathbb{E}[\|\epsilon'_{t+1}(v_t) + \nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t]$$

$$\le 2\mathbb{E}[\|\epsilon'_{t+1}(v_t)\|^2|\mathcal{F}_t] + 2\mathbb{E}[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t]$$

$$\le 2\sigma_1^2 + 2\mathbb{E}[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t].$$

Combining the above two inequalities, we get

$$\mathbb{E}[\|v_{t+1} - v_*\|^2|\mathcal{F}_t] = \mathbb{E}[\|v_t - \eta \epsilon'_{t+1}(v_t) - \eta \nabla f(v_t - \eta \epsilon'_{t+1}(v_t)) - v_*\|^2|\mathcal{F}_t]$$

$$= \|v_t - v_*\|^2 - 2\eta(v_t - v_*)^\top \mathbb{E}[\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))|\mathcal{F}_t]$$

$$+ \eta^2 \mathbb{E}[\|\epsilon'_{t+1}(v_t) + \nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t]$$

$$\le (1 - c\eta)\|v_t - v_*\|^2 + 2\eta^2 \sigma_1^2$$

$$+ 2\eta^2 \left(1 + \frac{2\eta \sigma_2^2}{c}\right) \mathbb{E}\left[\|\nabla f(v_t - \eta \epsilon'_{t+1}(v_t))\|^2|\mathcal{F}_t\right].$$

Taking the expectation regarding all histories and summing up over $t = 0, 1, \ldots, T$, we get

$$c\eta \sum_{t=0}^{T} \mathbb{E}[\|v_t - v_*\|^2] \leq \mathbb{E}[\|v_0 - v_*\|^2] - \mathbb{E}[\|v_{T+1} - v_*\|^2] + 2\eta^2 \sigma_1^2 (T+1)$$

$$+ 2\eta^2 \left(1 + \frac{2\eta\sigma_2^2}{c}\right) \sum_{t=0}^{T} \mathbb{E}\left[\|\nabla f(v_t - \eta \epsilon_{t+1}'(v_t))\|^2\right]$$

$$\leq \mathbb{E}[\|v_0 - v_*\|^2] - \mathbb{E}[\|v_{T+1} - v_*\|^2] + 2\eta^2 \sigma_1^2 (T+1)$$

$$+ \frac{8}{3}\eta \left(1 + \frac{2\eta\sigma_2^2}{c}\right) \mathbb{E}[f(w_0)] + \frac{4}{3}\eta^3 \left(1 + \frac{2\eta\sigma_2^2}{c}\right) \sigma_1^2 L (T+2),$$

where we used Proposition A.3. Therefore, we conclude

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|v_t - v_*\|^2] \leq \frac{1}{c\eta(T+1)} \mathbb{E}[\|v_0 - v_*\|^2] + \frac{8}{3c(T+1)} \left(1 + \frac{2\eta\sigma_2^2}{c}\right) \mathbb{E}[f(w_0)]$$

$$+ \frac{2\eta\sigma_1^2}{c} + \frac{8\eta^2}{3c} \left(1 + \frac{2\eta\sigma_2^2}{c}\right) \sigma_1^2 L$$

$$= O\left(T^{-1}\right) + \frac{2\eta\sigma_1^2}{c} + \frac{8\eta^2}{3c} \left(1 + \frac{2\eta\sigma_2^2}{c}\right) \sigma_1^2 L.$$

$\square$

We prove Proposition 3.5, which is restated below.

Proposition A.8. *Suppose that Assumptions **(A1)** and **(A2)** hold, and that there exists $\sigma_3 > 0$ such that for any $w \in \mathbb{R}^d$, $\mathbb{E}[\|\epsilon_{t+1}(w)\|^2] \geq \sigma_3^2$. Running SGD with $\eta \leq \frac{3\sigma_3^2}{32\sigma_1^2 L}$, we get*

$$\frac{\eta^2 \sigma_3^2}{8} \leq D_T^2 + O(T^{-1}).$$

Proof. Let us evaluate the term $\mathbb{E}[\|w_t - v_*\|^2]$ as follows. On one hand, by using the update rule of SGD, for $t \geq 1$,

$$\mathbb{E}\left[\|w_t - v_*\|^2\right] = \mathbb{E}\left[\|w_{t-1} - \eta \nabla f(w_{t-1}) - \eta \epsilon_t(w_{t-1}) - v_*\|^2\right]$$

$$= \mathbb{E}\left[\|w_{t-1} - \eta \nabla f(w_{t-1}) - v_*\|^2 - 2\eta \epsilon_t(w_{t-1})^\top (w_{t-1} - \eta \nabla f(w_{t-1}) - v_*) + \eta^2 \|\epsilon_t(w_{t-1})\|^2\right]$$

$$= \mathbb{E}\left[\|w_{t-1} - \eta \nabla f(w_{t-1}) - v_*\|^2 + \eta^2 \|\epsilon_t(w_{t-1})\|^2\right]$$

$$\geq \eta^2 \sigma_3^2,$$

where we used $\mathbb{E}[\epsilon_t(w_{t-1})^\top (w_{t-1} - \eta \nabla f(w_{t-1}) - v_*)] = \mathbb{E}[\mathbb{E}[\epsilon_t(w_{t-1})|\mathcal{F}_{t-1}]^\top (w_{t-1} - \eta \nabla f(w_{t-1}) - v_*)] = 0$. On the other hand, by using $v_t = w_t - \eta \nabla f(w_t)$,

$$\mathbb{E}\left[\|w_t - v_*\|^2\right] = \mathbb{E}\left[\|v_t + \eta \nabla f(w_t) - v_*\|^2\right]$$

$$\leq 2\mathbb{E}\left[\|v_t - v_*\|^2 + \eta^2 \|\nabla f(w_t)\|^2\right].$$

Hence, by taking the sum over $t \in \{0, 1, \ldots, T\}$, we have

$$\frac{1}{2}\|w_0 - v_*\|^2 + \frac{1}{2}\eta^2 \sigma_3^2 T \leq \sum_{t=0}^{T} \mathbb{E}\left[\|v_t - v_*\|^2\right] + \eta^2 \sum_{t=0}^{T} \mathbb{E}\left[\|\nabla f(w_t)\|^2\right]$$

$$\leq \sum_{t=0}^{T} \mathbb{E}\left[\|v_t - v_*\|^2\right] + \eta^2 \sum_{t=0}^{T} \mathbb{E}\left[\|\nabla f(w_{t+1})\|^2\right] + \eta^2 \|\nabla f(w_0)\|^2.$$

Since $w_{t+1} = v_t - \eta \epsilon_{t+1}(w_t)$, we get by apply Proposition A.3,

$$\frac{1}{2}\|w_0 - v_*\|^2 + \frac{1}{2}\eta^2 \sigma_3^2 T \leq \sum_{t=0}^{T} \mathbb{E}\left[\|v_t - v_*\|^2\right] + \frac{4}{3}\eta \mathbb{E}[f(w_0)] + \frac{2}{3}\eta^3 \sigma_1^2 L (T+2) + \eta^2 \|\nabla f(w_0)\|^2.$$
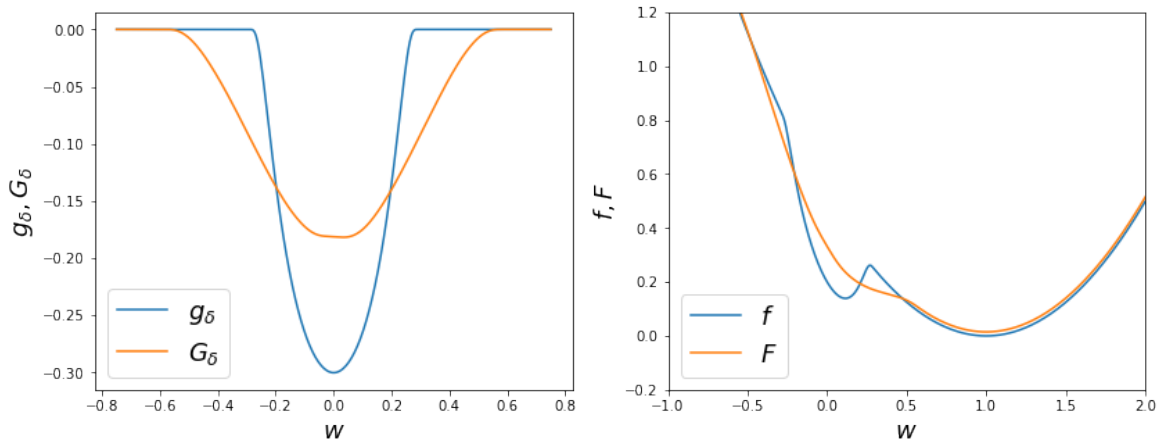
Figure 7: The left figure plots the mollifier g_δ (blue) and smoothed mollifier G_δ (orange), and the right figure plots the objective f (blue) and smoothed objective F (orange). Hyperparameters are set to $\delta = 0.3$, $r = 1$, and $\eta = 0.3$.

Therefore, we have

$$\frac{1}{4}\eta^2\sigma_3^2 - \frac{4}{3}\eta^3\sigma_1^2 L \leq \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|v_t - v_*\|^2\right] + \frac{1}{T+1}\left(\frac{4}{3}\eta\mathbb{E}[f(w_0)] + \eta^2\|\nabla f(w_0)\|^2 - \frac{1}{2}\|w_0 - v_*\|^2\right)$$

$$= \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|v_t - v_*\|^2\right] + O(T^{-1}).$$

Since $\eta \leq \frac{3\sigma_3^2}{32\sigma_1^2 L}$, we have $\frac{1}{8}\eta^2\sigma_3^2 \leq \frac{1}{4}\eta^2\sigma_3^2 - \frac{4}{3}\eta^3\sigma_1^2 L$. Therefore, the above inequality concludes the proof. $\qquad\square$

B   Motivating Example 1

We present an example that demonstrates the implicit bias of SGD and averaged SGD. Moreover, we theoretically and empirically verify that averaged SGD can get closer to a flat minimum than SGD.

B.1   Problem Setup

In this section, we present a motivating example that verifies the convergence to a flat minimum and a certain separation between SGD and averaged SGD. We consider a one-dimensional objective function $f : \mathbb{R} \to \mathbb{R}$ defined below: for $\delta > 0$,

$$f(w) = \frac{1}{2}(w - 1)^2 + g_\delta(w), \tag{9}$$

where $g_\delta : \mathbb{R} \to \mathbb{R}$ is a scaled mollifier:

$$g_\delta(w) = \begin{cases} -\delta\exp\left(1 - \frac{1}{1-\left(\frac{w}{\delta}\right)^2}\right) & (|w| < \delta), \\ 0 & (|w| \geq \delta). \end{cases}$$

$g_\delta(w) = \delta g_1(w/\delta)$ is a scaling of the well-known mollifier of g_1 which is an infinitely differentiable function with a compact support. That is, g_δ is a smooth function whose support is $[-\delta, \delta]$. Note that the function $f(w)$ has a local minimum in $[-\delta, \delta]$, which becomes sharp when δ is small. See Figure 7 which depicts graphs of g_δ, G_δ, f, and F.

The maximum values of the first and second derivatives of g_1 are bounded. Thus, we define constants C_1, C_2 by

$$C_1 = \max\left\{1, \max_w |g_1'(w)|\right\}, \quad C_2 = \max_w |g_1''(w)|.$$

Since $g_\delta''(w) = \frac{1}{\delta} g_1''(w/\delta)$, we see the second derivative of g_δ is bounded by $C_2 \delta^{-1}$. Hence, Lipschitz smoothness (boundedness of Hessian) L of f is $1 + C_2 \delta^{-1}$.

Next, we consider the uniform noise on the interval $[-r, r]$ for $r > 0$, i.e., $\epsilon \sim U[-r, r]$ and suppose $\epsilon(w, z) = \epsilon(z)(= \epsilon'(v, z))$ where $\Omega \ni z \mapsto \epsilon(w, z)$ is an explicit representation of the random noise. In other words, noise distribution does not change in w. In this case, we see $\sigma_1^2 = \sigma_3^2 = \mathbb{E}[\epsilon^2] \leq \frac{r^2}{3}$ and $\sigma_2 = 0$. The smoothed objective F with the noise ϵ' and step-size η is

$$
\begin{aligned}
F(v) &= \mathbb{E}[f(v - \eta \epsilon')] \\
&= \frac{1}{2}(v-1)^2 + \eta^2 \sigma_1^2 + \mathbb{E}[g_\delta(v - \eta \epsilon')] \\
&\sim \frac{1}{2}(v-1)^2 + \mathbb{E}[g_\delta(v - \eta \epsilon')].
\end{aligned}
$$

We consider the following problem setup:

$$
\delta \leq \frac{1}{4(1 + 2C_1)}, \tag{10}
$$

$$
r \geq \frac{64}{3} C_1 (\delta + C_2). \tag{11}
$$

Note that we can choose arbitrarily small $\delta > 0$ and large r which satisfy the above inequalities. Under these conditions, we can choose the following step size η for appropriate smoothing.

$$
\frac{2C_1 \delta}{r} \leq \eta \leq \min \left\{ \frac{1}{4r} - \frac{\delta}{r}, \frac{3\delta}{32(\delta + C_2)} \right\}. \tag{12}
$$

We note that the above step size satisfies $\eta \leq \min \left\{ \frac{1}{2L}, \frac{3\sigma_3^2}{32\sigma_1^2 L} \right\}$ required in Propositions 3.4 and 3.5 since $\frac{3\delta}{32(\delta + C_2)} = \frac{3}{32L} = \frac{3\sigma_3^2}{32\sigma_1^2 L} < \frac{1}{2L}$.

B.2   Evaluation of SGD and Averaged SGD

Under the above setup (10)–(12), we can estimate constants appearing in the convergence results of SGD and averaged SGD as follows (for the detail see the next subsection):

$$
L = 1 + \frac{C_2}{\delta}, \ \sigma_1^2 = \sigma_3^2 = \frac{r^2}{3}, \ \sigma_2 = 0, \mu = 1, \ c = \frac{1}{3}, \ \gamma = 0, \ M = \frac{8}{9}. \tag{13}
$$

Moreover, the minimum of the smoothed objective is $v_* = 1$, and a sharp minimum (~ 0) can be eliminated by smoothing.

Therefore, for SGD we obtain by Proposition 3.4,

$$
\begin{aligned}
D_T^2 &= \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|v_t - v_*\|^2] \\
&\leq O\left(T^{-1}\right) + \frac{2\eta \sigma_1^2}{c} + \frac{8\eta^2 \sigma_1^2 L}{3c}\left(1 + \frac{2\eta \sigma_2^2}{c}\right) \\
&\leq O\left(T^{-1}\right) + 6\eta r^2 + 8\eta^2 r^2 \left(1 + \frac{C_2}{\delta}\right).
\end{aligned}
$$

We see from this inequality, $\eta_* = \frac{2C_1 \delta}{r}$ is the best choice of the step-size, resulting in

$$
D_T \leq O\left(T^{-1/2}\right) + \sqrt{12 C_1 \delta r + 32 C_1^2 \delta (\delta + C_2)},
$$

and

$$
\frac{C_1 \delta}{\sqrt{6}} \leq D_\infty \leq 2\sqrt{(3r + 8C_1 C_2) C_1 \delta} + 4\sqrt{2} C_1 \delta,
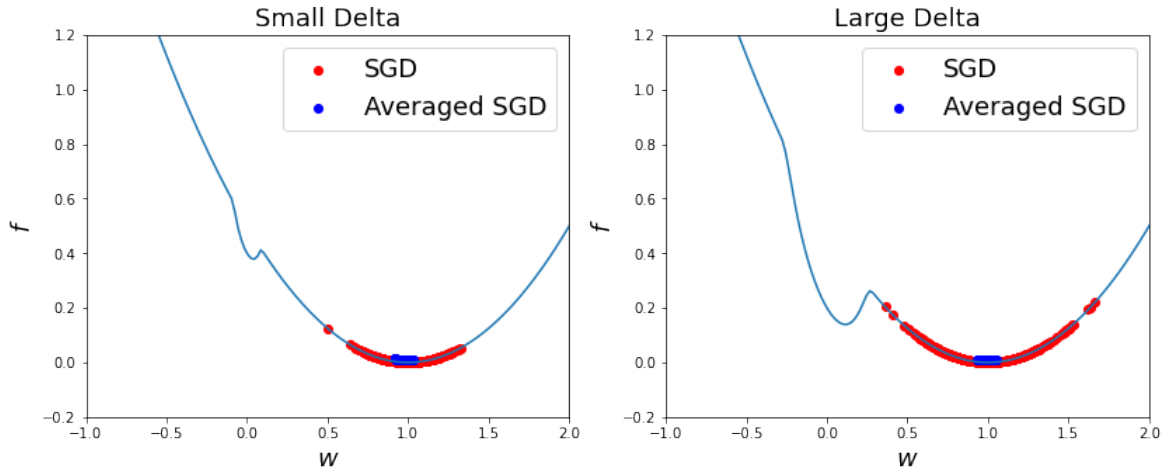$$

Figure 8: The figures plot the convergent points of SGD and averaged SGD for the problem (9). For the left case, we set $\delta = 0.1$, $r = 1$, $\eta = \delta/r = 0.1$ and for the right case, we set $\delta = 0.3$, $r = 1$, $\eta = \delta/r = 0.3$.

where we applied Proposition 3.5 for the lower bound. This result means SGD avoids a sharp minimum (i.e., $v \sim 0$ under small $\delta > 0$) and converges to a flat minimum $v_* = 1$, and a too large noise will affect the convergence to v_*.

Hence, Eq. (5): $0 \ll D_\infty \ll 9/8$ is satisfied under the above setting with mildly small δ, meaning that the error of averaged SGD is strictly smaller than that of SGD since $\lim_{T \to \infty} \mathbb{E}[\|\bar{v}_T - v_*\|] \leq \frac{8}{9}D_\infty^2$ by Eq. (4). We empirically observed this phenomenon in Figure 8 in which we run SGD and averaged SGD for problems with small $\delta = 0.1$ and relatively large $\delta = 0.3$.

B.3   Estimation of Constants

We verify the estimations of constants in (13). L, σ_1^2, and σ_2 are already obtained, thus, μ, c, and M remain.

Minimum and estimation of μ.   We first see that under our problem setting, the local minimum around the origin is eliminated and 1 is the optimal solution of F, i.e., $v_* = 1$.

The smoothed function $G_\delta(v) \overset{\text{def}}{=} \mathbb{E}[g_\delta(v - \eta\epsilon')]$ and its derivative $G_\delta'(v)$ are calculated as follows:

$$G_\delta(v) = \int_{-r}^{r} g_\delta(v - \eta t)\frac{1}{2r}\mathrm{d}t,$$

$$G_\delta'(v) = \int_{-r}^{r} g_\delta'(v - \eta t)\frac{1}{2r}\mathrm{d}t.$$

By taking into account $\mathrm{supp}(g_\delta) = [-\delta, \delta]$, the smoothed objective $G_\delta(v)$ is constant on $\{|v| \leq \eta r - \delta\} \cup \{|v| \geq \eta r + \delta\}$, and thus, G_δ' is non-zero only on $\mathrm{supp}(G_\delta') = [-\eta r - \delta, -\eta r + \delta] \cup [\eta r - \delta, \eta r + \delta]$. For graphs of g_δ and G_δ, see Figure 7 (left). Since $\eta r + \delta < 1/4$ under (12), G_δ is a constant around $v = 1$, that is, $v = 1$ is still a local minimum of F.

We evaluate the bound on G_δ' on $\mathrm{supp}(G_\delta')$ below. For $v \in [\eta r - \delta, \eta r + \delta]$ the support of $g_\delta'(v - \eta t)$ in $t \in \mathbb{R}$ is $[(v - \delta)/\eta, (v + \delta)/\eta]$, we get

$$0 \leq G_\delta'(v) = \int_{-r}^{r} g_\delta'(v - \eta t)\frac{1}{2r}\mathrm{d}t$$

$$\leq \int_{-\infty}^{\infty} |g_\delta'(v - \eta t)|\frac{1}{2r}\mathrm{d}t$$

$$= \int_{\frac{v-\delta}{\eta}}^{\frac{v+\delta}{\eta}} |g_\delta'(v - \eta t)|\frac{1}{2r}\mathrm{d}t$$

$$\leq C_1 \int_{\frac{v-\delta}{\eta}}^{\frac{v+\delta}{\eta}} \frac{1}{2r} \mathrm{d}t = \frac{C_1 \delta}{\eta r},$$

where we used $|g_\delta'(v)| = |g_1'(v/\delta)| \leq C_1$. A bound on $[-\eta r - \delta, -\eta r + \delta]$ is also obtained in the same way. Thus, we see

$$\begin{cases} -\frac{C_1 \delta}{\eta r} \leq & G_\delta'(v) \leq 0 & (v \in [-\eta r - \delta, -\eta r + \delta]), \\ 0 \leq & G_\delta'(v) \leq \frac{C_1 \delta}{\eta r} & (v \in [\eta r - \delta, \eta r + \delta]), \\ & G_\delta'(v) = 0 & (\text{else}). \end{cases}$$

If there are additional stationary points of F, they should exist in $[\eta r - \delta, \eta r + \delta] = \mathrm{supp}(G_\delta') \setminus [-\eta r - \delta, -\eta r + \delta]$ because of the sign of G_δ' and $\mathrm{supp}(G_\delta') \subset (-\infty, 1/4)$. However, since $\eta r + \delta \leq 1/4$ and $\frac{C_1 \delta}{\eta r} \leq 1/2$ under (12), we see

$$\max_{v \in [\eta r - \delta, \eta r + \delta]} F'(v) \leq (\eta r + \delta) - 1 + \frac{C_1 \delta}{\eta r} \leq \frac{1}{4} - 1 + \frac{1}{2} = -\frac{1}{4}.$$

Hence, $v_* = 1$ is the unique local minimum (i.e., optimal solution) of F and we can conclude $\mu = 1$.

Estimation of c. From the above argument, we get

$$F'(v)(v-1) = (v-1)^2 + G_\delta'(v)(v-1)$$

$$\geq \begin{cases} (v-1)^2 & (v \in [-\eta r - \delta, -\eta r + \delta]), \\ (v-1)^2 + \frac{C_1 \delta}{\eta r}(v-1) \geq (v-1)^2 + \frac{1}{2}(v-1) & (v \in [\eta r - \delta, \eta r + \delta]), \\ (v-1)^2 & (\text{else}). \end{cases}$$

Clearly, $1/2 \leq 2(1-v)/3$ for $v \leq \eta r + \delta \leq 1/4$. Thus, $F'(v)(v-1) \geq (v-1)^2/3$ on $v \in [\eta r - \delta, \eta r + \delta]$ and we conclude $c = 1/3$.

Estimation of γ and M. Noting $v_* = 1$ and $F''(1) = 1$, we have

$$|F'(v) - F''(v_*)(v - v_*)| = |(v-1) + G_\delta'(v) - (v-1)| = |G_\delta'(v)|.$$

Because of the problem setup, it is enough to verify $\gamma = 0$, $M = \frac{8}{9}$ satisfy $|G_\delta'(v)| \leq M|v-1|^2$ on $v \in [\eta r - \delta, \eta r + \delta]$. Since $|G_\delta'(v)| \leq 1/2$ and $v \leq 1/4$ for v in this interval, we have

$$|G_\delta'(v)| \leq \frac{1}{2} \leq M\frac{9}{16} \leq M(v-1)^2.$$

This concludes $\gamma = 0$, $M = \frac{8}{9}$.

C Motivating Example 2

We provide an example for which SGD and averaged SGD behave in significantly different ways. For this example, the convergence of SGD cannot be guaranteed whereas averaged SGD can converge.

C.1 Problem Setup

We consider a one-dimensional objective function $f : \mathbb{R} \to \mathbb{R}$ defined below: for $\delta > 0$,

$$f(w) = \frac{1}{2}w^2 + g_\delta(w), \tag{14}$$

where $g_\delta : \mathbb{R} \to \mathbb{R}$ is a scaled mollifier:

$$g_\delta(w) = \begin{cases} \delta^2 \exp\left(1 - \frac{1}{1 - \left(\frac{w}{\delta}\right)^2}\right) & (|w| < \delta), \\ 0 & (|w| \geq \delta). \end{cases}$$
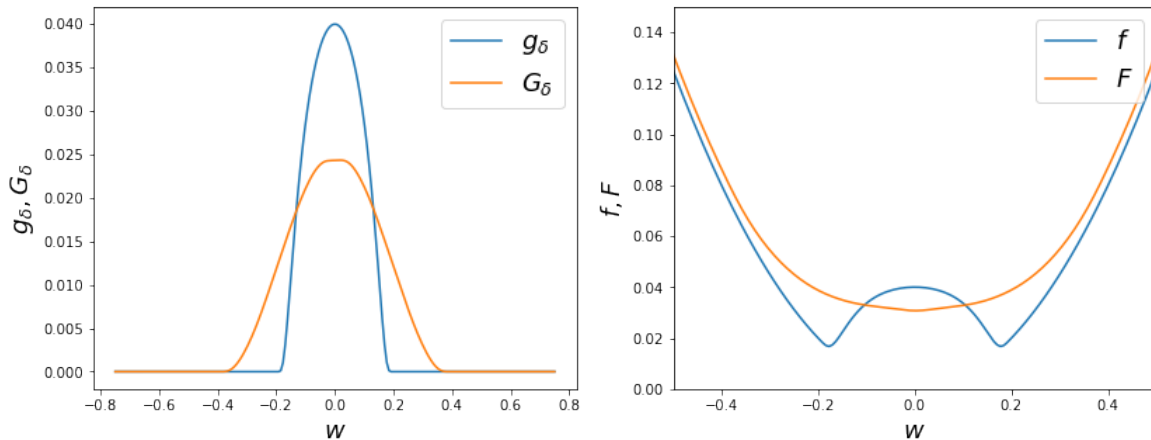
Figure 9: The left figure plots the mollifier g_δ (blue) and smoothed mollifier G_δ (orange), and the right figure plots the objective f (blue) and smoothed objective F (orange). Hyperparameters are set to $\delta = 0.2$, $r = 1$, and $\eta = 0.2$.

The function f is symmetry and has two local minima. Taking the derivative at $v = \alpha\delta$ ($\alpha > 0$), we see

$$f'(\alpha\delta) = \alpha\delta \left(1 - \frac{2}{(1-\alpha^2)^2} \exp\left(1 - \frac{1}{1-\alpha^2} \right) \right).$$

Therefore, the value of $\alpha > 0$ such that $f'(\alpha\delta) = 0$ is independent of δ, meaning optimal solutions of f are $O(\delta)$.

Let $\epsilon \sim U[-r, r]$ be the uniform distribution as in Section B. Then, the smoothed objective F with the noise ϵ' and step-size η is

$$F(v) = \frac{1}{2}v^2 + \eta^2\sigma_1^2 + \mathbb{E}[g_\delta(v - \eta\epsilon')]$$

$$\sim \frac{1}{2}v^2 + \mathbb{E}[g_\delta(v - \eta\epsilon')].$$

Figure 9 depicts the above functions. We note that g_δ is slightly different from that in the previous section.

We define constants C_1, C_2 as follows:

$$C_1 = \max_w |g_1'(w)|, \ \ C_2 = \max\{1, \max_w |g_1''(w)|\}.$$

For sufficient smoothing, we consider the following problem setup:

$$C_1\delta^2 \le \frac{\eta r}{2}(\eta r - \delta), \tag{15}$$

$$C_2\delta \le \frac{\eta r}{2}. \tag{16}$$

Note that for a given $\eta r > 0$, the above conditions can be satisfied if $\delta > 0$ is sufficiently small.

Lipschitz smoothness L of f is $L = 1 + C_2$, and hence η should satisfy $\eta \le \min\left\{ \frac{3\sigma_3^2}{32\sigma_1^2 L}, \frac{1}{2L} \right\} = \frac{3}{32L} = \frac{3}{32(1+C_2)}$. We note contrary to the case in the previous section, L does not suffer from δ. Hence we do not need to make η small depending on δ.

C.2 Evaluation of SGD and Averaged SGD

Under conditions (15) and (16), we can estimates constants appearing in the convergence results of SGD and averaged SGD as follows (for the detail see the next subsection):

$$L = 1 + C_2, \ \sigma_1^2 = \sigma_3^2 = \frac{r^2}{3}, \ \sigma_2 = 0, \mu = 1, \ c = \frac{1}{2}, \ \gamma = \frac{C_1 \delta^2}{\eta r}, \ M = 0. \tag{17}$$

Moreover, the smoothed objective F has a unique solution $v_* = 0$ as shown in the next subsection.

Therefore, for SGD we obtain by Proposition 3.4,

$$
\begin{aligned}
D_T^2 &= \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|v_t - v_*\|^2] \\
&\leq O\left(T^{-1}\right) + \frac{2\eta \sigma_1^2}{c} + \frac{8\eta^2 \sigma_1^2 L}{3c}\left(1 + \frac{2\eta \sigma_2^2}{c}\right) \\
&\leq O\left(T^{-1}\right) + 4\eta r^2 + \frac{16}{3}\eta^2 r^2 (1 + C_2).
\end{aligned}
$$

That is,

$$D_T \leq O\left(T^{-1/2}\right) + 2\sqrt{\eta r^2 + \frac{4}{3}\eta^2 r^2 (1 + C_2)}, \tag{18}$$

and

$$\frac{\eta r}{2\sqrt{6}} \leq D_\infty \leq 2\sqrt{\eta} r + 4\eta r \sqrt{\frac{1 + C_2}{3}},$$

where we applied Proposition 3.5 for the lower bound. Taking into account (16), we see the right hand side of this bound is $\Omega(\delta)$. Then, since the optimal solution of f is also $O(\delta)$, we can see that upper-bound (18) on SGD cannot distinguish solutions of the original function f and the smoothed objective F.

On the other hand, the condition Eq. (5): $C_1 \delta^2 / (\eta r) \ll D_\infty \ll \infty$ is satisfied under the setting $\delta^2 \ll \eta^2 r^2$. Then, we have $\lim_{T \to \infty} \mathbb{E}[\|\bar{v}_T - v_*\|] \leq C_1 \delta^2 / (\eta r)$ by by Eq. (4). Since, δ can be small independent of η and r, averaged SGD can approach a solution $v_* = 0$ of F up to $O(\delta^2)$.

This difference between SGD and averaged SGD is intuitively understandable. In the case of using a large step size compared to δ (i.e., $\delta^2 \ll \eta^2 r^2$), SGD does not converge to a point and hence oscillates in the valley. Then, its mean should be located almost at the origin (i.e., v_*), meaning the convergence of averaged SGD. In the case of using a small step size, both SGD and averaged SGD do not converge to v_* but converge to a minimizer of f. Indeed, we empirically observed this phenomenon in Figure 10. Both SGD and averaged SGD using small step size $\eta = 0.01$ compared to δ converge to the solutions of f because of the weakened bias and precise optimization. On the other hand, when using a relatively large step size $\eta = 0.2$, SGD cannot converge, whereas averaged SGD converges to the solution $v_* = 0$ of F very accurately.

C.3 Estimation of Constants

Minimum and estimation of μ. In a similar way to Section B, the smoothed function $G_\delta(v) = \mathbb{E}[g_\delta(v - \eta t)]$ and its derivative G_δ' satisfy the following. By taking into account $\mathrm{supp}(g_\delta) = [-\delta, \delta]$, the smoothed objective $G_\delta(v)$ is constant on $\{|v| \leq \eta r - \delta\} \cup \{|v| \geq \eta r + \delta\}$, and thus, G_δ' is non-zero only on $\mathrm{supp}(G_\delta') = [-\eta r - \delta, -\eta r + \delta] \cup [\eta r - \delta, \eta r + \delta]$. Specifically, we obtain

$$
\begin{cases}
0 \leq & G_\delta'(v) \leq \frac{C_1 \delta^2}{\eta r} & (v \in [-\eta r - \delta, -\eta r + \delta]), \\
-\frac{C_1 \delta^2}{\eta r} \leq & G_\delta'(v) \leq 0 & (v \in [\eta r - \delta, \eta r + \delta]), \\
& G_\delta'(v) = 0 & (\text{else}).
\end{cases}
$$

It is verified that the derivative of F vanishes only at $v = 0$ as follows. If there is a point so that $F'(v) = 0$ other than $v \neq 0$, v should be in $\mathrm{supp}(G_\delta)$. For $v \in [\eta r - \delta, \eta r + \delta]$, we get

$$F'(v) = v + G_\delta'(v) \geq \eta r - \delta - \frac{C_1 \delta^2}{\eta r} > 0,$$
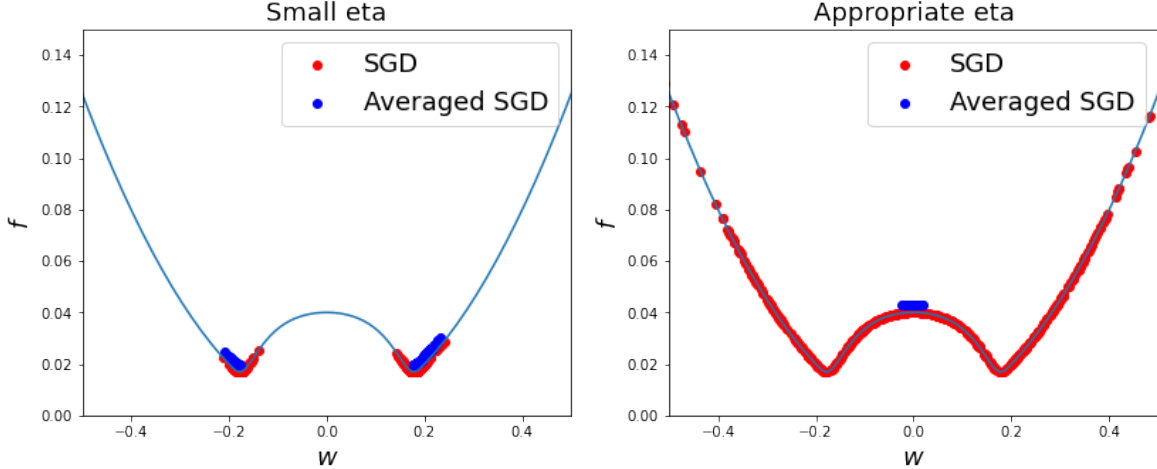
Figure 10: The figures plot the convergent points of SGD and averaged SGD using small step size $\eta = 0.01$ (left) and appropriately large $\eta = 0.2$ (right). For both cases, we set $\delta = 0.2$ and $r = 1$.

where we used (15). In a similar way, we have $F'(v) < 0$ on $v \in [-\eta r - \delta, -\eta r + \delta]$. Moreover, $F'(0) = 0$ clearly holds. Thus, $v_* = 0$ is a unique solution.

Since G_δ is constant on $[-\eta r + \delta, \eta r - \delta]$ under the condition (16), we see $G''_\delta(0) = 0$. Therefore, we get $F''(0) = v$, and hence $\mu = 1$.

Estimation of c. From the above argument, we get

$$F'(v)v = v^2 + G'_\delta(v)v$$

$$\geq \begin{cases} v^2 + \frac{C_1\delta^2}{\eta r}v & (v \in [-\eta r - \delta, -\eta r + \delta]), \\ v^2 - \frac{C_1\delta^2}{\eta r}v & (v \in [\eta r - \delta, \eta r + \delta]), \\ v^2 & (\text{else}). \end{cases}$$

Since $\frac{C_1\delta^2}{\eta r} \leq \frac{\eta r - \delta}{2} \leq -\frac{v}{2}$ on $[-\eta r - \delta, -\eta r + \delta]$ and $\frac{C_1\delta^2}{\eta r} \leq \frac{\eta r - \delta}{2} \leq \frac{v}{2}$ on $[\eta r - \delta, \eta r + \delta]$, we have $F'(v)v \geq \frac{v^2}{2}$. Thus, $c = \frac{1}{2}$.

Estimation of γ and M. Since $G''_\delta(0) = 0$, we have $F'(v) - F''(0)v = G'_\delta(v)$ for any $v \in \mathbb{R}$, leading to $|F'(v) - F''(0)v| = |G'_\delta(v)| \leq \frac{C_1\delta^2}{\eta r}$. Hence, $\gamma = \frac{C_1\delta^2}{\eta r}$ and $M = 0$.

D   Additional Experiments

Table 3: Comparison of test classification accuracies on CIFAR10 dataset. All methods adopt the multi-step strategy for the step size schedule.

		CIFAR10	
	η	ResNet-50	WRN-28-10
SGD	s	95.95 (0.10)	96.85 (0.16)
Averaged SGD	s	96.58 (0.14)	97.24 (0.07)
	m	**96.89 (0.05)**	**97.44 (0.04)**
	l	96.27 (0.16)	97.05 (0.09)

We run SGD and averaged SGD on CIFAR10 dataset with the step size strategy 'l' under the same settings as in Section 4. Table 3 lists the results including this case. We observe that the large step size 'l' does not work
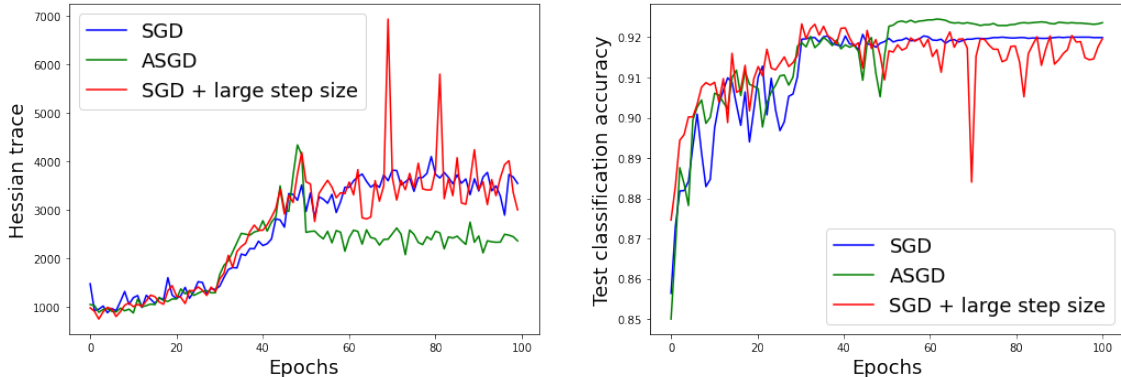
Figure 11: The figure depicts the curve of the trace of Hessian $\nabla^2 f(w)$ and test loss functions achieved by SGD, SGD with large step size, and averaged SGD. Each algorithm is run to train the standard convolutional neural network on Fashion MNIST dataset.

so well on CIFAR10 dataset compared to other schedules. We hypothesize this is because CIFAR10 is not so difficult dataset and does not require stronger bias induced by a larger step size.

We also validate the cosine annealing strategy for the step size, which is frequently used due to its excellent performance. We used the symbols 's', 'm', and 'l' for the cosine annealing depending on the last step sizes which are set to 0, 0.004, and 0.02, respectively. The parameter averaging for averaged SGD is taken over the last quarter of the training. From the table, we observe the usefulness of parameter averaging for the cosine annealing schedule as well.

Table 4: Comparison of test classification accuracies on CIFAR100 and CIFAR10 datasets. All methods adopt cosine annealing for the step-size schedule.

		CIFAR100				CIFAR10		
	η	ResNet-50	WRN-28-10	Pyramid	η	ResNet-50	WRN-28-10	Pyramid
SGD	s	82.26	82.68	82.97	s	96.58	97.00	96.66
Averaged	s	**83.89**	84.28	85.14	s	**97.01**	97.28	97.07
SGD	l	83.21	**84.49**	**85.47**	m	96.86	**97.51**	**97.32**
SAM	s	83.35	84.64	86.24	s	96.40	96.89	97.61
Averaged	s	83.18	84.94	86.79	s	**96.56**	97.14	**97.55**
SAM	l	**83.58**	**85.26**	**86.84**	m	96.51	**97.19**	97.48

Finally, we run SGD, SGD with a large step size, and averaged SGD to train the standard convolutional neural network on Fashion MNIST dataset to confirm how efficiently sharpness and classification accuracy can be optimized by each method. We note the large step size used for SGD is the same as that for averaged SGD. We plot the trace of Hessian $\nabla^2 f(w)$ and test loss functions in Figure 11. From this figure, we observe that the averaged SGD converges to a flatter region and achieves the highest classification accuracy on the test dataset as expected in our theory.

E Related Literature and Discussion

Flat Minimum. Keskar et al. (2017) and Hochreiter and Schmidhuber (1997) showed a flat minimum generalizes well and a sharp minimum generalizes poorly. However, Dinh et al. (2017) pointed out that the flatness solely cannot explain generalization because it can be easily manipulated. Neyshabur et al. (2017) rigorously proved the sharpness combined with ℓ_2-norm provides a generalization bound and Jiang et al. (2020) confirmed this strong correlation through large scale experiments. Keskar et al. (2017) also argued that SGD converges to a flat minimum and He et al. (2019) argued that the averaged SGD tends to converge to an asymmetric valley due to

the stochastic gradient noise. Indeed, several works (Kleinberg et al., 2018; Zhou et al., 2020) justified the implicit bias of SGD towards a flat region or asymmetric valley. Moreover, Izmailov et al. (2018); Foret et al. (2020); Damian et al. (2021); Orvieto et al. (2022); Kaddour et al. (2022) studied the techniques to further bring out the bias of SGD with improved performance. In particular, SAM and SWA achieved a significant improvement in generalization. In our paper, we show that parameter averaging stabilizes the convergence to a flat region or asymmetric valley, and suggest the usefulness of the combination with the large step size for the difficult dataset which needs a stronger regularization. Besides, several authors proposed methods that explicitly inject noise for improving generalization (Chaudhari et al., 2019), in particular for the large batch setting (Wen et al., 2018; Haruki et al., 2019; Lin et al., 2020).

Markov Chain Interpretation of SGD.   Dieuleveut et al. (2020); Yu et al. (2020) provided the Markov chain interpretation of SGD. They showed the marginal distribution of the parameter of SGD converges to an invariant distribution for convex and nonconvex optimization problems, respectively. Moreover, Dieuleveut et al. (2020) showed the mean of the invariant distribution, attained by the averaged SGD, is at distance $O(\eta)$ from the minimizer of the objective function, whereas SGD itself oscillates at distance $O(\sqrt{\eta})$ in the convex optimization settings. Izmailov et al. (2018) also attributed the success of SWA to such a phenomenon. That is, Izmailov et al. (2018) explained that SGD travels on the hypersphere because of the convergence to Gaussian distribution and the concentration on the sphere under a simplified setting, and thus averaging scheme allows us to go inside of the sphere which may be flat. We can say our contribution is to theoretically justify this intuition by extending the result obtained by Dieuleveut et al. (2020) to a nonconvex optimization setting. In the proof, we utilize the alternative view of SGD (Kleinberg et al., 2018) in a non-asymptotic way under some conditions not on the original objective but on the smoothed objective function. Combination with the Markov chain view for nonconvex objective (Yu et al., 2020) may be helpful in more detailed analyses.

Step size and Minibatch.   SGD with a large step size often suffers from stochastic gradient noise and becomes unstable. This is the reason why we should take a smaller step size so that SGD converges. In this sense, the minibatching of stochastic gradients clearly plays the same role as the step size and sometimes brings additional gains. For instance, Smith et al. (2017) empirically demonstrated that the number of parameter updates can be reduced, maintaining the learning curves on both training and test datasets by increasing minibatch size instead of decreasing step size. We remark that our analysis can incorporate the minibatch by dividing σ_1^2 and σ_2^2 in Proposition 3.4 and 3.2 by the minibatch size, and we can see certain improvements of optimization accuracy as well.

Edge of Stability.   Recently, Cohen et al. (2021) showed the deterministic gradient descent for deep neural networks enters Edge of Stability phase. In the traditional optimization theory, the step size is set to be smaller than $1/L$ to ensure stable convergence and we also make such a restriction. On the other hand, the Edge of Stability phase appears when using a higher step size than $2/L$. In this phase, the training loss behaves non-monotonically and the sharpness finally stabilizes around $2/\eta$. This can be explained as follows (Lewkowycz et al., 2020); if the sharpness around the current parameter is large compared to the step size, then gradient descent cannot stay in such a region and goes to a flatter region that can accommodate the large step size. There are works (Arora et al., 2022; Ahn et al., 2022) which attempted to rigorously justify the Edge of Stability phase. Interestingly, their analyses are based on a similar intuition to ours, but we consider a different regime of step sizes and a different factor (stochastic noise or larger step size than $2/L$) brings the implicit bias towards flat regions. We believe establishing a unified theory is interesting future research.

Averaged SGD.   The averaged SGD (Ruppert, 1988; Polyak and Juditsky, 1992) is a popular variant of SGD, which returns the average of parameters obtained by SGD aiming at stabilizing the convergence. Because of the better generalization performance, many works conducted convergence rate analysis in the expected risk minimization setting and derived the asymptotically optimal rates $O(1/\sqrt{T})$ and $O(1/T)$ for non-strongly convex and strongly convex problems (Nemirovski et al., 2009; Bach and Moulines, 2011; Rakhlin et al., 2012; Lacoste-Julien et al., 2012). However, the schedule of step size is basically designed to optimize the original objective function, and hence the implicit bias coming from the large step size will eventually disappear. When applying a non-diminishing step size schedule, the non-zero optimization error basically remains. What we do in this paper is to characterize it as the implicit bias toward a flat region.

E.1 Technical difference.

The proof idea of Proposition 3.4 relies on the alternative view of SGD (Kleinberg et al., 2018) which shows the existence of an associated SGD for the smoothed objective. However, since its stochastic gradient is a biased estimator, they showed the convergence not to the solution but to a point at which a sort of one-point strong convexity holds, and avoid the treatment of a biased estimator. Hence, the optimization of the smoothed objective is not guaranteed in their theory. On the other hand, optimization accuracy is the key in our theory, thus we need nontrivial refinement of the proof under a *normal* one-point strong convexity at the solution.