# Generalization Analysis of Stochastic Weight Averaging with General Sampling

**Peng Wang** [1 2] **Li Shen** [3 4] **Zerui Tao** [5 6] **Shuaida He** [7] **Dacheng Tao** [2]

## Abstract

Stochastic weight averaging (SWA) method has empirically proven its advantages compared to stochastic gradient descent (SGD). Despite it is widespread used, theoretical investigations have been limited, particularly in scenarios beyond the ideal setting of convex and sampling with replacement. However, non-convex cases and sampling without replacement are very practical in real-world applications. The main challenges under the above settings are two-folds: (i) All the historical gradient information introduced by SWA is considered, while the analysis of SGD using the tool of uniform stability requires only to bound the current gradient. (ii) The $(1+\alpha\beta)$-expansion property causes the boundary of each gradient step dependent on the previous step, making the boundary of each historical gradient in SWA nested and the theoretical analysis even harder. To address the theoretical challenges, we adopt mathematical induction to find a recursive representation that bounds the gradient at each step. Based on this, we establish stability bounds supporting sampling with and without replacement in the non-convex setting. Furthermore, the derived generalization bounds of SWA are sharper than SGD. At last, experimental results on several benchmarks verify our theoretical results.

## 1. Introduction

The generalization ability of deep neural networks is a primary concern, particularly given their capacity to memorize large datasets and potential for overfitting. A prevalent method to improve this ability is Stochastic Weight Averaging (SWA) (Izmailov et al., 2018), which involves averaging the model parameters obtained through Stochastic Gradient Descent (SGD). SWA has empirically demonstrated significant benefits in enhancing generalization across various applications, including large-scale network training (Izmailov et al., 2018; Lu et al., 2022), adversarial learning (Xiao et al., 2022), etc. Indeed, SWA and related model averaging techniques have become standard practices in training deep neural network models. Consequently, a theoretical analysis of SWA's generalization capabilities is vital to further its adoption in the deep learning community.

Due to the great complexity of neural networks, it is typically challenging to theoretically study the generalization properties of deep learning optimizers. To address the issue, one powerful tool is the sensitivity analysis (Bousquet & Elisseeff, 2002), which builds connections between generalization and stability. In specific, one can bound the generalization gap by analyzing stabilities of the algorithm. Based on this, Hardt et al. (2016) study the generalization property of SGD for both convex and non-convex functions. Since then, there are many theoretical works based on stability to study SGD and its variants. Hardt et al. (2016) also give generalization bounds of model averaging similar to SWA, but constrain to the convex case and sampling with replacement. Kuzborskij & Lampert (2018) study the generalization ability of SGD in the case of sampling without replacement. Yang et al. (2021) extend the analysis of SGD to online learning setting, where the data points are sampled without replacement. Xiao et al. (2022) study the generalization bounds of SWA in adversarial training, but still limit to the convex case. However, all these works do not study the generalization of SWA under the more practical non-convex settings. Moreover, the analysis of sampling without replacement of SWA is also missing. In this paper, we aim to give a thorough investigation of SWA under non-convex setting, while considering different sampling mechanisms.

To conduct these theoretical studies, there are three main challenges. ① The first challenge comes from the parameter averaging in SWA. For SGD, building the boundary for the step $T$ only requires to consider the gradient generated by the $T$-th sample. However, since each step of the SWA is averaged, all the historical information before the $T$-th step should be counted. To bound each step of the SWA, each step must take into account the above-mentioned operations,

[1]Huazhong University of Science and Technology, China [2]Nanyang Technological University, Singapore [3]Sun Yat-sen University, China [4]JD Explore Academy, China [5]Tokyo University of Agriculture and Technology, Japan [6]RIKEN AIP, Japan [7]Southern University of Science and Technology, China. Correspondence to: Li Shen <mathshenli@gmail.com>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

*Table 1.* Comparison of SWA with SGD on different settings. Here $T$ represents iterations, $n$ denotes the size of datasets. $\beta$ and $\gamma$ are Lipschitz constants, where $\gamma$ depends on the data-generating distribution and initialization point of SGD. c is a small constant. We can derive that SWA has sharper bounds compared to SGD in the different settings, particularly in convex optimization with sampling without replacement, where there is a significant improvement from $\mathcal{O}(T)$ to $\mathcal{O}(\ln T)$.

| SETTINGS | ALGORITHM | GENERALIZATION BOUND |
|---|---|---|
| CONVEX & SAMPLING WITH REPLACEMENT | SGD | $\mathcal{O}(2T/n)$ (HARDT ET AL., 2016) |
| | SWA | $\mathcal{O}(T/n)$ THEOREM 4.1 (HARDT ET AL., 2016; XIAO ET AL., 2022) |
| CONVEX & SAMPLING WITHOUT REPLACEMENT | SGD | $\mathcal{O}(T/n)$ (KUZBORSKIJ & LAMPERT, 2018) |
| | SWA | $\mathcal{O}(\ln T/n)$ THEOREM 4.2 |
| NON-CONVEX & SAMPLING WITH REPLACEMENT | SGD | $\mathcal{O}(T^{\frac{c\beta}{1+c\beta}}/n)$ (HARDT ET AL., 2016) |
| | SWA | $\mathcal{O}(T^{\frac{c\beta}{2+c\beta}}/n)$ THEOREM 5.1 |
| NON-CONVEX & SAMPLING WITHOUT REPLACEMENT | SGD | $\mathcal{O}(T^{\frac{c\gamma}{1+c\gamma}}/n)$ (KUZBORSKIJ & LAMPERT, 2018) |
| | SWA | $\mathcal{O}(T^{\frac{c\beta}{2+c\beta}}/n)$ THEOREM 5.2 |

which brings great analytical difficulty. ② Secondly, for non-convex functions, the $(1+\alpha\beta)$-expansion property causes the boundary of each gradient step dependent on the previous step, making the boundary of each historical gradient in SWA nested and the theoretical analysis even harder. In specific, due to the use of the $(1+\alpha\beta)$-expansion property in our analysis, we cannot establish the $T$-th gradient at once. We must also consider the effect of the $(T-1)$-th gradient and recursively compute the bound for the $T$-th step. ③ Finally, the sampling without replacement method also contains dependencies between events, which requires us to rethink the boundaries in this situation.

To mitigate these theoretical deficiencies, we find the recursive representation that bounds the gradient at each step using mathematical induction. Transform the problem of bounding each step into a problem of bounding a finite sum, which solves the difficulties in non-convex settings. What's more, we cast the case of sampling without replacement as a classic combinatorial probability problem in probability studies. We conduct a deep analysis of the impact on historical information at each step of sample selection based on this framework and establish stability bounds. Based on this, we present a thorough generalization analysis of SWA in various settings and establish stability bounds. Compared with existing generalization analysis of SGD methods (see Table 1), our results suggest that SWA can improve the generalization bound in various settings.

### 1.1. Our Contributions

This paper mainly focuses on theoretical exploration, and some experimental results are intended to verify its correctness. In specific, we establish a number of generalization bounds for the SWA to illustrate its good properties compared with SGD. The main theoretical results are summarized in Table 1. Our contributions are listed as follows.

- We focus on the theoretical exploration of the generalization ability of SWA and derive stability bounds for SWA using the notion of uniform stability. Based on this, we provide a theoretical perspective on why SWA improves generalization better than the SGD.

- We derive stability-based generalization bounds for SWA in the various settings – convex or non-convex, sampling with or without replacement – each of which is sharper than the bound of SGD (see Table 1).

- We construct a recursive representation that bounds the gradient at each step using mathematical induction, to establish stability bounds of SWA in the non-convex setting without replacement sampling based on combinatorial probability.

- We provide two experiments for SWA with or without replacement sampling cases to verify our results based on the metric parameter distance and generalization error on the MNIST, CIFAR10, and Adult datasets, respectively. The experimental results also coincide with our theoretical findings.

## 2. Related Work

**SWA algorithm.** Model averaging techniques were fristly used in convex optimization (Ruppert, 1988; Polyak & Juditsky, 1992), which showed advantages in generalization and convergence speed. Then, this idea was extended to deep neural networks (Neklyudov et al., 2018; Garipov et al., 2018). To bypass the overhead of training multiple models, Izmailov et al. (2018) proposed the stochastic weight averaging (SWA), which averages parameters on the learning trajectory of SGD. Cha et al. (2021) modify SWA and empirically show flatter minima can be found by the modification. More recently, trainable weight averaging (TWA) (Li et al., 2022) employs trainable averaging coefficients to further improve the efficiency of SWA. Despite the widespread use

of SWA, a thorough generalization analysis for non-convex functions and general sampling schemes is still missing.

**Generalization analysis.** Many theoretical tools have been established to study the generalization ability of algorithms, such as the VC-dimension (Blumer et al., 1989; Vapnik, 2006), the Rademacher complexity (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2000; Koltchinskii, 2001), and the PAC-Bayesian theory (McAllester, 1999a;b). Stability analysis tries to build the connection between generalization ability and algorithm stability (Devroye & Wagner, 1979; Kearns, 1989; Bousquet & Elisseeff, 2002; Mukherjee et al., 2006; Shalev-Shwartz et al., 2010). In specific, Bousquet & Elisseeff (2002) propose the algorithm stability from the perspective of statistical learning theory. Hardt et al. (2016) pioneer the use of algorithm stability to study the generalization bounds of SGD, and subsequent improved methods include (Charles & Papailiopoulos, 2018; Zhou et al., 2018b; Yuan et al., 2019; Lei & Ying, 2020b). This tool is then applied in various of applications including online learning (Yang et al., 2021), adversarial training (Xiao et al., 2022), decentralized learning (Zhu et al., 2023), and federated learning (Sun et al., 2023b;a). In particular, Hardt et al. (2016); Xiao et al. (2022) establish generalization and stability analysis of SWA, but constrained to convex functions and sampling with replacement.

**SGD without replacement sampling.** Traditional analysis of SGD usually assumes sampling with replacement with uniform probabilities (Bottou, 2009). Shamir (2016) gave convergence guarantees of SGD without replacement sampling. Extensions include Rajput et al. (2020); Nguyen et al. (2021); Das et al. (2022); Sherman et al. (2021); Nagaraj et al. (2019); Zhou et al. (2018a). In this work, we focus on generalization ability, while we also provide optimization bound. Notably, Kuzborskij & Lampert (2018) derive generalization bounds for the SGD algorithm based on stability, under the setting of without replacement sampling, while we focus on the analysis of SWA.

The most related work to ours is (Hardt et al., 2016). However, Hardt et al. (2016) merely provide a stability bound for SWA in the assumption of convex and sampling with replacement and they do not consider the results of SWA in other settings. In contrast, we focus on the generalization analysis for SWA and establish stability bounds in all cases, convexity and non-convexity, samples with and without replacement, respectively. In summary, we present a thorough analysis of SWA's generalization bounds and rigorously demonstrate the superiority of SWA over vanilla SGD.

## 3. Preliminary

The necessary notations, assumptions, definitions of stability and generalization are given in this section.

### 3.1. Problem Setup

Let $g(w, z)$ be a loss function that measures the loss of the predicted value of the parameter $w$ at a given sample $z$. There is an unknown distribution $\mathcal{D}$ over examples from some space $\mathcal{Z}$, and a sample dataset $S = (z_1, z_2, ..., z_n)$ of $n$ examples i.i.d. drawn from $\mathcal{D}$. Then the *population risk* and *empirical risk* are defined as

$$\textbf{Population Risk:} \quad \min_w \{R_{\mathcal{D}}[w] = E_{z\sim\mathcal{D}} g(w; z)\} \quad (1)$$

$$\textbf{Empirical Risk:} \quad \min_w \{R_S[w] = \frac{1}{n}\sum_{i=1}^{n} g(w; z_i)\}. \quad (2)$$

The generalization error of a model $w$ is the difference $\epsilon_{gen} = R_{\mathcal{D}}[w] - R_S[w]$. Moreover, we assume function $g$ satisfies the following *Lipschitz* and *smoothness* assumption.

**Definition 3.1** ($L$-Lipschitz)**.** For a fixed parameter $z \in Z$, a function $g(u, z) : u \in \Omega \subset R^n \to R$ is $L$-Lipschitz if for all $u, v \in \Omega$ and $\exists L \geq 0$ such that

$$|g(u; z) - g(v; z)| \leq L\|u - v\|, \quad (3)$$

where the constant $L$ is uniform for the parameter $z$ and $\|\cdot\|$ be Euclidean norm. And it implies that $\| \bigtriangledown g(u, z)\| \leq L$ if $g \in C^1(\Omega)$.

**Definition 3.2** ($\beta$-smooth)**.** For a fixed parameter $z \in Z$, a function $g(u, z) : u \in \Omega \subset R^n \to R$ is $\beta$-smooth if for all $u, v \in \Omega$ and $\exists \beta \geq 0$ such that

$$\| \bigtriangledown g(u, z) - \bigtriangledown g(v, z)\| \leq \beta\|u - v\|, \quad (4)$$

where the constant $\beta$ is uniform for the parameter $z$ and $\|\cdot\|$ be Euclidean norm.

**Definition 3.3** (convex function)**.** A function $g : \Omega \to R$ is convex if for all $u, v \in \Omega$, then we have

$$g(u; z) \leq g(v; z) + \langle \bigtriangledown g(v), u - v\rangle. \quad (5)$$

### 3.2. SGD and SWA Algorithm

**SGD.** For the given training set $S = (z_1, z_2, ..., z_n)$ and the target function $g$, the general update rule of the stochastic gradient descent (SGD) algorithm is formulated as

$$w_{t+1} = w_t - \alpha \nabla_w g(w_t, z_{i_t}), \quad (6)$$

where $\alpha$ is the fixed step size, $z_{i_t}$ is the sample chosen in iteration t. We consider two ways to choose samples from dataset $S$. Choosing $z_{i_t}$ with replacement is a standard way to train the model (Bottou, 2009). In addition, we also consider the setting of sampling without replacement (Bertsekas, 2011; Bottou, 2012; Gürbüzbalaban et al., 2019).

We consider two popular schemes that are commonly used for choosing the samples. One is to pick $i_t \sim$

$Uniform\{1, \cdots, n\}$ at each step. The other is to choose a random permutation over $\{1, \cdots, n\}$ and cycle through the examples repeatedly in the order determined by the permutation. This setting is commonly explored in analyzing the stability (Hardt et al., 2016; Xiao et al., 2022). On the other hand, sampling from $S$ without replacement, which requires the number of iterates of an algorithm is less than to the size of $S$ (Kuzborskij & Lampert, 2018). Our developed analysis framework is general and holds for both sampling with and without replacement.

**SWA.** According to the gradient update rule, recursively $w_t$ is represented as $w_t = w_0 - \alpha \sum_{i=1}^{t} \bigtriangledown g(w_{i-1}, z_i)$, where $w_0$ is the initial point. SWA is formulated as

$$\bar{w}_T = w_0 - \frac{\alpha}{T} \sum_{t=1}^{T} \sum_{i=1}^{t} \bigtriangledown g(w_{i-1}, z_i), \tag{7}$$

where $T$ is the maximum number of iterates. It is not difficult to find the relationship between $\bar{w}_T$ and $w_T$, i.e.,

$$\bar{w}_T = \frac{1}{T}\left(\sum_{t=1}^{T-1} w_t + w_T\right) = \frac{T-1}{T}\bar{w}_{T-1} + \frac{1}{T} w_T. \tag{8}$$

Furthermore, we have

$$\bar{w}_T = \bar{w}_{T-1} - \frac{\alpha}{T(T-1)} \sum_{i=1}^{T-1} i \cdot \nabla g(w_i, z_{i+1}), \tag{9}$$

and there is $\bar{w}_1 = w_1 = w_0 - \alpha \nabla g(w_0, z_1)$ when $T = 1$, which is often used in the proof section of this chapter and the proof of Eq. (9) is placed in the **Appendix** A.1.

### 3.3. Stability and Generalization Definition

Establishing a generalization error bound for certain algorithm and using it to find variable dependencies is a primary means of studying generalization ability. Hardt et al. (2016) link the *uniform stability* of the learning algorithm to the expected generalization error bound and derive the generalization error bound for SGD algorithm. The expected generalization error of a model $w = A_S$ trained by certain randomized algorithm $A$ defined as

$$\mathbb{E}_{S,A}\left[R_S\left[A_S\right] - R_{\mathcal{D}}\left[A_S\right]\right]. \tag{10}$$

Next, we employ the following notion of *uniform stability*.

**Definition 3.4** ($\epsilon$-Uniformly Stable)**.** A randomized algorithm $A$ is $\epsilon$-uniformly stable if for all data sets $S, S' \in Z^n$ such that $S$ and $S'$ differ in at most one example, we have

$$\sup_{z \in Z}\left\{\mathbb{E}_A\left[g(A_S; z) - g(A_{S'}; z)\right]\right\} \leq \epsilon. \tag{11}$$

We recall the important theorem that uniform stability implies *generalization in expectation* (Hardt et al., 2016). The proof is based on Bousquet & Elisseeff (2002, Lemma 7) and very similar to Shalev-Shwartz et al. (2010, Lemma 11).

**Theorem 3.5.** (Generalization in Expectation, Hardt et al. (2016, Theorem 2.2)) *Let $A$ be $\epsilon$-uniformly stable. Then,*

$$\left|\mathbb{E}_{S,A}\left[R_S\left[A_S\right] - R_{\mathcal{D}}\left[A_S\right]\right]\right| \leq \epsilon. \tag{12}$$

Here, the expectation is taken only over the internal randomness of $A$. This theorem clearly states that if an algorithm is uniform stability, then its generalization error is small. Therefore, we can characterize the generalization error bound of an algorithm by controlling its uniform stability according to this theorem.

**Property of SGD iterate.** Since stability is to investigate the impact of input perturbations on the output, we need to characterize the state of the sequence updates in two different scenarios. Let $w_T$ and $w'_T$ be the outputs of SGD after running $T$ steps on $S$ and $S'$, which are two datasets with only one sample difference. Next, we consider the expansion properties of $\|w_T - w'_T\|$ after updating one step.

**Lemma 3.6** (Lemma 3.6, Hardt et al. (2016))**.** *Assume that the function $g$ is $\beta$-smooth. Then,* **(1). ($(1+\alpha\beta)$-expansive)** $\|w_{T+1} - w'_{T+1}\| \leq (1+\alpha\beta)\|w_T - w'_T\|$*;* **(2). (non-expansive)** *Assume that $g$ is convex. Then for any $\alpha \leq \frac{2}{\beta}$, we have $\|w_{T+1} - w'_{T+1}\| \leq \|w_T - w'_T\|$.*

Lemma 3.6 tells us, in general, smoothness will imply that the gradient updates cannot be overly expansive. In addition, when the function is convex and the step size is sufficiently small, the gradient update becomes non-expansive. The proof of Lemma 3.6 is deferred to **Appendix** A.2. and the more results can be found in several literature (Hardt et al., 2016; Xiao et al., 2022). Notable references are Polyak (1987) and Nesterov (2004).

Then, the stability bounds will be constructed if we can control $\|\bar{w} - \bar{w}'\|$ in different scenarios, indicating that SWA is stable. We will provide how to control it recursively using the properties of gradient updates in the next section.

## 4. Generalization Bound under Convexity

Now we start with a stability bound that target function $g$ is convex. Note that the analysis of SWA is more challenging than that of SGD. To see this, recall that the bound of $w_{T+1}$ is obtained from $w_T$, bounding only $\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\|$. However, in the analysis of SWA, the term

$$\frac{\alpha}{T(T+1)} \sum_{i=1}^{T} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| \tag{13}$$

should be carefully bounded. This is because $\bar{w}_{T+1}$'s expression after a step is not only related to $\bar{w}_T$ but also the accumulation of gradients from previous steps, the problem we need to deal with is more complex.

Another obstacle arises in the scenario of sampling without replacement. Unlike sampling with replacement that follows a discrete uniform distribution, sampling without replacement involves a more complicated combinatorial problem that requires further case-by-case discussion. Due to limited space, we merely provide the **Proof Sketch** in the main file. The detailed proof is placed in the **Appendix** B.

### 4.1. Sampling with Replacement

By using the $L$-Lipschitz property of the target function, we have

$$\mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq L\mathbb{E}\bar{\delta}_T \tag{14}$$

for all $\bar{w}_T$ and $\bar{w}'_T$, where $\bar{\delta}_T = \|\bar{w}_T - \bar{w}'_T\|$. This implies that if the algorithm is stable, its generalization bound is immediately obtained because $\bar{\delta}_T$ is bounded. Below, we provide our first theoretical result, considering the stability bound in the case of sampling with replacement.

**Theorem 4.1.** *Assume that the loss function $g(w; z)$ is convex, $L$-Lipschitz and $\beta$-smooth for all given $z \in \mathcal{Z}$ with sizes $n$. Suppose we run SWA with step sizes $\alpha \leq \frac{2}{\beta}$ for $T$ steps, where each step samples $z_i$ from $\mathcal{Z}$ uniformly with replacement. Then, SWA has uniform stability of*

$$\epsilon_{gen} = \mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq \frac{\alpha L^2 T}{n}. \tag{15}$$

*Proof sketch.* To complete the proof, we need to recursively bound $\bar{\delta}_{T+1}$ and its expectation $\mathbb{E}[\bar{\delta}_{T+1}]$ according to Eq. (14).

First, to bound $\bar{\delta}_{T+1}$, we need to solve the challenge in Eq. (13) and divide it into two parts

$$\begin{aligned}\bar{\delta}_{T+1} \leq \bar{\delta}_T + \frac{\alpha}{T+1}\|\nabla g(w_T, z_{T+1}) - \nabla g'(w'_T, z_{T+1})\| \\ + \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|.\end{aligned} \tag{16}$$

Next, we try to estimate the last two terms separately. For ease of notation, we define the event $A$ as the selection of distinguished sample pairs $(z, z')$ from the datasets $S$ and $S'$. When sampling with replacement, we have $p(A) = \frac{1}{n}$.

**(1)** Bounding $\|\nabla g(w_T, z_{T+1}) - \nabla g'(w'_T, z_{T+1})\|$. On one hand, when event $A$ occurs at step $T+1$ with probability $\frac{1}{n}$, we only need to use $L$-Lipschitz to bound $\nabla g(w_T)$ and $\nabla g'(w'_T)$ respectively. On the other hand, with probability $1 - \frac{1}{n}$ that $A$ does not occur, we can use the *non*-expansive update rule from Lemma 3.6, based on the fact that the objective function is convex and $\alpha \leq \frac{2}{\beta}$. In summary, $\|\nabla g(w_T, z_{T+1}) - \nabla g'(w'_T, z_{T+1})\| \leq \frac{2L}{n}$.

**(2)** Bounding the third term on the historical gradient.

$$\begin{aligned}&\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| \\ &= \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1}\frac{2Li}{n} = \frac{(T-1)\alpha L}{n(T+1)}.\end{aligned} \tag{17}$$

Since each step $i \in [1, \cdots, T-1]$ executes sampling with replacement, we can bound them in the way above.

Second, by merging the above two results, we derive

$$\begin{aligned}\mathbb{E}\left[\bar{\delta}_{T+1}\right] &\leq (1 - \frac{1}{n})\bar{\delta}_T + \frac{1}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T+1}\right) + \frac{(T-1)\alpha L}{n(T+1)} \\ &\leq \mathbb{E}\left[\bar{\delta}_T\right] + \frac{\alpha L}{n}.\end{aligned}$$

Taking summation over $T$ steps, we get $\mathbb{E}\left[\bar{\delta}_T\right] \leq \frac{\alpha L T}{n}$. Finally, substituting it to Eq. (14) yields the desired result. We leave the details of this proof to Appendix B.1. □

### 4.2. Sampling without Replacement

Next, we provide the generalization error bound of SWA under sampling without replacement.

**Theorem 4.2.** *Assume that the loss function $g(w; z)$ is convex, $L$-Lipschitz and $\beta$-smooth for all given $z \in \mathcal{Z}$ with sizes $n$. Suppose we run SWA with step sizes $\alpha \leq \frac{2}{\beta}$ for $T(T \leq n)$ steps, where each step samples $z_i$ from $\mathcal{Z}$ without replacement. Then, SWA has uniform stability of*

$$\epsilon_{gen} = \mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq \frac{4\alpha L^2 \ln T}{n}. \tag{18}$$

*Proof sketch.* When applying SWA via sampling without replacement, event $A$ may occur only once during the training procedure, but at any step. Note that the probability of $A$ occurring varies across different training stages and can be determined by computing combinatorial probability. We split the event into three cases and discuss them in detail.

**(1)** If $A$ occurs in step $T+1$ with a probability of $\frac{1}{n}$, we will bound the Eq. (13) with $2L$. Since the same sample will be selected for the first $T$ steps of updates, we bound all the historical information using the *non*-expansive property. Then we obtain $\bar{\delta}_{T+1} \leq \bar{\delta}_T + \frac{2\alpha L}{T+1}$.

**(2)** If the event $A$ occurs after step $T+1$ with a probability of $\frac{n-T-1}{n}$, we bound all the first $T+1$ steps using the *non*-expansive property. And we have $\bar{\delta}_{T+1} \leq \bar{\delta}_T$.

**(3)** If the event $A$ occurs in the previous $T$ steps with a probability of $\frac{T}{n}$, we select step $i$ from 1 to $T$ with a probability of $\frac{1}{T}$ and bound the Eq. (13) using $2L$. The *non*-expansive property is then applied to obtain $\bar{\delta}_{T+1} \leq \bar{\delta}_T + \frac{2\alpha L}{T(T+1)}$.

Then, we obtain the expectation

$$\mathbb{E}\left[\bar{\delta}_{T+1}\right] \leq \frac{T}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T(T+1)}\right) + \frac{n-T-1}{n}\bar{\delta}_T + \frac{1}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T+1}\right) \leq \mathbb{E}\left[\bar{\delta}_T\right] + \frac{2\alpha L}{n}\cdot\frac{2}{T+1}$$

and the bound

$$\mathbb{E}\left[\bar{\delta}_T\right] \leq \frac{2\alpha L}{n}\cdot\sum_{t=1}^{T}\frac{2}{t} \leq \frac{2\alpha L}{n}\cdot 2\ln T. \tag{19}$$

Then, substituting it to Eq. (14) yields the desired result. We leave the details of this proof to Appendix B.2. □

*Remark* 4.3. The stability bound of SWA under the convex assumption has already been explored in previous work. For instance, Hardt et al. (2016) present a result similar to ours in their online draft, while Xiao et al. (2022) investigate the stability of SGD in the context of adversarial training. However, their results are limited to considering sampling with replacement only. We extent existing results in two ways: by providing a sharper bound, which is only half of SGD under the same setting, and by addressing the gap of sampling without replacement with a new bound.

*Remark* 4.4. A theoretical study for SGD (Kuzborskij & Lampert, 2018), which gives the data-dependent stability bound $\mathcal{O}(T)$ in the convex. This bound depends on the choice of initial points and requires that the variance of the random gradient should not too large. If then, there will be no improvement over the case of sampling with replacement. In comparison, our results suggest there is a sharper bound $\mathcal{O}(\ln T)$, which is a significant improvement for SWA.

*Remark* 4.5. Note that in Eq. (6) of (Hoffer et al., 2017) and Eq. (2) of (Ziyin et al., 2021), the covariance of the noise of SGD is proven to be the same for sampling with or without replacement, while the number of data is much larger than the size of mini-batch. However, in our Theorems 4.1 and 4.2, sampling without replacement has much a sharper bound. The main reason for the sharper generalization error rates comes from our analytical process under the assumption of $T \leq n$ rather than from the different sampling methods. The reason for this setting is twofold: 1) To allow us to make a direct comparison with the bound of SGD (Kuzborskij & Lampert, 2018) in the same setting ($T \leq n$). 2) To simplify our theoretical analysis. Also, it should be noted that our main purpose is not to compare these two sampling strategies, but to show that SWA has better generalization bounds under these two sampling strategies.

## 5. Generalization Bound under Non-Convexity

In this part, we consider the case where target function $g$ is non-convex and assume that $g$ is $L$-Lipschitz and $\beta$-smooth as defined previously. Moreover, differing in the ideas of the convex case, the research scheme of the non-convex case is motivated by the arguments in (Hardt et al., 2016), which divides the objective into two parts whether to converge or not, and then bounding each of these separately, i.e.,

$$\mathbb{E}|g(\bar{w};z) - g(\bar{w}';z)| \leq \frac{t_0}{n} + L\mathbb{E}[\bar{\delta}_T|\bar{\delta}_{t_0}=0], \tag{20}$$

where $\bar{\delta}_{t_0}=0$ denotes coverage after $t_0$, $t_0 \in \{0,1,\cdots,n\}$. The detailed proof is placed in Appendix C.1.

Below, we first provide stability bounds using $(1+\alpha\beta)$-expansive property where a key technical challenge is solved. The dependence relationship between $\bar{\delta}_{T+1}$ and $\bar{\delta}_T$ introduced by the $(1+\alpha\beta)$-expansive property makes calculations more difficult, especially for historical gradients,

$$\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1}) - \nabla g(w_i',z_{i+1})\|. \tag{21}$$

For this, we first recursively bound $\|\nabla g(w_i,z_{i+1}) - \nabla g(w_i',z_{i+1})\|$. We find the recursive representation that bounds the gradient at each step by using mathematical induction. Then, we transform the problem of bounding each step into bounding a finite-sum problem, which solves the difficulties in non-convex settings.

### 5.1. Sampling with Replacement

**Theorem 5.1.** *Assume that the loss function $g(w;z) \in [0,1]$ is $L$-Lipschitz and $\beta$-smooth for all given $z \in \mathcal{Z}$ with sizes $n$. Suppose we run SWA with non-increasing step sizes $\alpha \leq \frac{c}{t}$ for $T$ steps, where each step samples $z_i$ from $\mathcal{Z}$ uniformly with replacement. To simplify, omitting constant factors that depend on $\beta$, $c$ and $L$, SWA has uniform stability as*

$$\epsilon_{gen} = \mathbb{E}|g(\bar{w}_T;z) - g(\bar{w}_T';z)| \leq \mathcal{O}\left(\frac{T^{\frac{\beta c}{\beta c+2}}}{n}\right). \tag{22}$$

*Proof sketch.* First, we find the recursive relationship as follows $\|w_i - w_i'\| \leq (1+\alpha\beta)\|w_{i-1} - w_{i-1}'\| + \frac{2\alpha L}{n}$. Based on this, we can get the equivalent form of Eq. (21)

$$\begin{aligned}&\frac{\alpha\beta}{T(T+1)}(\|w_1 - w_1'\| + \cdots + (T-1)\|w_{T-1} - w_{T-1}'\|)\\ &\leq \frac{\alpha\beta}{T(T+1)}\frac{2\alpha L}{n}\cdot\frac{T(T-1)}{2}\cdot\sum_{i=1}^{T-2}(1+\alpha\beta)^i\\ &+ \frac{\alpha\beta}{T(T+1)}\frac{2\alpha L}{n}\cdot\frac{T(T-1)}{2} \leq \frac{\alpha L}{n}(1+\alpha\beta)^{T-1}\end{aligned} \tag{23}$$

Second, let $\alpha = \frac{c}{t}$ and we obtain the expectation

$$\mathbb{E}\left[\bar{\delta}_{T+1}\right] \leq \exp\left((1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{2cL}{n}\cdot\frac{1+e^{c\beta}}{t}$$

where the limitation $\lim_{t\to\infty}(1+\frac{c\beta}{t})^t = e^{c\beta}$ is used. It's worth mentioning that we use the key inequality $\frac{T-t}{tT} \leq \frac{1}{2}\log(\frac{T}{t}), 1 \leq t \leq T$, which helps us complete and improve the stability bound. Therefore,

$$\begin{aligned}
\mathbb{E}\bar{\delta}_t &\leq \sum_{t=t_0+1}^{T} \exp\left((1-\frac{1}{n})c\beta \sum_{k=t+1}^{T}\frac{1}{k(k-1)}\right)\frac{cL}{n}\cdot M \\
&\leq \sum_{t=t_0+1}^{T} \exp\left((1-\frac{1}{n})c\beta\cdot\frac{T-t}{tT}\right)\frac{cL}{n}\cdot M \\
&\leq \sum_{t=t_0+1}^{T} \exp\left(\log(\frac{T}{t})\cdot\frac{(1-\frac{1}{n})c\beta}{2}\right)\frac{cL}{n}\cdot M \\
&\leq \frac{4L(1+e^{c\beta})}{(n-1)\beta}\cdot\left(\frac{T}{t_0}\right)^{\frac{c\beta}{2}}, \quad \text{where} \quad M=\frac{1+e^{c\beta}}{t}.
\end{aligned}$$

Last, plugging this final term back into Eq. (20) and minimizing the $t_0$, we obtain the bound and finish the proof. □

### 5.2. Sampling without Replacement

Finally, we consider the case where the function $g$ is non-convex and the setting of sampling without replacement.

**Theorem 5.2.** *Assume that loss function $g(w;z) \in [0,1]$ is $L$-Lipschitz and $\beta$-smooth for all given $z \in \mathcal{Z}$ with sizes $n$. Suppose we run SWA with non-increasing step sizes $\alpha \leq \frac{c}{t}$ for $T(T \leq n)$ steps, where each step samples $z_i$ from $\mathcal{Z}$ without replacement. To simplify, omitting constant factors that depend on $\beta$, $c$ and $L$, SWA has uniform stability of to simplify, omitting constant factors that depend on $\beta$, $c$ and $L$, we get*

$$\epsilon_{gen} = \mathbb{E}|g(\bar{w}_T;z) - g(\bar{w}'_T;z)| \leq \mathcal{O}\left(T^{\frac{\beta c}{\beta c+2}}/n\right). \quad (24)$$

In this part, a major difficulty is to obtain the bound for the event $A$ occurs in the previous step $T+1$ with probability $\frac{n-T-1}{n}$ based on the $(1+\alpha\beta)$-expansive property.

*Proof sketch.* There are two extreme situations, where $A$ occurs in step 1 or step $T$ with probability $\frac{1}{T}$, respectively. First, we consider that $A$ occurs in step 1. We bound step 1 with $2L$ and other steps with the $(1+\alpha\beta)$-expansive property. Then, we have

$$\begin{aligned}
&\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1}) - \nabla g(w'_i,z_{i+1})\| \\
&\leq \frac{2\alpha L}{T(T+1)}\left(\sum_{k=1}^{T-1}\frac{(T-k)(T+k-1)}{2}(\alpha\beta)^k\right) \quad (25) \\
&\leq \frac{\alpha L}{2}\sum_{k=1}^{T-1}(\alpha\beta)^k.
\end{aligned}$$

Let $\alpha = \frac{c}{t}$, we get $\sum_{k=1}^{T-1}\left(\frac{c\beta}{t}\right)^k \leq \frac{c\beta}{t-c\beta}$ based on the fact $t \gg c\beta$. Then, we obtain the expectation

$$\begin{aligned}
\mathbb{E}\left[\bar{\delta}_{t+1}\right] \leq& \frac{n-T-1}{n}\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{1}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T+1}\right) \\
&+\frac{T}{n}\left(\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{1}{T}\left(\frac{\alpha L}{2}\sum_{k=1}^{T-1}(\alpha\beta)^k(1+\alpha\beta)\right)\right) \\
&\leq \exp\left((1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{cL}{n}\cdot\frac{4+c\beta+c^2\beta^2}{t-c\beta}.
\end{aligned}$$

Combining the above with the key inequality $\frac{T-t}{tT} \leq \frac{1}{2}\log(\frac{T}{t}), 1 \leq t \leq T$, the expectation will be obtained.

Second, we discuss that $A$ occurs in step $T$. It's bounded by $2L$ at step $T$ and $(1+\alpha\beta)$-expansive at step $T+1$, respectively. We can eliminate others because $w_0 = w'_0$ and the same samples are selected for each step. Then, we get

$$\mathbb{E}\left[\bar{\delta}_{t+1}\right] \leq \left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{2\alpha^2\beta L(T-1)}{T(T+1)} + \frac{2\alpha L(T-1)}{T(T+1)}.$$

Next, a standard calculation process for obtaining expectation is executed. Finally, by choosing the maximum expectation of these two cases and minimizing it w.r.t. $t_0$, we obtain the desired results. And this completes the proof. □

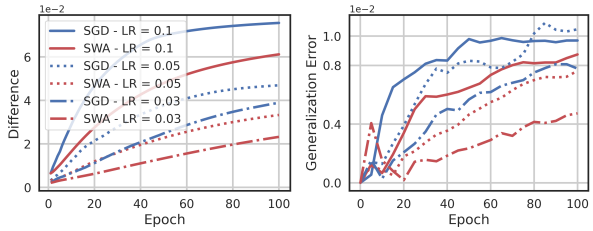
*Remark* 5.3. The assumption that $g(w;z) \in [0,1]$ in Theorem 5.1 and Theorem 5.2 is made for ease of presentation. Removing this condition will not alter the final results, but merely scales them by a constant. Further details can be found in Appendix C.2 and C.3.

*Remark* 5.4. The generalization bound $\mathcal{O}(T^{\frac{c\beta}{c\beta+2}})$ established for SWA in Theorem 5.1 represent a slight improvement compared to existing work $\mathcal{O}(T^{\frac{c\beta}{c\beta+1}})$ according to (Hardt et al., 2016). However, this subtle enhancement suggests that SWA can transform the $(1+\alpha\beta)$-expansive of SGD into $(1+\frac{\alpha\beta}{T})$-expansive, where $T$ is the number of training steps. Thus, this improvement is fundamental.

*Remark* 5.5. In the same setting as theorem 5.2, Kuzborskij & Lampert (2018) provide a bound $\mathcal{O}(T^{\frac{c\gamma}{1+c\gamma}})$ for SGD, where $\gamma$ is the Lipschitz parameter of the gradient. As discussed in the previous remark, our bound $\mathcal{O}(T^{\frac{c\beta}{c\beta+2}})$ again demonstrates the benefit of SWA. Moreover, the parameter $\gamma$ in Kuzborskij & Lampert (2018) is not merely a Lipschitz constant, as $\beta$ in our Theorem 5.2, but rather depends on the potential data distribution and the initial point used in SGD. Considering this, we provide a better result with fewer theoretical restrictions.

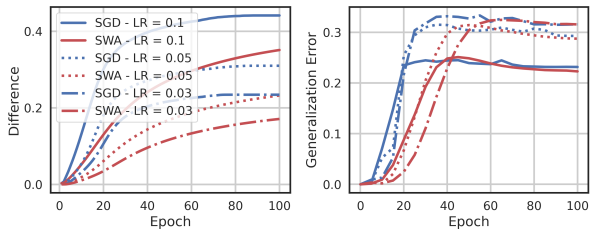
## 6. Experimental Evaluation

The purpose of the experiment is to verify that SWA can improve the generalization ability. The stable boundary heavily




(a) Parameter difference (b) Generalization error

*Figure 1.* Stability results of SGD and SWA with replacement sampling on the MNIST dataset.




(a) Parameter difference (b) Generalization error

*Figure 2.* Stability results of SGD and SWA with replacement sampling on the CIFAR10 dataset.

depends on the training time, measured by the number of iterates, so we compare the generalization ability of SWA and SGD under the same iterates. We train deep neural networks on different datasets to verify the theory.

### 6.1. Training with Replacement Sampling

We first investigate the stability of SGD and SWA under the standard training setting, *i.e.*, selecting samples with replacement for multiple epochs. Following the setting of Hardt et al. (2016), we train a LeNet (LeCun et al., 1998) with two convolutional layers on the MNIST dataset (Deng, 2012), and a VGG16 (Simonyan & Zisserman, 2014) on the CIFAR10 dataset (Krizhevsky et al., 2009). To construct two different datasets, we randomly remove one sample from the training set to construct dataset $S$. Then, another dataset $S'$ is constructed by replacing one random data point in $S$ with the deleted one. We train the model on these two datasets with the same initialization and settings. The batch size is set as 128. Although our theoretical results are established for SWA with batch size 1, we believe that the use of mini-batches in our experiments applies equally well to our theoretical results. When the batch size is set greater than 1, we just need to adjust the number of steps to be averaged in the SWA expression accordingly. The learning rate is chosen from $\{0.1, 0.05, 0.03\}$. Similar as Hardt et al. (2016), to validate the theoretical results, neither data augmentation nor learning rate scheduling is adopted.

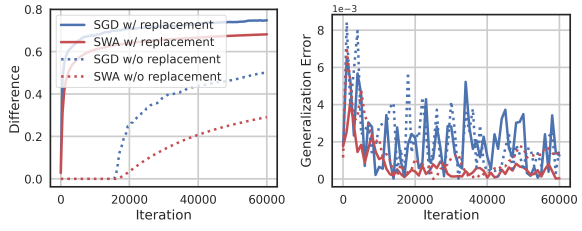
To evaluate the stability of the two algorithms, we compare the *parameter distance* and *generalization error*. Concretely, the parameter difference is defined as the Euclidean distance between parameters of the two model, *e.g.*, $\sqrt{\|w - w'\|^2/(\|w\|^2 + \|w'\|^2)}$, where $w$ and $w'$ denote all the parameters of models trained on $S$ and $S'$ respectively. The generalization error is defined as the absolute value of the difference between the training error and test error.

The results are shown in Figures 1 and 2 for MNIST and CIFAR10, respectively. In specific, Figures 1(a) and 2(a) show that SWA always achieves smaller parameter differences, which validates our analysis. Meanwhile, SWA also tends to have smaller generalization error than vanilla SGD, as shown in Figures 1(b) and 2(b). These results validate our theoretical results in Theorem 5.1.

### 6.2. Training without Replacement Sampling

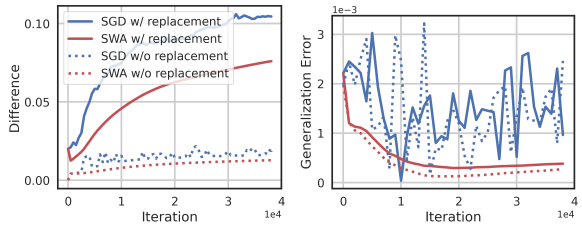
Below, we validate the theory of training deep neural networks with SGD and SWA by sampling without replacement. Specifically, we train LeNet on the MNIST dataset and a multi-layer perceptron (MLP) on the Adult dataset in the UCI repository (Becker & Kohavi, 1996). The Adult dataset is normalized using $l_1$ norm. The MLP consists of three layers with 50 hidden units each and the Tanh activation function. The datasets are constructed in the same manner as Section 6.1. For all the experiments in this subsection, we set the mini-batch size as 1 and learning rate as 0.03 for MNIST dataset and 0.01 for Adult dataset, respectively. As in Section 6.1, we report the parameter differences and generalization errors for both SGD and SWA.

Results are shown in Figures 3 and 4. As we can see in Figures 3(a) and 4(a), SWA still yields lower parameter differences in general. Moreover, when training without replacement sampling, the parameter differences are smaller, which satisfies our theoretical results in Theorem 5.2. These observations are similar in the generalization error, shown in Figures 3(b) and 4(b). In summary, the experiments coincide with our theoretical analysis, i.e., SWA achieves better generalization than vanilla SGD optimizer. Moreover, compared to SGD, SWA is better at finding flat landscape in real-world tasks, and flat minima tend to have better test results (Izmailov et al., 2018). Therefore, SWA works better than SGD in terms of the test accuracy. From our theoretical analysis, the average coefficient $1/T$ of the SWA with ergodic averaging plays an important role. It can directly lead to the reduction of the generalization error in both convex and non-convex cases. We show results of test error in Appendix D. SWA generally leads to more stable curves and better test error. These results coincide with our theoretical analysis in Theorems 5.1 and 5.2.




(a) Parameter difference (b) Generalization error

*Figure 3.* Stability results of SGD and SWA without replacement sampling on the MNIST dataset.




(a) Parameter difference (b) Generalization error

*Figure 4.* Stability results of SGD and SWA without replacement sampling on the Adult dataset.

## 7. Conclusion

We have established generalization bounds for SWA based on stability in different cases. Based on this analysis framework, by comparing the SGD algorithm and further analysis, we theoretically explain the natural reason why SWA can improve the generalization ability. From the perspective of expansion properties, our research process not only generates sharper boundaries, but also directly reflects the reason that the expansion caused by SWA becomes the original $\frac{1}{T}$. Our theoretical results show that the generalization bound depends directly on the training time and requires finding the solution as fast as possible. Combining the geometric features of the loss surface to accelerate the training process is also something we need to explore further. Our analysis requires a rigorous discussion under the assumptions of convexity and non-convexity. How to build a model that can avoid these theoretical assumptions in practical applications will be our next research direction.

**Limitations:** The theoretical analysis of algorithms are carried out under some classical assumptions such as $L$-Lipschitz and $\beta$-smooth, which seem to be standard assumptions in the analysis of stability and generalization. However, they may be restricted in practive. Many recent studies try to remove these assumptions, in both algorithmic convergence (Nguyen et al., 2019; Li et al., 2024) and stability analysis (Lei & Ying, 2020a). It would be an interesting direction to further enhance our work in the future.

## Acknowledgement

This work is supported by STI 2030—Major Projects (No. 2021ZD0201405), the National Natural Science Foundation of China under grants 12171178. Dr Tao's research is partially supported by NTU RSR and Start Up Grants.

## Impact Statement

SWA can improve the generalization ability in many tasks and has been widely used in practical applications. Its generalization research can be further combined with the landscape of loss function, which is beneficial to better understand the model selection.

## References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Bertsekas, D. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129 (2):163–195, October 2011. ISSN 0025-5610. doi: 10.1007/s10107-011-0472-0.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Bottou, L. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pp. 2624–2633, 2009.

Bottou, L. Stochastic gradient descent tricks. *Springer Berlin Heidelberg*, 2012.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Charles, Z. and Papailiopoulos, D. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pp. 745–754. PMLR, 2018.

Das, A., Schölkopf, B., and Muehlebach, M. Sampling without replacement leads to faster rates in finite-sum

minimax optimization. *Advances in Neural Information Processing Systems*, 35:6749–6762, 2022.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Devroye, L. and Wagner, T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. Convergence rate of incremental gradient and incremental newton methods. *SIAM Journal on Optimization*, 29 (4):2542–2565, 2019. ISSN 1052-6234. doi: https://doi.org/10.1137/17M1147846. Publisher Copyright: © 2019 Society for Industrial and Applied Mathematics.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

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

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Koltchinskii, V. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Koltchinskii, V. and Panchenko, D. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pp. 443–457. Springer, 2000.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020a.

Lei, Y. and Ying, Y. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2020b.

Li, H., Rakhlin, A., and Jadbabaie, A. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2024.

Li, T., Huang, Z., Tao, Q., Wu, Y., and Huang, X. Trainable weight averaging: Efficient training by optimizing historical solutions. In *The Eleventh International Conference on Learning Representations*, 2022.

Lu, P., Kobyzev, I., Rezagholizadeh, M., Rashid, A., Ghodsi, A., and Langlais, P. Improving generalization of pre-trained language models via stochastic weight averaging. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4948–4954, 2022.

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

McAllester, D. A. Some pac-bayesian theorems. *Machine Learning*, 3(37):355–363, 1999b.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

Nagaraj, D., Jain, P., and Netrapalli, P. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pp. 4703–4711. PMLR, 2019.

Neklyudov, K., Molchanov, D., Ashukha, A., and Vetrov, D. Variance networks: When expectation does not meet your expectations. In *International Conference on Learning Representations*, 2018.

Nesterov, Y. Introductory lectures on convex optimization. *Applied Optimization*, 87, 2004.

Nguyen, L. M., Nguyen, P. H., Richtárik, P., Scheinberg, K., Takáč, M., and van Dijk, M. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019.

Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and Van Dijk, M. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

Polyak, B. T. Introduction to optimization. 1987.

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

Rajput, S., Gupta, A., and Papailiopoulos, D. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pp. 7964–7973. PMLR, 2020.

Ruppert, D. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Shamir, O. Without-replacement sampling for stochastic gradient methods. *Advances in neural information processing systems*, 29, 2016.

Sherman, U., Koren, T., and Mansour, Y. Optimal rates for random order online optimization. *Advances in Neural Information Processing Systems*, 34:2097–2108, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sun, Y., Shen, L., and Tao, D. Which mode is better for federated learning? centralized or decentralized. *arXiv preprint arXiv:2310.03461*, 2023a.

Sun, Y., Shen, L., and Tao, D. Understanding how consistency works in federated learning via stage-wise relaxed initialization. *arXiv preprint arXiv:2306.05706*, 2023b.

Vapnik, V. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Xiao, J., Fan, Y., Sun, R., Wang, J., and Luo, Z.-Q. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35: 15446–15459, 2022.

Yang, Z., Lei, Y., Wang, P., Yang, T., and Ying, Y. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021.

Yuan, Z., Yan, Y., Jin, R., and Yang, T. Stagewise training accelerates convergence of testing error over sgd. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhou, P., Yuan, X., and Feng, J. New insight into hybrid stochastic gradient descent: Beyond with-replacement sampling and convexity. *Advances in Neural Information Processing Systems*, 31, 2018a.

Zhou, Y., Liang, Y., and Zhang, H. Generalization error bounds with probabilistic guarantee for sgd in nonconvex optimization. *arXiv preprint arXiv:1802.06903*, 2018b.

Zhu, M., Shen, L., Du, B., and Tao, D. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ziyin, L., Liu, K., Mori, T., and Ueda, M. Strength of minibatch noise in sgd. *arXiv preprint arXiv:2102.05375*, 2021.

# Supplementary Material for "Generalization Analysis of Stochastic Weight Averaging with General Sampling"

## A. Proof of Some Basic Properties

### A.1. Proof of Eq. (9)

Using the definition of $\bar{w}_T$ gives

$$\bar{w}_T = \frac{1}{T}\left(\sum_{t=1}^{T-1} w_t + w_T\right) = \frac{T-1}{T}\bar{w}_{T-1} + \frac{1}{T}w_T, \tag{26}$$

combining the update rule of gradient

$$w_T = \bar{w}_{T-1} - \frac{\alpha}{T-1}\sum_{i=1}^{T-1} i \cdot \nabla g(w_i, z_{i+1}), \tag{27}$$

we then have

$$\bar{w}_T = \bar{w}_{T-1} - \frac{\alpha}{T(T-1)}\sum_{i=1}^{T-1} i \cdot \nabla g(w_i, z_{i+1}). \tag{28}$$

### A.2. Proof of Lemma 3.6

**$(1+\alpha\beta)$-expansive.** According to triangle inequality and $\beta$-smoothness,

$$\begin{aligned}\|w_{T+1} - w'_{T+1}\| &\leq \|w_T - w'_T\| + \alpha\|\nabla g(w_T) - \nabla g(w'_T)\| \\ &\leq \|w_T - w'_T\| + \alpha\beta\|w_T - w'_T\| \\ &= (1+\alpha\beta)\|w_T - w'_T\|.\end{aligned} \tag{29}$$

***Non*-expansive.** Function is convexity and $\beta$-smoothness that implies

$$\langle \nabla g(w) - \nabla g(v), w - v\rangle \geq \frac{1}{\beta}\|\nabla g(w) - \nabla g(v)\|^2. \tag{30}$$

We conclude that

$$\begin{aligned}\|w_{T+1} - w'_{T+1}\| &= \sqrt{\|w_T - \alpha\nabla g(w_T) - w'_T + \alpha\nabla g(w'_T)\|^2} \\ &= \sqrt{\|w_T - w'_T\|^2 - 2\alpha\langle \nabla g(w_T) - \nabla g(w'_T), w_T - w'_T\rangle + \alpha^2\|\nabla g(w_T) - \nabla g(w'_T)\|^2} \\ &\leq \sqrt{\|w_T - w'_T\|^2 - \left(\frac{2\alpha}{\beta} - \alpha^2\right)\|\nabla g(w_T) - \nabla g(w'_T)\|^2} \\ &\leq \|w_T - w'_T\|.\end{aligned} \tag{31}$$

## B. Generalization Bound under Convexity

### B.1. Proof of Theorem 4.1

First, we consider that the different sample are selected to update with probability $\frac{1}{n}$ at the step $T+1$.

$$\begin{aligned}\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\ &\leq \|\frac{T}{T+1}(\bar{w}_T - \bar{w}'_T) + \frac{1}{T+1}\left(\bar{w}_T - \bar{w}'_T + \frac{\alpha}{T}\sum_{i=1}^{T} i\cdot(\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1}))\right)\| \\ &\leq \bar{\delta}_T + \frac{\alpha}{T+1}\left(\|\nabla g(w_T, z_{T+1}) - \nabla g'(w'_T, z'_{T+1})\| + \frac{1}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\ &\leq \bar{\delta}_T + \frac{2\alpha L}{T+1} + \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|,\end{aligned} \tag{32}$$

where the proof follows from the Eq. (9), triangle inequality, $\|\alpha\nabla g(w_T, z_{T+1})\| \leq \alpha L$ for step $T+1$ step. And $\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|$ will be controlled in the late.

Second, another situation need be considered in case of the same sample are selected to update with probability $1-\frac{1}{n}$ at the step $T+1$.

$$\begin{aligned}\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\ &\leq \bar{\delta}_T + \frac{\alpha}{T+1}\left(\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\| + \frac{1}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\ &\leq \bar{\delta}_T + \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|,\end{aligned} \tag{33}$$

where $\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\| = 0$ in the second inequality because the non-expansive property of convex function.

For each $\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|$ in the sense of expectation, We consider two situations using $\alpha L$ bound and the non-expansive property. Then

$$\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| = \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1}\frac{2Li}{n} = \frac{(T-1)\alpha L}{n(T+1)} \tag{34}$$

Then we obtain the expectation based on the above analysis

$$\begin{aligned}\mathbb{E}\left[\bar{\delta}_{T+1}\right] &\leq (1-\frac{1}{n})\bar{\delta}_T + \frac{1}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T+1}\right) + \frac{(T-1)\alpha L}{n(T+1)} \\ &\leq \mathbb{E}\left[\bar{\delta}_T\right] + \frac{\alpha L}{n}\end{aligned} \tag{35}$$

recursively, we can get

$$\mathbb{E}\left[\bar{\delta}_T\right] \leq \frac{\alpha L T}{n}. \tag{36}$$

Plugging this back into Eq. (14), we obtain

$$\epsilon_{gen} = \epsilon_{stab} = \mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq \frac{\alpha L^2 T}{n}. \tag{37}$$

And we finish the proof.

### B.2. Proof of Theorem 4.2

First, we consider that the different sample are selected to update with probability $\frac{1}{n}$ at the step $T+1$.

$$\begin{aligned}
\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\
&\leq \|\frac{T}{T+1}\left(\bar{w}_T - \bar{w}'_T\right) + \frac{1}{T+1}\left(\bar{w}_T - \bar{w}'_T + \frac{\alpha}{T}\sum_{i=1}^{T} i\cdot\left(\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\right)\right)\| \\
&\leq \frac{T}{T+1}\|\bar{w}_T - \bar{w}'_T\| + \frac{1}{T+1}\left(\|\bar{w}_T - \bar{w}'_T\| + \alpha\|\nabla g(w_T, z_{T+1}) - \nabla g'(w'_T, z_{T+1})\|\right) \\
&\leq \bar{\delta}_T + \frac{2\alpha L}{T+1},
\end{aligned} \tag{38}$$

where the proof follows from the Eq. (9), triangle inequality, $\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| = 0$ for the previous $T$ step and the Lipschitz assumption.

Second, There are two situations that need be further considered in case of the same sample are selected to update at the step $T+1$ with probability $1-\frac{1}{n}$. Assuming that the same sample is selected for the first $T$ steps of updates with probability $\frac{n-T-1}{n}$, we get

$$\begin{aligned}
\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\
&\leq \frac{T}{T+1}\|\bar{w}_T - \bar{w}'_T\| + \frac{1}{T+1}\left(\|\bar{w}_T - \bar{w}'_T\| + \alpha\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\|\right) \\
&\leq \bar{\delta}_T,
\end{aligned} \tag{39}$$

where $\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| = 0$ in the first inequality because the non-expensive properties of convex function.

Besides, the case that the different samples have already occurred in the first $T$ steps with probability $\frac{T}{n}$, then the different samples are selected at step $i$ with probability $\frac{1}{T}$

$$\begin{aligned}
\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w'_{T+1}\| \\
&\leq \bar{\delta}_T + \frac{\alpha}{T+1}\left(\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\| + \frac{1}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&\leq \bar{\delta}_T + \frac{\alpha}{T(T+1)}\cdot\frac{i}{T}\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g'(w'_i, z'_{i+1})\|, \qquad i\in[1,\cdots T] \\
&\leq \bar{\delta}_T + \frac{2\alpha L}{T(T+1)}
\end{aligned} \tag{40}$$

where the triangle inequality, the non-expansion properties of convex function and the Lipschitz assumption are used in inequalities form 1 to 3, respectively.

Then we obtain the expectation consider the above analysis

$$\begin{aligned}
\mathbb{E}\left[\bar{\delta}_{T+1}\right] &\leq \frac{n-T-1}{n}\bar{\delta}_T + \frac{T}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T(T+1)}\right) + \frac{1}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T+1}\right) \\
&\leq \mathbb{E}\left[\bar{\delta}_T\right] + \frac{2\alpha L}{n}\cdot\frac{2}{T+1}
\end{aligned} \tag{41}$$

recursively, we can get

$$\mathbb{E}\left[\bar{\delta}_T\right] \leq \frac{2\alpha L}{n}\cdot\sum_{t=1}^{T}\frac{2}{t} \leq \frac{2\alpha L}{n}\cdot 2\ln T. \tag{42}$$

Plugging this back into Eq. (14), we obtain

$$\epsilon_{gen} = \epsilon_{stab} = \mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq \frac{2\alpha L^2}{n} \cdot 2\ln T. \tag{43}$$

And we finish the proof.

## C. Generalization Bound under Non-Convexity

### C.1. Proof of Eq. (20)

We consider that $S$ and $S'$ be two sample of size $n$ differing in only a single example. Let $\xi$ denote the event $\bar{\delta}_{t_0} = 0$. Let $z$ be an arbitrary example and consider the random variable $I$ assuming the index of the first time step using the different sample. then we have

$$\begin{aligned}\mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| &= P\{\xi\}\,\mathbb{E}[|g(\bar{w}_T; z) - g(\bar{w}'_T; z)||\xi] + P\{\xi^c\}\,E[|g(\bar{w}_T; z) - g(\bar{w}'_T; z)||\xi^c] \\ &\leq P\{I \geq t_0\} \cdot \mathbb{E}[|g(\bar{w}_T; z) - g(\bar{w}'_T; z)||\xi] + P\{I \leq t_0\} \cdot \sup_{w,x} g(w; x),\end{aligned} \tag{44}$$

where $\xi^c$ denotes the complement of $\xi$.

Note that Note that when $I \geq t_0$, then we must have that $\bar{\delta}_{t_0} = 0$, since the execution on $S$ and $S'$ is identical until step $t_0$. We can get $LE[\|\bar{w}_T - \bar{w}'_T\||\xi]$ combined the Lipschitz continuity of $g$. Furthermore, we know $P\{\xi^c\} = P\left\{\bar{\delta}_{t_0} = 0\right\} \leq P\{I \leq t_0\}$, for the random selection rule, we have

$$P\{I \leq t_0\} \leq \sum_{t=1}^{t_0} P\{I = t_0\} = \frac{t_0}{n}. \tag{45}$$

We can combine the above two parts and $g \in [0, 1]$ to derive the stated bound $L\mathbb{E}[\|\bar{w}_T - \bar{w}'_T\||\xi] + \frac{t_0}{n}$, which completes the proof.

### C.2. Proof of Theorem 5.1

First, we consider that the different sample are selected to update with probability $\frac{1}{n}$ at the step $T + 1$.

$$\begin{aligned}\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\ &\leq \|\frac{T}{T+1}(\bar{w}_T - \bar{w}'_T) + \frac{1}{T+1}\left(\bar{w}_T - \bar{w}'_T + \frac{\alpha}{T}\sum_{i=1}^{T} i \cdot (\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1}))\right)\| \\ &\leq \bar{\delta}_T + \frac{\alpha}{T+1}\left(\|\nabla g(w_T, z_{T+1}) - \nabla g'(w'_T, z'_{T+1})\| + \frac{1}{T}\sum_{i=1}^{T-1} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\ &\leq \bar{\delta}_T + \frac{2\alpha L}{T+1} + \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|,\end{aligned} \tag{46}$$

where the proof follows from the Eq. (9), triangle inequality, $\|\alpha\nabla g(w_T, z_{T+1})\| \leq \alpha L$ for step $T + 1$ step. And $\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|$ will be controlled in the late.

Second, another situation need be considered in case of the same sample are selected to update with probability $1 - \frac{1}{n}$ at the

step $T+1$.

$$\begin{aligned}
\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\
&\leq \bar{\delta}_T + \frac{\alpha}{T+1}\left(\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\| + \frac{1}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&\leq \bar{\delta}_T + \frac{\alpha}{T+1}\left(\beta\|w_T - w'_T\| + \frac{1}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&\leq \bar{\delta}_T + \frac{\alpha\beta}{T+1}\left(\|\bar{w}_T - \bar{w}'_T\| + \frac{1}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&+ \frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| \\
&\leq \left(1 + \frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{(1+\alpha\beta)\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|,
\end{aligned} \tag{47}$$

where $\|\nabla g(w_{T+1}) - \nabla g(w'_{T+1})\|$ is controlled by $(1+\alpha\beta)$-expansive property of non-convex function in the second inequality.

Now we build the bound of $\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|$.

$$\begin{aligned}
&\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\| \\
&= \frac{\alpha}{T(T+1)}(\|\nabla g(w_1, z_2) - \nabla g(w'_1, z_2)\| + 2\|\nabla g(w_2, z_3) - \nabla g(w'_2, z_3)\| \\
&+ \cdots + (T-1)\|\nabla g(w_{T-1}, z_T) - \nabla g(w'_{T-1}, z_T)\|)
\end{aligned} \tag{48}$$

For each $\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|$ in the sense of expectation, We consider two situations using $\alpha L$ bound with probability $\frac{1}{n}$ and the $(1+\alpha\beta)$-expansive property with probability $1-\frac{1}{n}$. Then

$$\begin{aligned}
&\frac{1}{T(T+1)}(\frac{2\alpha L}{n} + (1-\frac{1}{n})\alpha\beta\|w_1 - w'_1\| + 2\left(\frac{2\alpha L}{n} + (1-\frac{1}{n})\alpha\beta\|w_2 - w'_2\|\right) \\
&+ \cdots + (T-1)\left(\frac{2\alpha L}{n} + (1-\frac{1}{n})\alpha\beta\|w_{T-1} - w'_{T-1}\|\right)) \\
&\leq \frac{\alpha L}{n} + \frac{\alpha\beta}{T(T+1)}(\|w_1 - w'_1\| + 2\|w_2 - w'_2\| + \cdots + (T-1)\|w_{T-1} - w'_{T-1}\|)
\end{aligned} \tag{49}$$

Next, we establish the bounds for each $\|w_i - w'_i\|$ based on gradient update rules using $\alpha L$ bound and the $(1+\alpha\beta)$-expansive property. And we have recursively

$$\begin{aligned}
\|w_i - w'_i\| &\leq (1-\frac{1}{n})(1+\alpha\beta)\|w_{i-1} - w'_{i-1}\| + \frac{1}{n}(2\alpha L + \|w_{i-1} - w'_{i-1}\|) \\
&\leq (1+\alpha\beta)\|w_{i-1} - w'_{i-1}\| + \frac{2\alpha L}{n}.
\end{aligned} \tag{50}$$

Then, combined with the above recursive relationship

$$\begin{aligned}
&\frac{\alpha\beta}{T(T+1)}(\|w_1-w_1'\|+2\|w_2-w_2'\|+\cdots+(T-1)\|w_{T-1}-w_{T-1}'\|)\\
&\leq \frac{\alpha\beta}{T(T+1)}(\frac{2\alpha L}{n}\cdot\frac{T(T-1)}{2}+\frac{2\alpha L}{n}\cdot\frac{T(T-1)}{2}\cdot\sum_{i=1}^{T-2}(1+\alpha\beta)^i)\\
&\leq \frac{\alpha^2\beta L}{n}(1-\frac{(1+\alpha\beta)-(1+\alpha\beta)^{T-1}}{\alpha\beta})\\
&\leq \frac{\alpha L}{n}((1+\alpha\beta)^{T-1}-1)
\end{aligned} \tag{51}$$

finally, we finished the task

$$\frac{\alpha}{T(T+1)}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1})-\nabla g(w_i',z_{i+1})\|\leq\frac{\alpha L}{n}(1+\alpha\beta)^{T-1}. \tag{52}$$

Then we obtain the expectation consider the above analysis

$$\begin{aligned}
\mathbb{E}\left[\bar{\delta}_{T+1}\right]&\leq(1-\frac{1}{n})\left(\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T+\frac{\alpha L}{n}(1+\alpha\beta)^T\right)+\frac{1}{n}\left(\bar{\delta}_T+\frac{2\alpha L}{T+1}+\frac{\alpha L}{n}(1+\alpha\beta)^{T-1}\right)\\
&\leq\left(\frac{1}{n}+(1-\frac{1}{n})\left(1+\frac{\alpha\beta}{T+1}\right)\right)\bar{\delta}_T+\frac{2\alpha L}{n(T+1)}+\frac{2\alpha L}{n}(1+\alpha\beta)^T
\end{aligned} \tag{53}$$

let $\alpha=\frac{c}{t}$, then

$$\begin{aligned}
&=\left(1+(1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t+\frac{2cL}{nt(t+1)}+\frac{2cL}{tn}(1+\frac{c\beta}{t})^t\\
&\leq\exp\left((1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t+\frac{2cL}{n}\cdot\frac{1+e^{c\beta}}{t}.
\end{aligned} \tag{54}$$

Here we used that $\lim_{x\to\infty}(1+\frac{1}{x})^x=e$.

Using the fact that $\bar{\delta}_0=0$, we can unwind this recurrence relation from $T$ down to $t_0+1$.

$$\begin{aligned}
\mathbb{E}\bar{\delta}_t&\leq\sum_{t=t_0+1}^{T}\left(\prod_{k=t+1}^{T}\exp\left((1-\frac{1}{n})\frac{c\beta}{k(k-1)}\right)\right)\frac{2cL}{n}\cdot\frac{1+e^{c\beta}}{t}\\
&=\sum_{t=t_0+1}^{T}\exp\left((1-\frac{1}{n})c\beta\sum_{k=t+1}^{T}\frac{1}{k(k-1)}\right)\frac{2cL}{n}\cdot\frac{1+e^{c\beta}}{t}\\
&\leq\sum_{t=t_0+1}^{T}\exp\left((1-\frac{1}{n})c\beta\frac{T-t}{tT}\right)\frac{2cL}{n}\cdot\frac{1+e^{c\beta}}{t}\\
&\leq\sum_{t=t_0+1}^{T}\exp\left(\log(\frac{T}{t})\cdot\frac{(1-\frac{1}{n})c\beta}{2}\right)\frac{2cL}{n}\cdot\frac{1+e^{c\beta}}{t}\\
&\leq T^{\frac{(1-\frac{1}{n})c\beta}{2}}\cdot\sum_{t=t_0+1}^{T}\left(\frac{1}{t-1}\right)^{\frac{(1-\frac{1}{n})c\beta}{2}+1}\cdot\frac{2cL(1+e^{c\beta})}{n}\\
&\leq\frac{2}{(1-\frac{1}{n})c\beta}\cdot\frac{2cL(1+e^{c\beta})}{n}\cdot\left(\frac{T}{t_0}\right)^{\frac{(1-\frac{1}{n})c\beta}{2}}\\
&\leq\frac{4L(1+e^{c\beta})}{(n-1)\beta}\cdot\left(\frac{T}{t_0}\right)^{\frac{c\beta}{2}}
\end{aligned} \tag{55}$$

Plugging this back into Eq. (20), we obtain

$$\mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq \frac{t_0}{n} + \frac{4L^2(1+e^{c\beta})}{(n-1)\beta} \cdot \left(\frac{T}{t_0}\right)^{\frac{c\beta}{2}}. \tag{56}$$

By taking the extremum, we obtain the minimum

$$t_0 = \left(2cL^2(1+e^{c\beta})\right)^{\frac{2}{c\beta+2}} \cdot T^{\frac{c\beta}{c\beta+2}} \tag{57}$$

finally, this setting get

$$\epsilon_{gen} = \epsilon_{stab} = \mathbb{E}|g(\bar{w}_T; z) - g(\bar{w}'_T; z)| \leq \frac{1+\frac{2}{c\beta}}{n-1}\left(2cL^2(1+e^{c\beta})\right)^{\frac{2}{c\beta+2}} \cdot T^{\frac{c\beta}{c\beta+2}}. \tag{58}$$

to simplify, omitting constant factors that depend on $\beta$, c and L, we get

$$\epsilon_{stab} \lessapprox \frac{T^{\frac{\beta c}{\beta c+2}}}{n}. \tag{59}$$

And we finish the proof.

### C.3. Proof of Theorem 5.2

In the case of non-convexity, we consider that the different sample are selected to update with probability $\frac{1}{n}$ at the step $T+1$,

$$\bar{\delta}_{T+1} = \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \leq \bar{\delta}_T + \frac{2\alpha L}{T+1}. \tag{60}$$

Next, we consider the case that the same sample are selected to update with probability $1-\frac{1}{n}$ at the step $T+1$. Assuming that the same sample is selected for the first $T$ steps of updates and using $(1+\alpha\beta)$-expansive, we get

$$\begin{aligned}
\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w_{T+1}\| \\
&\leq \frac{T}{T+1}\bar{\delta}_T + \frac{1}{T+1}\left(\|\bar{w}_T - \bar{w}'_T\| + \frac{\alpha}{T}\sum_{i=1}^{T} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&\leq \bar{\delta}_T + \frac{\alpha}{T+1}\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\| \\
&\leq \bar{\delta}_T + \frac{\alpha\beta}{T+1}\|w_T - w'_T\| \\
&\leq \bar{\delta}_T + \frac{\alpha\beta}{T+1}\left(\|\bar{w}_T - \bar{w}'_T\| + \frac{1}{T}\sum_{i=1}^{T-1} i \cdot \|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&\leq \left(1 + \frac{\alpha\beta}{T+1}\right)\bar{\delta}_T.
\end{aligned} \tag{61}$$

There are different bounds because the locations of different samples occurred. We have shown two extreme cases

respectively. Assuming the different samples are selected at step $T$,

$$\begin{aligned}
\bar{\delta}_{T+1} &= \|\frac{T}{T+1}\bar{w}_T + \frac{1}{T+1}w_{T+1} - \frac{T}{T+1}\bar{w}'_T - \frac{1}{T+1}w'_{T+1}\| \\
&\leq \bar{\delta}_T + \frac{1}{T+1}\frac{T-1}{T}\alpha\|\nabla g(w_{T-1}, z_T) - \nabla g'(w'_{T-1}, z'_T)\| \\
&+ \frac{\alpha}{T+1}\|\nabla g(w_T, z_{T+1}) - \nabla g(w'_T, z_{T+1})\| \\
&\leq \bar{\delta}_T + \frac{2\alpha L(T-1)}{T(T+1)} + \frac{\alpha\beta}{T+1}\left(\|\bar{w}_T - \bar{w}'_T\| + \frac{\alpha}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i, z_{i+1}) - \nabla g(w'_i, z_{i+1})\|\right) \\
&\leq \left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{\alpha\beta}{T+1}\left(\frac{(T-1)\alpha}{T}\|\nabla g(w_{T-1}, z_T) - \nabla g'(w'_{T-1}, z'_T)\|\right) + \frac{2\alpha L(T-1)}{T(T+1)} \\
&\leq \left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{2\alpha^2\beta L(T-1)}{T(T+1)} + \frac{2\alpha L(T-1)}{T(T+1)}
\end{aligned} \tag{62}$$

The expectation will be obtained,

$$\begin{aligned}
\mathbb{E}\left[\bar{\delta}_{T+1}\right] &\leq \frac{n-T-1}{n}\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{1}{n}\left(\bar{\delta}_T + \frac{2\alpha L}{T+1}\right) \\
&+ \frac{T}{n}\left(\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T + \frac{1}{T}\left(\frac{2\alpha^2\beta L(T-1)}{T(T+1)} + \frac{2\alpha L(T-1)}{T(T+1)}\right)\right) \\
&\leq \left(\frac{1}{n} + (1-\frac{1}{n})\left(1+\frac{\alpha\beta}{T+1}\right)\right)\bar{\delta}_T + \frac{2\alpha L}{n}\left(\frac{2+\alpha\beta}{T+1}\right)
\end{aligned} \tag{63}$$

let $\alpha = \frac{c}{t}$, then

$$\begin{aligned}
&= \left(1 + (1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{2cL}{n}\cdot\frac{2t+c\beta}{t(t+1)} \\
&\leq \exp\left((1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{2cL}{n}\cdot\frac{2t+c\beta}{t(t+1)}.
\end{aligned} \tag{64}$$

Using the fact that $\bar{\delta}_0 = 0$, we can unwind this recurrence relation from $T$ down to $t_0 + 1$.

$$\begin{aligned}
\mathbb{E}\bar{\delta}_t &\leq \sum_{t=t_0+1}^{T}\left(\prod_{k=t+1}^{T}\exp\left((1-\frac{1}{n})\frac{c\beta}{k(k-1)}\right)\right)\frac{2cL}{n}\cdot\frac{2+c\beta}{t-1} \\
&= \sum_{t=t_0+1}^{T}\exp\left((1-\frac{1}{n})c\beta\sum_{k=t+1}^{T}\frac{1}{k(k-1)}\right)\frac{2cL}{n}\cdot\frac{2+c\beta}{t-1} \\
&\leq \sum_{t=t_0+1}^{T}\exp\left((1-\frac{1}{n})c\beta\frac{T-t}{tT}\right)\frac{2cL}{n}\cdot\frac{2+c\beta}{t-1} \\
&\leq \sum_{t=t_0+1}^{T}\exp\left(\log(\frac{T}{t})\cdot\frac{(1-\frac{1}{n})c\beta}{2}\right)\frac{2cL}{n}\cdot\frac{2+c\beta}{t-1} \\
&\leq T^{\frac{(1-\frac{1}{n})c\beta}{2}}\cdot\sum_{t=t_0+1}^{T}\left(\frac{1}{t-1}\right)^{\frac{(1-\frac{1}{n})c\beta}{2}+1}\cdot\frac{2cL(2+c\beta)}{n} \\
&\leq \frac{2}{(1-\frac{1}{n})c\beta}\cdot\frac{2cL(1+c\beta)}{n}\cdot\left(\frac{T}{t_0}\right)^{\frac{(1-\frac{1}{n})c\beta}{2}} \\
&\leq \frac{4L(2+c\beta)}{(n-1)\beta}\cdot\left(\frac{T}{t_0}\right)^{\frac{c\beta}{2}}
\end{aligned} \tag{65}$$

Plugging this back into Eq. (20), we obtain

$$\mathbb{E}|g(\bar{w}_T;z)-g(\bar{w}'_T;z)| \leq \frac{t_0}{n}+\frac{4L(2+c\beta)}{(n-1)\beta}\cdot\left(\frac{T}{t_0}\right)^{\frac{c\beta}{2}}. \tag{66}$$

By taking the extremum, we obtain the minimum

$$t_0=\left(2cL^2(2+c\beta)\right)^{\frac{2}{c\beta+2}}\cdot T^{\frac{c\beta}{c\beta+2}} \tag{67}$$

finally, this setting get

$$\epsilon_{gen}=\epsilon_{stab}=\mathbb{E}|g(\bar{w}_T;z)-g(\bar{w}'_T;z)| \leq \frac{1+\frac{2}{c\beta}}{n-1}\left(2cL^2(2+c\beta)\right)^{\frac{2}{c\beta+2}}\cdot T^{\frac{c\beta}{c\beta+2}}. \tag{68}$$

to simplify, omitting constant factors that depend on $\beta$, c and L, we get

$$\epsilon_{stab}\lessapprox \frac{T^{\frac{\beta c}{\beta c+2}}}{n}. \tag{69}$$

On the other side, assuming the different samples are selected at step 1,

$$\begin{aligned}
\bar{\delta}_{T+1} &\leq \bar{\delta}_T+\frac{1}{T+1}\left(\alpha\beta\|w_T-w'_T\|+\frac{\alpha}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1})-\nabla g(w'_i,z_{i+1})\|\right)\\
&\leq \bar{\delta}_T+\frac{\alpha}{T(T+1)}\left(\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1})-\nabla g(w'_i,z_{i+1})\|\right)\\
&+\frac{\alpha\beta}{T+1}\left(\|\bar{w}_T-\bar{w}'_T\|+\frac{\alpha}{T}\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1})-\nabla g(w'_i,z_{i+1})\|\right)\\
&\leq \left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T+\frac{1+\alpha\beta}{T(T+1)}\left(\alpha\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1})-\nabla g(w'_i,z_{i+1})\|\right)
\end{aligned} \tag{70}$$

we use the $\alpha L$ bounding the $\|\nabla g(w_0,z_1)-\nabla g'(w'_0,z'_1)\|$ at step 1 and $(1+\alpha\beta)$-expansive property for another $\|\nabla g(w_i,z_{i+1})-\nabla g(w'_i,z_{i+1})\|$.

$$\begin{aligned}
\alpha\sum_{i=1}^{T-1} i\cdot\|\nabla g(w_i,z_{i+1})-\nabla g(w'_i,z_{i+1})\| &\leq 2\alpha L\left(\sum_{k=1}^{T-1}\frac{(T-k)(T+k-1)}{2}(\alpha\beta)^k\right)\\
&= 2\alpha L\left(\sum_{k=1}^{T-1}\frac{T(T-1)-k(k-1)}{2}(\alpha\beta)^k\right)\\
&\leq 2\alpha L\frac{T(T-1)}{4}\sum_{k=1}^{T-1}(\alpha\beta)^k
\end{aligned} \tag{71}$$

then, return to inequality Eq. (70), we get

$$\bar{\delta}_{T+1}\leq\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T+\frac{\alpha L}{2}\sum_{k=1}^{T-1}(\alpha\beta)^k(1+\alpha\beta). \tag{72}$$

The expectation will be obtained,

$$\begin{aligned}
\mathbb{E}\left[\bar{\delta}_{T+1}\right] &\leq \frac{n-T-1}{n}\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T+\frac{1}{n}\left(\bar{\delta}_T+\frac{2\alpha L}{T+1}\right)\\
&+\frac{T}{n}\left(\left(1+\frac{\alpha\beta}{T+1}\right)\bar{\delta}_T+\frac{1}{T}\left(\frac{\alpha L}{2}\sum_{k=1}^{T-1}(\alpha\beta)^k(1+\alpha\beta)\right)\right)\\
&\leq \left(\frac{1}{n}+(1-\frac{1}{n})\left(1+\frac{\alpha\beta}{T+1}\right)\right)\bar{\delta}_T+\frac{2\alpha L}{n}\left(\frac{1}{T+1}+\frac{1}{4}\sum_{k=1}^{T-1}(\alpha\beta)^k(1+\alpha\beta)\right)
\end{aligned} \tag{73}$$

let $\alpha = \frac{c}{t}$, then

$$\begin{aligned}
\mathbb{E}\bar{\delta}_{t+1} &= \left(1 + (1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{2cL}{n}\left(\frac{1}{t(t+1)} + \frac{1}{4}\left(\sum_{k=1}^{T-1}\left(\frac{c\beta}{t}\right)^k + \sum_{k=1}^{T-1}\left(\frac{c\beta}{t}\right)^{k+1}\right)\right) \\
&\leq \exp\left((1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{2cL}{n}\left(\frac{1}{t(t+1)} + \frac{1}{4}\left(\frac{c\beta\left(1-\left(\frac{c\beta}{t}\right)^{T-1}\right)}{t(t-c\beta)} + \frac{c^2\beta^2\left(1-\left(\frac{c\beta}{t}\right)^{T-1}\right)}{t^2(t-c\beta)}\right)\right) \\
&\leq \exp\left((1-\frac{1}{n})\frac{c\beta}{t(t+1)}\right)\bar{\delta}_t + \frac{cL}{n}\cdot\frac{4+c\beta+c^2\beta^2}{t-c\beta}
\end{aligned} \tag{74}$$

Using the fact that $\bar{\delta}_0 = 0$, we can unwind this recurrence relation from $T$ down to $t_0 + 1$.

$$\begin{aligned}
\mathbb{E}\bar{\delta}_t &\leq \sum_{t=t_0+1}^{T}\left(\prod_{k=t+1}^{T}\exp\left((1-\frac{1}{n})\frac{c\beta}{k(k-1)}\right)\right)\frac{cL}{n}\cdot\frac{4+c\beta+c^2\beta^2}{t-c\beta} \\
&= \sum_{t=t_0+1}^{T}\exp\left((1-\frac{1}{n})c\beta\sum_{k=t+1}^{T}\frac{1}{k(k-1)}\right)\frac{cL}{n}\cdot\frac{4+c\beta+c^2\beta^2}{t-c\beta} \\
&\leq \sum_{t=t_0+1}^{T}\exp\left((1-\frac{1}{n})c\beta\frac{T-t}{tT}\right)\frac{cL}{n}\cdot\frac{4+c\beta+c^2\beta^2}{t-c\beta} \\
&\leq \sum_{t=t_0+1}^{T}\exp\left(\log(\frac{T}{t})\cdot\frac{(1-\frac{1}{n})c\beta}{2}\right)\frac{cL}{n}\cdot\frac{4+c\beta+c^2\beta^2}{t-c\beta} \\
&\leq T^{\frac{(1-\frac{1}{n})c\beta}{2}}\cdot\sum_{t=t_0+1}^{T}\left(\frac{1}{t-c\beta-1}\right)^{\frac{(1-\frac{1}{n})c\beta}{2}+1}\cdot\frac{cL(4+c\beta+c^2\beta^2)}{n} \\
&\leq \frac{2}{(1-\frac{1}{n})c\beta}\cdot\frac{cL(4+c\beta+c^2\beta^2)}{n}\cdot\left(\frac{T}{t_0-c\beta}\right)^{\frac{(1-\frac{1}{n})c\beta}{2}} \\
&\leq \frac{2L(4+c\beta+c^2\beta^2)}{(n-1)\beta}\cdot\left(\frac{T}{t_0-c\beta}\right)^{\frac{c\beta}{2}}
\end{aligned} \tag{75}$$

Plugging this back into Eq. (20), we obtain

$$\mathbb{E}|g(\bar{w}_T;z) - g(\bar{w}'_T;z)| \leq \frac{t_0}{n} + \frac{2L^2(4+c\beta+c^2\beta^2)}{(n-1)\beta}\cdot\left(\frac{T}{t_0}\right)^{\frac{c\beta}{2}}. \tag{76}$$

By taking the extremum, we obtain the minimum

$$t_0 = c\beta + \left(cL^2(4+c\beta+c^2\beta^2)\right)^{\frac{2}{c\beta+2}}\cdot T^{\frac{c\beta}{c\beta+2}} \tag{77}$$

finally, this setting get

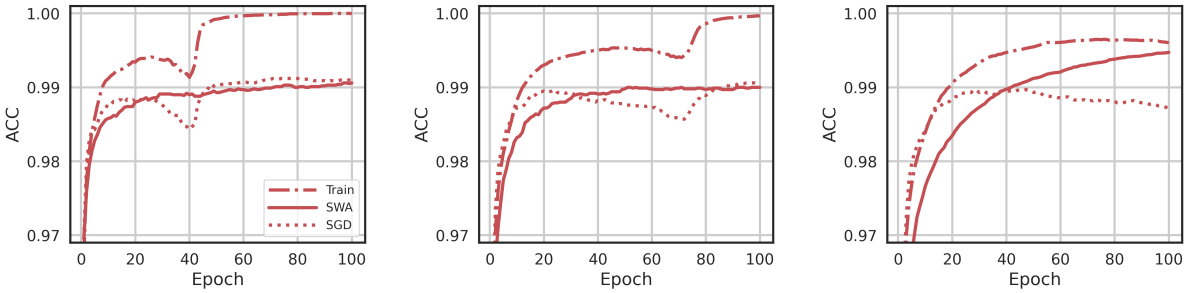
$$\epsilon_{gen} = \epsilon_{stab} = \mathbb{E}|g(\bar{w}_T;z) - g(\bar{w}'_T;z)| \leq \frac{1}{n-1}\left(c\beta + \left(\frac{1+c\beta}{c\beta}\right)cL^2(4+c\beta+c^2\beta^2)^{\frac{2}{c\beta+2}}\cdot T^{\frac{c\beta}{c\beta+2}}\right). \tag{78}$$

to simplify, omitting constant factors that depend on $\beta$, c and L, we get

$$\epsilon_{gen} \lessapprox \frac{T^{\frac{\beta c}{\beta c+2}}}{n}. \tag{79}$$

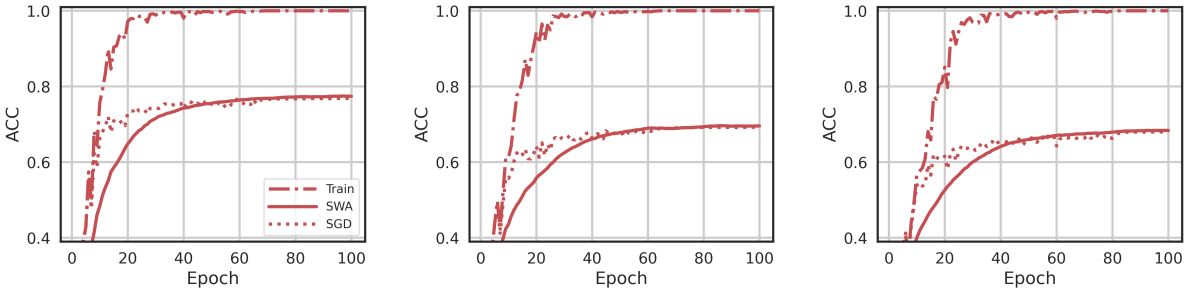
And we finish the proof.

$$\epsilon_{gen} \leq \frac{c\beta + 2\left(cL^2(4+c\beta+c^2\beta^2)\right)^{\frac{2}{c\beta+2}}\cdot T^{\frac{c\beta}{c\beta+2}}}{n-1} \tag{80}$$




(a) LR = 0.1 (b) LR = 0.05 (c) LR = 0.03

*Figure 5.* Train and test accuracy on the MNIST dataset, sample with replacement.



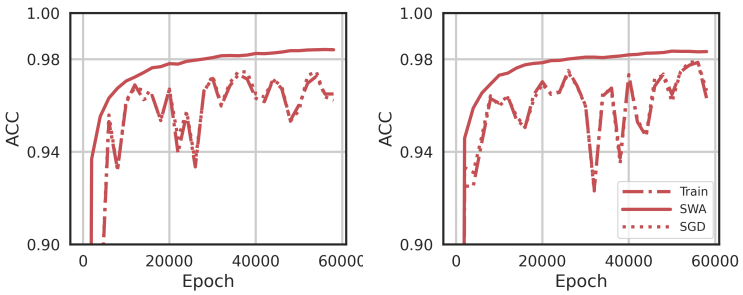

(a) LR = 0.1 (b) LR = 0.05 (c) LR = 0.03

*Figure 6.* Train and test accuracy on the CIFAR10 dataset, sample with replacement.

## D. Experiments

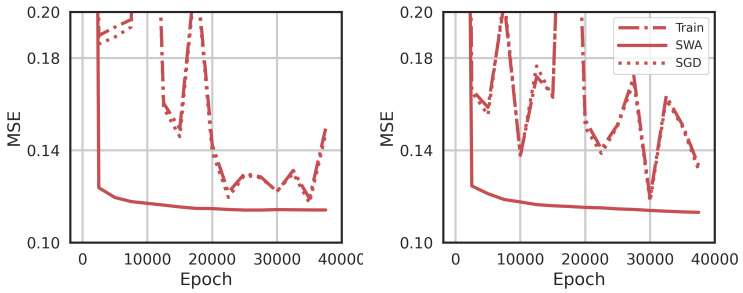
For experiments in Section 6.1, the results of train and test accuracy are shown in Figures 5 and 6.

For experiments in Section 6.2, the results of train and test accuracy are shown in Figures 7 and 8.




(a) With Replacement (b) Without Replacement

*Figure 7.* Train and test accuracy on the MNIST dataset.




(a) With Replacement (b) Without Replacement

*Figure 8.* Train and test accuracy on the Adult dataset.