GENERALIZATION BOUNDS OF STOCHASTIC GRADIENT DESCENT IN HOMOGENEOUS NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Algorithmic stability is among the most potent techniques in generalization analysis. However, its derivation usually requires a stepsize $\eta_t = \mathcal{O}(1/t)$ under non-convex training regimes, where t denotes iterations. This rigid decay of the stepsize potentially impedes optimization and may not align with practical scenarios. In this paper, we derive the generalization bounds under the homogeneous neural network regimes, proving that this regime enables a family of slower stepsize decay of order $\Omega(1/\sqrt{t})$ in expectation, under mild assumptions. We further extend the theoretical results from several aspects, *e.g.*, non-Lipschitz regimes. This finding is broadly applicable, as homogeneous neural networks encompass fully-connected and convolutional neural networks with ReLU and LeakyReLU activations.

1 INTRODUCTION

Stochastic gradient descent (SGD) has become a cornerstone in the field of deep learning. Extensive applications have demonstrated its remarkable success in both optimization and generalization (Goyal, 2017; Brown, 2020). On the theoretical side, researchers have proposed various approaches to address the challenge of generalization (McAllester, 1999; Russo & Zou, 2016; Bartlett et al., 2020). Among these, one of the most popular is uniform convergence (Bartlett et al., 2017; Wei & Ma, 2020), which, unfortunately, has proven inadequate for understanding generalization in deep learning (Shalev-Shwartz et al., 2010; Nagarajan & Kolter, 2019; Glasgow et al., 2023).

As an alternative to uniform convergence, algorithmic stability has emerged as a potential solution. In convex training scenarios, it provides guaranteed generalization for a reasonable number of training iterations (Bousquet & Elisseeff, 2002; Hardt et al., 2016). However, real-world applications typically occur in non-convex landscapes, where the generalization behavior of SGD remains elusive. Existing works on algorithmic stability under non-convex training regimes usually apply to the case with a stepsize of $\mathcal{O}(1/t)$, where t denotes the training iteration (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Zhang et al., 2022). Unfortunately, this stepsize is rarely used in practice due to its slow training speed (see Section 3.4 for theoretical insights and Figure 2 for illustration). This creates a gap between theory and practice, significantly reducing the practical utility of algorithmic stability.

To bridge this gap, there are three main approaches: (a) employing alternative algorithms such as Stochastic Gradient Langevin Dynamics (SGLD) instead of SGD (Mou et al., 2018; Li et al., 2020; Farghly & Rebeschini, 2021; Bassily et al., 2021); (b) utilizing Polyak-Lojasiewicz (PL) conditions, which are closely associated with convexity (Charles & Papailiopoulos, 2018; Nikolakakis et al., 2022; Zhu et al., 2024); (c) adopting alternative generalization metrics such as the gradient norm rather than the traditional generalization gap (Lei, 2023; Zhang et al., 2024). However, (a) is challenging to extend to SGD settings; (b) relies on a restrictive assumption that is difficult to relax; and (c) typically fails to generalize to multi-pass scenarios. Consequently, it remains under-explored how SGD effectively generalizes in multi-epoch regimes without convexity-related assumptions beyond $\mathcal{O}(1/t)$ stepsize.

In this paper, we take a step forward in bridging the gap by proving that SGD can generalize with stepsizes of order $\Omega(1/\sqrt{t})$. To the best of our knowledge, this is the first result that relaxes the stepsize requirement while allowing multiple uses of each sample under non-convex SGD training regimes without relying on additional assumptions like PL conditions. The derivation of this result relies on a notion called *homogeneous* neural networks (Du et al., 2018; Lyu & Li, 2020). Informally,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a neural network Φ is H-homogeneous if

$$\forall c > 0 : \Phi(c\boldsymbol{w}; X) = c^H \Phi(\boldsymbol{w}; X) \text{ for all } \boldsymbol{w} \text{ and } X,$$

where $\boldsymbol{w}$ denotes the weight and X denotes the input. Our paper proves that for homogeneous neural networks with a loss function that preserves homogeneity, if the loss is Lipschitz and approximately smooth, its generalization gap can be bounded over a reasonable number of training iterations.

1.1 Our Results in More Detail.

In particular, we consider a binary classification task with noisy labels $\{-1, 1\}$, leading to a non-zero Bayesian optimal within the sphere[1]. This condition is imposed to regulate the decay rate of the weight norm. To harness the homogeneity of H-homogeneous neural networks defined above, we focus on a loss function of the form $\ell(y, \hat{y}) = \max\{0, -y\hat{y}\}$. Despite the fact that the loss function might encourage a small prediction $\hat{y}$ and correspondingly small weights, we concentrate on the direction of the weights rather than their absolute values, which is valid in classification regimes. To simplify our discussion, we define all generalization metrics within the sphere, as shown in (3). Besides homogeneity, we anchor our analysis on two mild assumptions:

- **Bounded Loss** (Assumption 1) We assume that the individual loss is upper bounded, and the whole training loss is lower bounded by its Bayesian optimal value. This assumption is necessary because, without it, achieving a consistent generalization gap in the presence of noisy labels would be unattainable.
- **Lipschitz and Approximately Smooth** (Assumption 2) This assumption is standard in algorithmic stability analysis. We extend the previous smoothness assumption to approximate smoothness, which aligns better with ReLU activations.

We derive Theorem 5 based on the above assumptions, elucidating the correlation between homogeneous neural networks and generalization:

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] \leq [Lm_2]^{\frac{1}{m_1+1}} \left[1 + \frac{1}{m_1}\right] \frac{T^{\frac{m_1}{m_1+1}}}{n},$$

where T is the number of iterations, n is the sample size, and m_1, m_2 are problem-specific constants. We use $\mathcal{A}$ and $\mathcal{D}$ to represent the randomized algorithm and the distribution of the training dataset, respectively. A comprehensive summary of the notation used throughout this paper is provided in Appendix B.1. This bound is derived under stepsizes satisfying $\mathbb{E}\eta_t = \Omega(t^{\frac{H-4}{2}})$ and applies to the multi-pass training regime. For the specific case of $H = 3$, the required stepsize simplifies to $\mathbb{E}\eta_t = \Omega(1/\sqrt{t})$. Notably, these homogeneous neural networks can be constructed by normalizing multiple layers while leaving the final several layers unnormalized (see Proposition 2). As such, our results encompass a wide range of scenarios.

Extending the Reach of Our Findings. Our findings can be extended in several directions. *First*, we demonstrate the broad applicability of our framework. While our main result focuses on the loss function $\ell(y, \hat{y}) = \max\{0, -y\hat{y}\}$, Corollary 6 extends our findings to other loss functions, including smooth homogeneous variants, *e.g.*, $\ell(y, \hat{y}) = \max\{0, -y\hat{y}\}^u$ with $u \geq 1$, and functions with relaxed homogeneity. Similarly, we expand the scope of activation functions beyond ReLU, for instance $\sigma(\cdot) = \max\{0, \cdot\}^u$, and architectures beyond linear connections, such as normalized ResNet models, as described in Proposition 2. *Second*, for the specific case of three-homogeneous networks, Corollary 7 demonstrates that the test error consistently converges to the Bayesian optimal if the training loss approximates the Bayesian optimal within a reasonable number of iterations. *Third*, Theorem 8 shows a separation in optimization performance, indicating that our proposed stepsizes of $\mathbb{E}\eta_t = \Omega(1/\sqrt{t})$ achieve polynomial convergence, potentially outperforming the logarithmic convergence of $\mathbb{E}\eta_t = \mathcal{O}(1/t)$. Example 1 further illustrates the compatibility of our setting regarding generalization and optimization. *Finally*, Section 4 proves that our framework ensures generalization without Lipschitz assumptions via an adapted on-average model stability, while maintaining the applicability to the multi-pass training regime.

Proof Key Insights. Our key insights stem from a straightforward observation: For classification problems with homogeneous neural networks, accuracy depends solely on the direction of the weights.

[1]Here, the noisy label setting indicates the presence of irreducible noise, meaning the population loss of the Bayesian optimal predictor (i.e., the Bayesian optimal value) is strictly positive.

Table 1: Comparison of our method with related works on non-convex training regimes with SGD. A discussion of related works is deferred to Appendix A.1.

	Stepsize Order	PL Condition	Multi-pass	Overpara-meterization	Generalization Metric
Hardt et al. (2016)	$\mathcal{O}(1/t)$	No	Yes	Yes	Loss
Kuzborskij & Lampert (2018)	$\mathcal{O}(1/t)$	No	Yes	Yes	Loss
Zhang et al. (2022)	$\mathcal{O}(1/t)$	No	Yes	Yes	Loss
Charles & Papailiopoulos (2018)	$\mathcal{O}(1)$	Yes	Yes	Yes	Loss
Nikolakakis et al. (2022)	$\mathcal{O}(1)$	Yes	Yes	Yes	Loss
Zhu et al. (2024)	$\mathcal{O}(1)$	Yes	Yes	Yes	Loss
Lei & Tang (2021)	$o(1)$	No	Yes	No	Gradient
Lei (2023)	$\mathcal{O}(1)$	No	No	Yes	Gradient
Zhang et al. (2024)	$\mathcal{O}(1)$	No	No	Yes	Gradient
This Paper	$\Omega(1/\sqrt{t})$	No	Yes	Yes	Loss

To utilize this insight, we project the training trajectory onto a sphere. While the observed stepsize is in order $\Omega(1/\sqrt{t})$, the effective stepsize of the projected trajectory can be in order $\Theta(1/t)$. This leads to generalization, as demonstrated by algorithmic stability. One may question whether this effective stepsize $\Theta(1/t)$ negatively affects optimization. To address this concern, we provide additional discussions on optimization performance in Section 3.4 and present the training loss curve across different stepsize regimes in Figure 3 (left), which illustrates that the training loss can be optimized efficiently with the stepsize $\Omega(1/\sqrt{t})$.

2 Homogeneity

Distributions, Datasets and Algorithms. Let $(X, Y) \sim \mathcal{P} \in \mathbb{R}^d \times \mathbb{R}$ denote the feature-response pair, where $\mathcal{P}$ denotes the joint distribution. We consider a noisy classification problem where $Y \in \{-1, +1\}$ contains label noise. The model $f_{\boldsymbol{w}}(\cdot)$ is trained on an n-sample dataset $S = \{(X_i, Y_i)\}_{i\in[n]}$ where the samples are drawn i.i.d. from the distribution $\mathcal{P}$. Here $[n]$ denotes the set $\{1, 2, \cdots, n\}$. We denote $\mathcal{D}$ as the distribution of dataset S.

We consider a non-negative loss function $\ell(\boldsymbol{w}; X, Y)$ with the population loss $\mathcal{L}(\boldsymbol{w}) \triangleq \mathbb{E}_{(X,Y)\sim\mathcal{P}} \ell(\boldsymbol{w}; X, Y)$ and the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w}) \triangleq \frac{1}{n}\sum_{i\in[n]} \ell(\boldsymbol{w}; X_i, Y_i)$. We assume that the Bayesian optimal on the unit sphere is non-zero, namely $\underline{\sigma}^2 \triangleq \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}/\|\boldsymbol{w}\|) \neq 0$. This condition generally holds in scenarios with label noise.

The model is optimized via the SGD algorithm $\mathcal{A}$. Let $\boldsymbol{w}_t$ denote the trained weight at iteration t, and let $\boldsymbol{v}_t \triangleq \boldsymbol{w}_t/\|\boldsymbol{w}_t\|$ be its corresponding direction. Starting from an initialization $\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2 = \mathcal{O}(1)$, the weight update follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla_{\boldsymbol{w}_t} \ell_t(\boldsymbol{w}_t), \tag{1}$$

where η_t is the stepsize and ℓ_t is the loss function at iteration t. In this paper, we consider SGD with replacement, and therefore, $\ell_t(\cdot) \triangleq \ell(\cdot; X_i, Y_i)$ with probability $1/n$ for each index $i \in [n]$. Let $\mathbb{E}_I$ denote the expectation over the choice of the index i at the current iteration, given the parameters from all previous iterations. Let $\mathbb{E}_{\mathcal{D}}$ denote the expectation over the training dataset $S \sim \mathcal{D}$, and $\mathbb{E}_{\mathcal{A}}$ denote the expectation over the randomness of the algorithm $\mathcal{A}$, including the initialization and the sampling of indices across all previous iterations. Asymptotic notations are defined in Appendix B.1.

Homogeneity. Our results are based on the homogeneity of neural networks. Next, we provide a detailed definition of homogeneous neural networks and explain how they can be constructed.

Definition 1 (Homogeneous function and homogeneous neural network) *A function h is H-homogeneous ($H \geq 0$) if $h(c\boldsymbol{w}) = c^H h(\boldsymbol{w})$. Specifically, a neural network $\Phi(\boldsymbol{w}; X)$ is H-homogeneous ($H \geq 0$) if*

$$\Phi(c\boldsymbol{w}; X) = c^H \Phi(\boldsymbol{w}; X).$$

It's worth noting that the concept of homogeneity is evident across a range of neural network architectures. Specifically, neural networks employing ReLU, LeakyReLU, linear layers, and max-pooling layers inherently exhibit homogeneity. One could construct an H-homogeneous neural network based on Proposition 2, and we refer to Figure 1 in Appendix A.2 for an illustration.

Proposition 2 (Homogeneity Construction) *Consider two types of layers:*

1. ***Normalized layers***[2] $\mathcal{N}_{\boldsymbol{w}}(a) = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\sigma_n(a)$*, with arbitrary activation* σ_n*,* e.g.*, ReLU, sigmoid;*
2. ***Unnormalized layers*** $\mathcal{U}_{\boldsymbol{w}}(a) \triangleq \boldsymbol{w}\sigma_u(a)$*, where* σ_u *are ReLU, LeakyReLU, and Linear activations.*

For convenience, we omit the bias when the context is clear. The h*-layer neural network* $\Phi_H(\boldsymbol{w}; X)$ *with* $h - H$ *normalized layers* $\mathcal{N}_{\boldsymbol{w}}$ *and* H *unnormalized layers* $\mathcal{U}_{\boldsymbol{w}}$ *is* H*-homogeneous, where*

$$\Phi_H(\boldsymbol{w}; X) = \underbrace{\mathcal{U}_{\boldsymbol{w}_h} \circ \mathcal{U}_{\boldsymbol{w}_{h-1}} \circ \cdots \circ \mathcal{U}_{\boldsymbol{w}_{h-H+1}}}_{H \text{ unnormalized layers}} \circ \underbrace{\mathcal{N}_{\boldsymbol{w}_{h-H}} \circ \cdots \circ \mathcal{N}_{\boldsymbol{w}_1}(X)}_{h-H \text{ normalized layers}}.$$

Loss Function. Following the H-homogeneous neural network $\Phi(\boldsymbol{w}; X)$, the main result in Theorem 5 focuses on the loss function

$$\ell(\boldsymbol{w}; X, Y) = \max\{-Y\Phi(\boldsymbol{w}; X), 0\}. \tag{2}$$

Extensions to other loss functions are further discussed in the Remark of Theorem 5.

The Validity of the Loss Function. The loss function in (2) is valid because (a) the training direction depends only on the weight direction in classification problems with homogeneous neural networks, and (b) the generalization metric is defined on the unit sphere, meaning that the bound of the generalization gap does not necessarily converge to zero as $\|\boldsymbol{w}\|$ approaches zero. We choose this loss function because it harnesses the homogeneity of neural networks, namely, applying ℓ on an H-homogeneous neural network returns an H-homogeneous loss. Our theorems can be generalized to other loss functions that similarly harness homogeneity, *e.g.*, $(\max\{-Y\Phi(\boldsymbol{w}; X), 0\})^u$ with $u \geq 1$. To proceed, we derive Lemma 3 based on the homogeneity:

Lemma 3 *For any iteration* $t > 0$ *and a non-negative* H*-homogeneous function* ℓ_t *where* $H \neq 0$*, define the effective stepsize* $\tilde{\eta}_t$ *as the stepsize projected onto the sphere, namely, on* $\boldsymbol{v}_t = \boldsymbol{w}_t/\|\boldsymbol{w}_t\|$*. Then, the following properties hold:*

1. *Effective Stepsize on the sphere* [3]*:* $\tilde{\eta}_t = \eta_t \|\boldsymbol{w}_t\|^{H-2}$*;*
2. *Inner Product:* $\boldsymbol{w}_t^\top \nabla_{\boldsymbol{w}_t} \ell_t(\boldsymbol{w}_t) = H\ell_t(\boldsymbol{w}_t)$*;*
3. *Norm Iteration:* $\|\boldsymbol{w}_{t+1}\|^2 = \|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\tilde{\eta}_t\ell_t(\boldsymbol{v}_t))$.

Notably, the effective stepsize is normalized by the weight norm. This normalization enables the possibility of having a large stepsize while maintaining a small effective stepsize. This observation constitutes a key aspect of our analysis. We defer the proof of Lemma 3 to Appendix B.5.

In the context of homogeneity, the weight direction matters much more than its precise values. This phenomenon aligns with common practices in classification, where we mainly focus on the sign of the prediction instead of its precise value. Consequently, we evaluate the population loss and empirical loss on the sphere, as outlined in (3).

$$\mathcal{L}^s(\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w}/\|\boldsymbol{w}\|) = \mathcal{L}(\boldsymbol{v}), \quad \hat{\mathcal{L}}^s(\boldsymbol{w}) = \hat{\mathcal{L}}(\boldsymbol{w}/\|\boldsymbol{w}\|) = \hat{\mathcal{L}}(\boldsymbol{v}). \tag{3}$$

Algorithmic Stability. Algorithmic stability is a popular technique for bounding the generalization error by measuring an algorithm's sensitivity to perturbations in the training data (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016). This paper relies on this framework to motivate our analysis. We summarize the formal definitions of uniform stability (ϵ_T-stable) and the classic stability bounds for SGD in Appendix B.9.1.

[2]We only present linear connections here for simplicity but it allows for different structures, *e.g.*, ResNet connections $\mathcal{N}_{\boldsymbol{w}}(a) \triangleq a + \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\sigma_n(a)$ and max-pooling layers.

[3]The update rule formulated with the effective stepsize is given in Equation (17) within the proof of Lemma 16.

Why Algorithmic Stability Requires an $\mathcal{O}(1/t)$ Stepsize under Non-Convexity? The distinction between convex and non-convex optimization under algorithmic stability is fundamentally linked to the expansive property, as detailed in Lemma 3.6 in Hardt et al. (2016). In the context of convex training, if the same training sample is selected (with a probability of $1 - \frac{1}{n}$), the following inequality holds: $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}'_{t+1}\| \leq \|\boldsymbol{w}_t - \boldsymbol{w}'_t\|$, where the primed iterates denote the weights obtained from training on a perturbed dataset, as detailed in Appendix B.9.1. However, in the case of non-convex training, the inequality is modified to: $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}'_{t+1}\| \leq (1 + \beta\eta_t)\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|$. This modification implies that errors can accumulate over iterations in non-convex training, leading to an accumulation term $\prod_{t\in[T]}(1 + \beta\eta_t) \approx \exp(\beta \sum_{t\in[T]} \eta_t)$. To ensure that this term grows polynomially with the number of iterations T, we require that the sum of the stepsizes over the iterations is $\sum_{t\in[T]} \eta_t = \mathcal{O}(\log T)$, which in turn necessitates a stepsize of $\eta_t = \mathcal{O}(1/t)$.

3 Generalization Under Homogeneity

In this section, we present our main theorems, which establish the connection between homogeneity and generalization. We begin by introducing the necessary assumptions in Section 3.1. Next, we present our main generalization bound for general H-homogeneous neural networks in Section 3.2. We then specialize this result to the important case of three-homogeneous neural networks in Section 3.3. Finally, we compare the optimization performance in Section 3.4 to demonstrate the separation between different stepsize choices.

3.1 Assumptions

Before delving into the main theorem, we introduce the following assumptions concerning the loss function and the optimization process.

Assumption 1 (Loss Bound) *Assume that the individual loss is upper bounded by a constant $\bar{\sigma}^2$ on the unit sphere for any sample, namely*

$$\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \sup_{X,Y} \ell(\boldsymbol{v}; X, Y) \leq \bar{\sigma}^2.$$

Additionally, assume that the training loss is lower bounded by half of the Bayesian optimal during the training process, namely, for $t \in [0, T]$ with a given iteration T,

$$\inf_t \hat{\mathcal{L}}(\boldsymbol{v}_t) \geq \frac{1}{2}\underline{\sigma}^2.$$

Assumption 1 provides bounds for the training loss. The upper bound is standard in related analyses, ensuring that the training procedure does not result in excessively high loss values that could lead to an unstable optimization process. The lower bound is necessary to obtain a consistent generalization bound in noisy label settings, since without it, the generalization gap would not converge to zero, as demonstrated in Proposition 4. The proof is provided in Appendix B.2.

Proposition 4 *Given the Bayesian optimal $\underline{\sigma}^2$, for any given iteration t, if the training loss satisfies $\hat{\mathcal{L}}(\boldsymbol{v}_t) \leq \frac{1}{2}\underline{\sigma}^2$, the corresponding generalization gap satisfies $|\mathcal{L}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}_t)| \geq \frac{1}{2}\underline{\sigma}^2$.*

Assumption 2 (Lipschitz, Approximately Smooth) *Assume that the loss function $\ell(\cdot; X, Y)$ is L-Lipschitz on the sphere for any sample (X, Y), namely,*

$$\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \|\nabla_{\boldsymbol{v}} \ell(\boldsymbol{v}; X, Y)\| \leq L.$$

Additionally, we assume that the loss function is (γ, β)-approximately smooth on the sphere for any sample (X, Y), meaning that for each $\ell(\boldsymbol{v}; X, Y)$, there exists a globally β-smooth function $\bar{\ell}(\boldsymbol{v}; X, Y)$ such that for each $\boldsymbol{v}$

$$\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \|\nabla_{\boldsymbol{v}} \ell(\boldsymbol{v}; X, Y) - \nabla_{\boldsymbol{v}} \bar{\ell}(\boldsymbol{v}; X, Y)\| \leq \gamma.$$

The Assumption 2 introduces constraints on the loss function to ensure its Lipschitz and approximate smooth properties, which are widely used in related literature (Hardt et al., 2016). Note that

Assumption 2 degenerates into the standard smoothness assumption when $\gamma = 0$. We extend the previous smoothness assumption to approximate smoothness to accommodate the non-smooth behavior of ReLU activations. It is worth noting that there exist smooth activation functions that maintain homogeneity, *e.g.*, $\sigma(\cdot) = (\max\{\cdot, 0\})^u$ with $u > 1$, which leads to $\gamma = 0$.

3.2 Generalization Under Homogeneity

This section provides the generalization bounds for general homogeneous neural networks. We defer the complete proof to Appendix B.3.

Theorem 5 (Generalization Under Homogeneity) *Assume that the neural network $\Phi(\boldsymbol{w}; X)$ is H-homogeneous for any X, with $H > 2$. Given any fixed $T > 0$, let Assumption 1 hold with loss bounds $\underline{\sigma}^2$ and $\bar{\sigma}^2$, and Assumption 2 hold with L-Lipschitz and (γ, β)-approximately smooth loss, where $\gamma = o(T/n)$. Then, there exist infinitely many stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \Omega(t^{\frac{H-4}{2}})$, for which the generalization bound holds:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] \leq [Lm_2]^{\frac{1}{m_1+1}} \left[1 + \frac{1}{m_1}\right] \frac{T^{\frac{m_1}{m_1+1}}}{n},$$

where $m_1 = \frac{4c_2}{\underline{\sigma}^2}\left[\bar{\sigma}^2 + \frac{\beta}{H}\right]$ and $m_2 = \frac{8c_2\bar{\sigma}^{2m_1}}{H\underline{\sigma}^2}[L + \gamma n]$, constant $c_2 \geq 1$ related to the stepsize.

Remark: Validity for Smaller Stepsizes. While Theorem 5 specifies stepsizes satisfying $\mathbb{E}\eta_t = \Omega(t^{\frac{H-4}{2}})$ to achieve $\Theta(1/t)$ effective stepsizes, the generalization result remains valid for smaller stepsizes, such as $\Theta(t^{\frac{H-4}{2}})$. This is because a smaller stepsize induces a smaller effective stepsize, which naturally ensures generalization, even if it may lead to slower optimization.

Remark: Extensions of Loss Functions. The core derivation can be extended to other loss functions beyond $\ell_t(z) = \max\{-Y_t z, 0\}$. For a general loss form $\ell_t(\Phi_t(\boldsymbol{w}_t))$ with a homogeneous neural network $\Phi_t(\boldsymbol{w}_t)$, the effective stepsize $\tilde{\eta}_t$ is derived as

$$\tilde{\eta}_t = \rho_t \|\boldsymbol{w}_t\|^{H-2}\eta_t,$$

where $\rho_t \triangleq \ell_t'(\Phi_t(\boldsymbol{w}_t))/\ell_t'(\Phi_t(\boldsymbol{v}_t))$. Notably, the effective stepsize in Lemma 3 is a specific case with $\rho_t = 1$. For the hinge loss $\ell_t(z) = \max\{0, 1 - Y_t z\}$, we assume that the zero-loss sets $\{\boldsymbol{w}_t : \ell_t(\Phi_t(\boldsymbol{w}_t)) = 0\}$ and $\{\boldsymbol{v}_t : \ell_t(\Phi_t(\boldsymbol{v}_t)) = 0\}$ approximately overlap. This implies the ratio $\rho_t = \ell_t'(\Phi_t(\boldsymbol{w}_t))/\ell_t'(\Phi_t(\boldsymbol{v}_t)) = 1$ and the property is preserved whenever $\ell_t(\Phi_t(\boldsymbol{w}_t)) \neq 0$ and $\ell_t(\Phi_t(\boldsymbol{v}_t)) \neq 0$. We next provide in Corollary 6 a relaxed version of Theorem 5 regarding loss form. The proof is deferred to Appendix B.4.

Corollary 6 *Under the assumptions of Theorem 5, if the individual loss function $\ell(\cdot)$ satisfies $0 < \rho_t = \ell_t'(\Phi_t(\boldsymbol{w}_t))/\ell_t'(\Phi_t(\boldsymbol{v}_t)) \leq k_1$ with constant $k_1 \geq 1$, and $\ell_t(z) \leq z\ell_t'(z) \leq k_2\ell_t(z)$ with constant $k_2 > 1$ uniformly for all $t \in [T]$, the generalization bound in Theorem 5 remains valid with modified constants:*

$$m_1 = \frac{4c_2k_1}{\underline{\sigma}^2}\left[k_2\bar{\sigma}^2 + \frac{\beta}{H}\right], \quad m_2 = \frac{8c_2\bar{\sigma}^{2m_1}k_1}{H\underline{\sigma}^2}(L + \gamma n).$$

Notably, these conditions can be satisfied by many homogeneous loss functions, including $\ell_t(z) = (\max\{-Y_t z, 0\})^u$ with $u \geq 1$.

Comparison with T/n-type Bound. Under reasonable assumptions, one can construct T/n-type generalization bounds via algorithmic stability. The intuition is that for small iteration T, the difference between S and S' is selected with probability T/n (Hardt et al., 2016). However, this type of bound remains valid only when $T = o(n)$, indicating that not all samples are utilized during training. Clearly, this does not accurately reflect real-world scenarios. As for the results in Theorem 5, when L, H, β, $\bar{\sigma}^2$, and $\underline{\sigma}^2$ are all on a constant scale, the bound is approximately of order $\tilde{\gamma}^{\frac{1}{m_1+1}}[T/n]^{\frac{m_1}{m_1+1}}$, where $\tilde{\gamma} = \max\{\gamma, 1/n\}$. Notably, this new bound remains consistent when setting $T = o(n/\tilde{\gamma}^{1/m_1})$, allowing for multiple uses of each sample during the training process.

3.3 Generalization Under Three-Homogeneity

This section specializes the general bound in Theorem 5 to three-homogeneous neural networks (Corollary 7). We focus on this specific case for two main reasons: (a) setting $H = 3$ could yield the stepsize $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \Omega(1/\sqrt{t})$, a rate that commonly appears in theoretical analysis (Nemirovski et al., 2009); and (b) three-homogeneous neural networks themselves are widely adopted in practice, such as VGG19 (Simonyan & Zisserman, 2014). This alignment is somewhat surprising since the above two reasons are independent. We leave further discussion on this alignment for future work.

Corollary 7 *Consider a three-homogeneous neural network as constructed in Proposition 2 with $H = 3$. Let T_a denote the iteration at which the training error achieves a non-zero Bayesian optimal value[4], namely,*

$$T_a = \arg\min_t \left\{ \mathbb{E}_{\mathcal{A},\mathcal{D}} \hat{\mathcal{L}}(\boldsymbol{v}_t) \le \underline{\sigma}^2 \right\},$$

where we omit the dependency on stepsize and dataset for simplicity. Let Assumption 1 and Assumption 2 hold with (γ, β)-approximately smooth loss, where $\gamma = o(T_a/n)$. Then, there exist infinitely many stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \Omega(t^{-1/2})$, such that if $T_a = o(n/\tilde{\gamma}^{1/m_1})$, where $\tilde{\gamma} = \max\{\gamma, 1/n\}$ and m_1 denote a constant related to loss bound and smoothness, it holds that

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{A},\mathcal{D}} \mathcal{L}^s(\boldsymbol{v}_{T_a}) = \underline{\sigma}^2.$$

Corollary 7 demonstrates that the generalization gap would be small within $o(n/\tilde{\gamma}^{1/m_1})$ iterations. Therefore, if the training loss achieves the Bayesian optimal within $o(n/\tilde{\gamma}^{1/m_1})$ iterations, the corresponding test loss would also achieve the Bayesian optimal value. Additionally, $o(n/\tilde{\gamma}^{1/m_1})$ enables multiple uses of each sample during the training given $\gamma = o(1)$, which is rarely achieved in previous works with a similar order of stepsize (Lei & Tang, 2021; Lei, 2023).

The Role of Normalization. Doubts might arise regarding whether the results in Corollary 7 suggest that a three-layer MLP suffices in real-world scenarios, potentially rendering deeper architectures obsolete. To address this, we emphasize the pivotal role that normalization layers play. The significance lies in two key aspects. Firstly, they introduce non-linearity, effectively reducing approximation errors. Secondly, increasing the depth of a network might facilitate the optimization process (Arora et al., 2018a), leading to faster attainment of Bayesian optimality in training error (smaller T_a).

Loss Curve. One can observe from Theorem 5 that the generalization gap remains within $o(1)$ for cases where $T = o(n/\tilde{\gamma}^{1/m_1})$. This observation implies that if T_a is relatively small, indicating a fast convergence of the training loss to its Bayesian optimal, it cannot sustain this decrease until the point where T reaches $\Theta(n/\tilde{\gamma}^{1/m_1})$. Otherwise, such a trend would result in a test error smaller than the Bayesian optimal, which is impossible. We refer to Figure 3 (right) for an illustration. This phenomenon is empirically observed in Wen et al. (2023).

3.4 Additional Discussions on Optimization Performance

This section demonstrates a separation in optimization performance between the stepsize in this paper and those in prior works. Following the setting $H = 3$ in Section 3.3, we prove that a stepsize of order $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \Omega(1/\sqrt{t})$ potentially optimizes better compared to $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \mathcal{O}(1/t)$. For the convenience of discussion, we consider the class of stepsizes that could eliminate the randomness of the corresponding effective stepsize. The result is provided in Theorem 8.

Theorem 8 (Separation Between Stepsize $\Omega(1/\sqrt{t})$ and $\mathcal{O}(1/t)$) *Under the settings of Theorem 5 with $H = 3$, assume that the individual loss function $\ell_i(\cdot) = \ell(\cdot; X_i, Y_i)$ on the unit sphere satisfies*

1. *Smoothness: $\|\nabla_{\boldsymbol{v}_1}\ell_i(\boldsymbol{v}_1) - \nabla_{\boldsymbol{v}_2}\ell_i(\boldsymbol{v}_2)\| \le \beta\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|$;*
2. *PL condition: $\frac{1}{2}\|\nabla_{\boldsymbol{v}}\ell_i(\boldsymbol{v})\|^2 \ge \mu(\ell_i(\boldsymbol{v}) - \min_{\boldsymbol{v}} \ell_i(\boldsymbol{v}))$;*
3. *Strong growth: $\max_{i\in[n]}\{\|\nabla_{\boldsymbol{v}}\ell_i(\boldsymbol{v})\|\} \le B\|\nabla_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v})\|$;*
4. *$\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) \ge (1+\alpha)/\alpha\hat{\mathcal{L}}(\boldsymbol{v}^*)$,*

[4]Here we allow for an $o(1)$ relaxation.

then for SGD with $\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2 = \mathcal{O}(1)$ and $\tilde{\eta}_t \leq \min\{\frac{1}{\beta B^2}, \frac{H\underline{\sigma}^2}{2L^2}\}$, if $\mu > 4B^2H^2\bar{\sigma}^2(1+\alpha)$, it holds that for some given constant $c > 0$,

- *For stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \Omega(1/\sqrt{t})$ used in Theorem 5, the convergence rate would be*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\hat{\mathcal{L}}(\boldsymbol{v}_T) - \min_{\boldsymbol{v}} \hat{\mathcal{L}}(\boldsymbol{v})] \leq \frac{\mathbb{E}_{\mathcal{A},\mathcal{D}}[\hat{\mathcal{L}}(\boldsymbol{v}_0) - \min_{\boldsymbol{v}} \hat{\mathcal{L}}(\boldsymbol{v})]}{T^{c\lambda}},$$

- *There exist infinitely many stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \mathcal{O}(1/t)$, leading to the convergence rate*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\hat{\mathcal{L}}(\boldsymbol{v}_T) - \min_{\boldsymbol{v}} \hat{\mathcal{L}}(\boldsymbol{v})] \leq \frac{\mathbb{E}_{\mathcal{A},\mathcal{D}}[\hat{\mathcal{L}}(\boldsymbol{v}_0) - \min_{\boldsymbol{v}} \hat{\mathcal{L}}(\boldsymbol{v})]}{[\log T]^{c\lambda}}.$$

where $\lambda = \frac{\mu}{B^2} - 4H^2\bar{\sigma}^2(1+\alpha)$.

Theorem 8 establishes a separation between the effects of different stepsizes during the optimization process. It asserts that there are certain training landscapes for which an optimization algorithm with stepsize $\Omega(1/\sqrt{t})$ achieves a significantly better convergence rate compared to one with stepsize $\mathcal{O}(1/t)$. The PL condition is used to ensure convergence, while the strong growth condition is utilized to enhance the convergence rate of SGD. Notably, this result pertains to optimization performance, distinguishing it from prior works that apply the PL condition to derive generalization results via algorithmic stability. We provide the proof in Appendix B.6.

Example 1 (Compatibility between Optimization and Generalization) *Consider the loss function that satisfies homogeneity with $H = 3$:*

$$\ell(\boldsymbol{v}) = (v_1^2 + 2v_2^2)^{3/2}.$$

This case illustrates that the optimization guarantees in Theorem 8 and the generalization requirements in Corollary 7 can be satisfied simultaneously. Notably, this does not fall into the Neural Tangent Kernel (NTK) regime. In the unit sphere, this loss satisfies the required regularity conditions with $\underline{\sigma}^2 = 1 \leq \ell(\boldsymbol{v}) \leq 2^{3/2} = \bar{\sigma}^2$, $L = 6\sqrt{2}$, $\gamma = 0$, $\beta = 12\sqrt{2}$, and $\mu \approx 19.7$. Without affecting asymptotic behavior, relax the definition of hitting time in Corollary 7 to

$$T_a(\tau(n)) = \min\{t : \mathbb{E}\hat{\mathcal{L}}(\boldsymbol{v}_t) \leq \underline{\sigma}^2 + \tau(n)\}$$

for some $\tau(n) \to 0$. Theorem 8 implies $\ell(\boldsymbol{v}_T) - \underline{\sigma}^2 = \mathcal{O}(T^{-c\mu})$, which yields $T_a \leq \tau(n)^{-1/c\mu}$ for a constant c. To satisfy the generalization requirement $T_a = o(n/\tilde{\gamma}^{1/m_1})$ with $\tilde{\gamma} = 1/n$, we choose

$$\tau(n) = n^{-c\mu(1+\epsilon/m_1)}$$

for some $\epsilon \in (0,1)$. This ensures $T_a \leq n^{1+\epsilon/m_1} = o(n^{1+1/m_1})$. This argument naturally extends to $\ell(\boldsymbol{v}) = (v_1^2 + 2\sum_{i=2}^{D} v_i^2)^{3/2}$.

4 Generalization Beyond Lipschitz

This section demonstrates that the stepsize regime in Theorem 5 ensures generalization even without the Lipschitz assumption. The analysis proceeds by first defining a stability metric and establishing its relationship with generalization in Section 4.1, subsequently deriving the non-Lipschitz generalization bound in Section 4.2.

4.1 Stability and Generalization

This section starts from the definition of the conditional on-average stability, adapted from the *on-average model stability* (Lei & Ying, 2020) in Appendix B.9.2.

Definition 9 (Conditional On-Average Stability) *Let $S = \{z_1, \ldots, z_n\}$ and $S^{(i)} = \{z_1, \ldots, z_i', \ldots, z_n\}$ be two datasets differing only at the i-th index, where $z_i \triangleq (X_i, Y_i)$. $\boldsymbol{v}_T$ and $\boldsymbol{v}_T^{(i)}$ are trained on S and $S^{(i)}$ over T iterations, respectively. For any given $t_0 \in [T]$, let $\mathcal{E}_{(i)}$ be the event that the*

index i is not selected during the first t_0 iterations. The algorithm $\mathcal{A}$ is said to be ϵ_{stab}-conditionally stable on average if:

$$\epsilon^2_{avg,T|t_0} \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{A},\mathcal{D}}\left[\|\boldsymbol{v}_T - \boldsymbol{v}_T^{(i)}\|^2 \mid \mathcal{E}_{(i)}\right] \leq \epsilon_{stab}.$$

The following Lemma 10 connects the generalization gap to the conditional on-average stability. We focus on the setting where $\gamma = 0$. Notably, the assumptions of homogeneity and β-smoothness can hold simultaneously, *e.g.*, with the loss $\ell_t(z) = (\max\{-Y_t z, 0\})^2$ and activation $\sigma(\cdot) = \max\{0, \cdot\}^2$.

Lemma 10 *Under Bounded Loss (Assumption 1) and global β-smoothness, for any $t_0 > 0$ and $\zeta > 0$, the generalization bound holds:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] \leq \left(\frac{t_0}{n} + \frac{\beta}{\zeta}\right)\bar{\sigma}^2 + \frac{\beta+\zeta}{2}\epsilon^2_{avg,T|t_0},$$

where ζ is a free parameter. The proof is deferred to Appendix B.8.1.

4.2 Generalization Beyond Lipschitz

This section derives the generalization bounds without the Lipschitz assumption.

Theorem 11 (Generalization Beyond Lipschitz) *Given any fixed $T > 0$, let Bounded Loss (Assumption 1) and global β-smoothness hold. Then, for an H-homogeneous network ($H > 2$) and the stepsize regime in Theorem 5, the generalization gap satisfies:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] = \mathcal{O}\left(n^{-\frac{m_1'+2}{m_1'+3}}\mathcal{K}_T^{\frac{1}{m_1'+3}}T^{\frac{m_1'}{m_1'+3}}\right),$$

where $\mathcal{K}_T = 1 + T/n$ and $m_1' = 8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2} + \frac{4c_2\beta}{H\underline{\sigma}^2}$.

Notably, this bound still supports multi-pass training. We defer the proof to Appendix B.8.2.

Remark: Comparison of Bounds. Theorem 11 removes the Lipschitz requirement in Assumption 2, better reflecting practical scenarios. Interestingly, despite this relaxation, it sometimes yields a faster growth in the iteration upper bound compared to that in Theorem 5. Specifically, it supports $T = o(n^{1+2/(m_1'+1)})$ while ensuring generalization, which surpasses $o(n^{1+1/m_1})$ when $4c_2\beta > \underline{\sigma}^2 H$, allowing for more training epochs. This improvement potentially stems from the different stability metrics utilized. It reveals that, in our setting, uniform stability may contain redundancy relative to the Assumptions 1 and 2. We leave the application of our stepsize regime to other stability metrics for further exploration.

5 Conclusion

In this paper, we establish generalization bounds for homogeneous neural networks. We prove in Theorem 5 that for H-homogeneous neural networks, there exist stepsizes of order $\Omega(t^{\frac{H-4}{2}})$ that ensure generalization via algorithmic stability. This is grounded in the observation that while the stepsize may be large, its effective stepsize on the unit sphere might be considerably small. Applying the main result to the specific case of $H = 3$, Corollary 7 establishes generalization for stepsizes of order $\Omega(1/\sqrt{t})$, relaxing the previous requirement of $\mathcal{O}(1/t)$. Our results also extend to a broader range of loss functions, activation functions, and architectures, as detailed in Corollary 6 and Proposition 2. Besides, we provide in Theorem 8 an optimization result on the separation between stepsize $\mathcal{O}(1/t)$ and stepsize $\Omega(1/\sqrt{t})$. Furthermore, we strengthen our framework in Section 4 by deriving non-Lipschitz generalization bounds, better aligning our theory with real-world scenarios. Overall, our findings highlight homogeneity as a helpful property that enhances algorithmic stability, yielding benefits in both generalization and optimization. Following convention in the line of work in algorithmic stability (Hardt et al., 2016; Charles & Papailiopoulos, 2018; Lei & Ying, 2020), we treat the characterization of the constants in the assumptions with respect to network structure as future work.

AI USE STATEMENT

Writing. We used a generative AI assistant to edit the manuscript for readability, reducing idiosyncratic phrasing that could otherwise impede clear communication.

Theory. We used a generative AI assistant to check the accuracy of proofs written by the authors and to help identify detail-level errors. All theoretical claims and proof steps remain the authors' own and were verified by the authors.

Experiments. Under the authors' control and supervision, a generative AI assistant was used to write code for the experiments in this paper.

We have reviewed all AI-assisted edits and verified that they do not alter any experimental results, theoretical claims, or data. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The theoretical results in this paper are stated with their full assumptions in the main text (Sections 1, 3, and 4), and complete proofs are provided in the Appendix. The illustrative empirical results on CIFAR-10 with ResNet-18 (Appendix A.3, Figure 2) are described qualitatively in the text.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 242–252. PMLR, 2019.

Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34:25033–25043, 2021.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. *Advances in neural information processing systems*, 35:22505–22517, 2022.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *International conference on machine learning*, pp. 244–253, 2018a.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *International conference on machine learning*, pp. 254–263, 2018b.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. *International Conference on Machine Learning*, pp. 393–403, 2021.

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Robert J. N. Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 10876–10889, 2021.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 2020a.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33: 4381–4391, 2020b.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *Conference on Learning Theory*, 125:610–626, 2020.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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

Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning. *arXiv preprint arXiv:2203.10036*, 2022.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, pp. 382–393, 2018.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 1675–1685. PMLR, 2019.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Conference on Uncertainty in Artificial Intelligence*, 2017.

Tyler Farghly and Patrick Rebeschini. Time-independent generalization bounds for sgld in non-convex settings. *Advances in Neural Information Processing Systems*, 34:19836–19846, 2021.

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29, 2016.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *Conference on Learning Theory*, 99:1270–1279, 2019.

Margalit Glasgow, Colin Wei, Mary Wootters, and Tengyu Ma. Max-margin works while large margin fails: Generalization without uniform convergence. *International Conference on Learning Representations*, 2023.

P Goyal. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935, 2020.

Mahdi Haghifam, Shay Moran, Daniel M. Roy, and Gintare Karolina Dziugaite. Understanding generalization via leave-one-out conditional mutual information. *IEEE International Symposium on Information Theory*, pp. 2487–2492, 2022.

Mahdi Haghifam, Borja Rodríguez Gálvez, Ragnar Thobaben, Mikael Skoglund, Daniel M. Roy, and Gintare Karolina Dziugaite. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. *International Conference on Algorithmic Learning Theory*, 201:663–706, 2023.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *International Conference on Machine Learning*, 48:1225–1234, 2016.

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

Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. *International Conference on Learning Representations*, 2021.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 2020b.

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:19605–19617, 2022.

Akash Kumar, Hanqi Zhang, Adish Singla, and Yuxin Chen. The teaching dimension of kernel perceptron. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 2071–2079. PMLR, 2021.

Daniel Kunin, Atsushi Yamamura, Chao Ma, and Surya Ganguli. The asymmetric maximum margin bias of quasi-homogeneous neural networks. *International Conference on Learning Representations*, 2023.

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

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. *The Thirty Sixth Annual Conference on Learning Theory*, pp. 191–227, 2023.

Yunwen Lei and Ke Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4505–4511, 2021.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. *International Conference on Machine Learning*, 119:5809–5819, 2020.

Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22(25):1–41, 2021a.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pp. 6175–6186. PMLR, 2021b.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *International Conference on Learning Representations*, 2020.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *International Conference on Learning Representations*, 2021.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

Andrew Lowy and Meisam Razaviyayn. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. *International Conference on Algorithmic Learning Theory*, pp. 986–1054, 2023.

Gábor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pp. 3524–3546. PMLR, 2022.

Gabor Lugosi and Eulalia Nualart. Convergence of continuous-time stochastic gradient descent with applications to linear deep neural networks. *arXiv preprint arXiv:2409.07401*, 2024.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *International Conference on Learning Representations*, 2020.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Yi-An Ma, Teodor Vanislavov Marinov, and Tong Zhang. Dimension independent generalization of dp-sgd for overparameterized smooth convex optimization. *arXiv preprint arXiv:2206.01836*, 2022.

David A. McAllester. Pac-bayesian model averaging. *Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. *Conference On Learning Theory*, 75:605–638, 2018.

Mor Shpigel Nacson, Suriya Gunasekar, Jason D. Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *International Conference on Machine Learning*, 97:4683–4692, 2019.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, pp. 11611–11622, 2019.

Nikolaos Nakis, Niels Raunkjær Holm, Andreas Lyhne Fiehn, and Morten Mørup. How low can you go? searching for the intrinsic dimensionality of complex networks using metric node embeddings. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *Advances in Neural Information Processing Systems*, pp. 11013–11023, 2019.

Arkadi Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

Alexandre Lemire Paquin, Brahim Chaib-draa, and Philippe Giguère. Stability analysis of stochastic gradient descent for homogeneous neural networks and linear classifiers. *Neural Networks*, 164: 382–394, 2023.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. *Conference on Learning Theory*, pp. 250–296, 2018a.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018b.

Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogeneous classifiers and the role of weight decay. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4243–4247, 2022.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. *International Conference on Artificial Intelligence and Statistics*, 51:1232–1240, 2016.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34:27422–27433, 2021.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

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

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Jiaye Teng, Jianhao Ma, and Yang Yuan. Towards understanding generalization via decomposing excess risk dynamics. *International Conference on Learning Representations*, 2022.

Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Stability and generalization for markov chain stochastic gradient methods. *Advances in Neural Information Processing Systems*, 35:37735–37748, 2022a.

Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(2):187–212, 2022b.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. *International Conference on Learning Representations*, 2020.

Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. *Advances in Neural Information Processing Systems*, pp. 9709–9721, 2019.

Kaiyue Wen, Jiaye Teng, and Jingzhao Zhang. Benign overfitting in classification: Provably counter label noise with larger models. *International Conference on Learning Representations*, 2023.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, pp. 2524–2533, 2017.

Jing Xu, Jiaye Teng, Yang Yuan, and Andrew Chi-Chih Yao. When do models generalize? a perspective from data-algorithm compatibility. *arXiv preprint arXiv:2202.06054*, 2022.

Ming Yang, Xiyuan Wei, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic compositional gradient descent algorithms. *International Conference on Machine Learning*, 2024.

Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. *International Conference on Artificial Intelligence and Statistics*, pp. 2026–2034, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Peiyuan Zhang, Jiaye Teng, and Jingzhao Zhang. Generalization lower bounds for GD and SGD in smooth stochastic convex optimization. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2025*, volume 258 of *Proceedings of Machine Learning Research*, pp. 334–342. PMLR, 2025.

Siqi Zhang, Yifan Hu, Liang Zhang, and Niao He. Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization. *International Conference on Artificial Intelligence and Statistics*, pp. 694–702, 2024.

Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of sgd: Tightness analysis and improved bounds. *Uncertainty in artificial intelligence*, pp. 2364–2373, 2022.

Bowei Zhu, Shaojie Li, and Yong Liu. Stability and sharper risk bounds with convergence rate $o(1/n^2)$. *arXiv preprint arXiv:2410.09766*, 2024.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *Conference on Learning Theory*, pp. 4633–4635, 2021.

Appendix

A Additional Discussions

This appendix provides a discussion of related works in Section A.1 on algorithmic stability, generalization analysis, homogeneous neural networks, and the justification of our loss form. Subsequently, Section A.2 offers a detailed explanation and visualization (Figure 1) of the homogeneous neural network construction introduced in Proposition 2. Finally, Section A.3 illustrates the effect of larger order of stepsize to practical optimization and generalization with Figure 2.

A.1 Related Works

Algorithmic Stability under Non-convex Training. To derive algorithmic stability in non-convex regimes with SGD, a typical requirement is to employ stepsize $\mathcal{O}(1/t)$ (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Zhang et al., 2022). To relax the stepsize requirement, one possible approach is to utilize alternative algorithms (*e.g.*, SGLD) (Mou et al., 2018; Li et al., 2020; Farghly & Rebeschini, 2021; Bassily et al., 2021). When restricted to SGD training with $\omega(1/t)$ stepsize, a line of work leverages the PL condition, which closely relates to convexity (Charles & Papailiopoulos, 2018; Nikolakakis et al., 2022; Zhu et al., 2024). Another line of work concentrates on alternative generalization metrics, such as the gradient norm, rather than the conventional generalization gap. However, they may face challenges when applied to multi-pass training regimes (Lei, 2023; Zhang

et al., 2024). In contrast, this paper introduces a novel approach parallel to the aforementioned two lines, based on the notion of *homogeneity*. Table 1 summarizes the comparison.

Algorithmic Stability. Algorithmic stability is a prominent generalization technique that may generalize to a broader range of loss functions and models in multi-pass settings (Bousquet & Elisseeff, 2002; Feldman & Vondrák, 2019; Bousquet et al., 2020; Bassily et al., 2020b; Teng et al., 2022; Yang et al., 2024). This argument shows that if the model and the training method are not overly sensitive to data perturbations, the generalization error can be effectively bounded. It is provably effective under Lipschitz, convex, and smooth regimes (Hardt et al., 2016; Asi et al., 2021; Zhang et al., 2025). There is a line of work focusing on relaxing the constraints of Lipschitz condition (Lei & Ying, 2020; Arora et al., 2022; Nikolakakis et al., 2022) and smoothness (Yang et al., 2021; Wang et al., 2022a; Lowy & Razaviyayn, 2023).

Scope of the Assumptions. In this paper, homogeneity is introduced to justify the boundedness, Lipschitzness, and approximate-smoothness assumptions on the sphere-projected loss (Assumptions 1 and 2), not to derive how their constants scale with network width or depth. This is consistent with the algorithmic-stability line of work our analysis belongs to: papers covering stochastic gradient descent (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Charles & Papailiopoulos, 2018; Lei & Ying, 2020; Bassily et al., 2020a; Teng et al., 2022), full-batch gradient descent (Charles & Papailiopoulos, 2018; Bassily et al., 2020a; Lowy & Razaviyayn, 2023), noisy gradient methods such as SGLD (Mou et al., 2018; Li et al., 2020), and minimax models such as GANs (Farnia & Ozdaglar, 2021; Lei et al., 2021b), all impose abstract assumptions of exactly this kind (bounded loss, Lipschitzness, smoothness, or the PL condition) without characterizing how the corresponding constants scale with the width or depth of a concrete network. Establishing such a scaling relation for homogeneous, normalized architectures is an interesting and largely open question, but it belongs to a different line of work (*e.g.*, feature-learning or NTK-style analyses that explicitly track width and depth) and is out of scope for the present paper; we leave it for future work.

Generalization in Stochastic Optimization. Generalization in stochastic optimization has been thoroughly studied (Shalev-Shwartz et al., 2010), across various scenarios such as one-pass SGD (Pillaud-Vivien et al., 2018a; Lugosi & Nualart, 2024), multi-pass SGD (Pillaud-Vivien et al., 2018b; Sekhari et al., 2021; Lei et al., 2021a), DPSGD (Bassily et al., 2019; Ma et al., 2022), and ERM solutions (Feldman, 2016; Aubin et al., 2020). Early research primarily focused on gradient descent in convex learning problems (Amir et al., 2021), where generalization can be bounded by uniform convergence (Bartlett et al., 2017; Nagarajan & Kolter, 2019; Wei & Ma, 2020; Glasgow et al., 2023). However, this approach struggles in high-dimensional spaces and can be inadequate (Shalev-Shwartz et al., 2010; Feldman, 2016). An alternative is online-to-batch conversion (Nemirovskij & Yudin, 1983), which mitigates the high-dimensional challenge and achieves minimax sample complexity, but it is limited to single-pass training, whereas in practice, longer training often results in better generalization (Hoffer et al., 2017). Recent works aim to close this gap by bounding generalization in multi-pass settings (Soudry et al., 2018; Li et al., 2021; Lyu et al., 2021; Sekhari et al., 2021), showing that gradient descent benefits from implicit bias. However, characterizing this implicit bias for more complex models remains an open problem. In addition to uniform convergence and implicit bias, various other approaches to generalization have been proposed, including PAC-Bayes (McAllester, 1999; Dziugaite & Roy, 2017; Haddouche et al., 2021; Lotfi et al., 2022), information-theoretic methods (Russo & Zou, 2016; Xu & Raginsky, 2017; Negrea et al., 2019; Haghifam et al., 2020; 2022; 2023; Lugosi & Neu, 2022), compression-based techniques (Arora et al., 2018b; Hsu et al., 2021), and benign overfitting (Bartlett et al., 2020; Zou et al., 2021; Koren et al., 2022; Xu et al., 2022; Wen et al., 2023). Notably, as one of the most fundamental issues in machine learning theory, generalization theory focuses on the mechanism of deep learning, distinguishing itself from practical approaches such as validation tricks (Zhang et al., 2021; Chatterjee & Zielinski, 2022). The NTK regime considered in Ji & Telgarsky (2020a) corresponds to $H = 2$, and requires a clearer elaboration of the neural network's architecture.

Homogeneous Neural Networks. The concept of homogeneous neural networks has been explored in the literature (Wei et al., 2019; Rangamani & Banburski-Fahey, 2022; Vardi et al., 2022). A line of work focuses on the implicit bias towards max-margin solutions in homogeneous neural networks (Nacson et al., 2019; Lyu & Li, 2020; Ji & Telgarsky, 2020b) and its variants (Kunin et al., 2023). Of particular relevance here is (Paquin et al., 2023), which investigates algorithmic stability for SGD within the context of homogeneous neural networks. Different from their analysis, this paper (a) aims

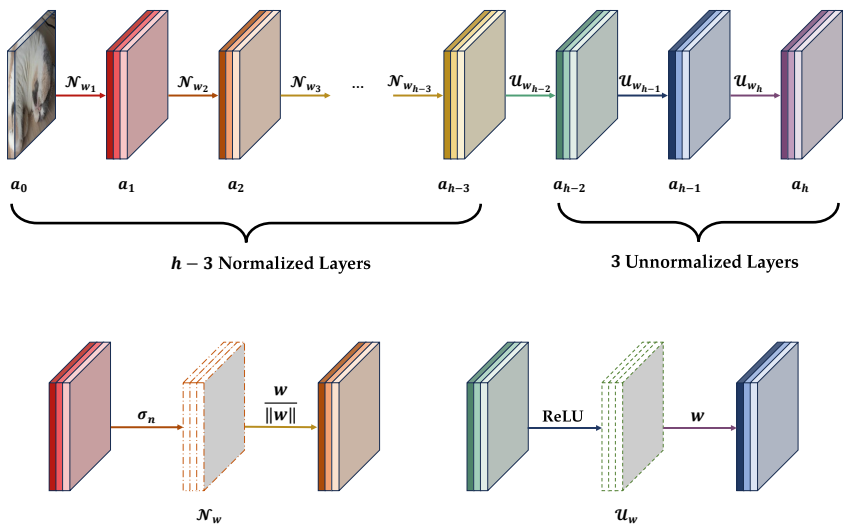


Figure 1: An instance of construction for $H = 3$. Following the rules in Proposition 2, normalized layers (left) accommodate arbitrary activations and connections with 0-homogeneity, while the subsequent unnormalized layers (right) compose a 3-layer MLP to achieve the desired three-homogeneity.

to improve stepsize requirements in stability analysis and (b) introduces the concept of approximate smoothness to extend the scope of previous regimes.

Loss Form. Constructing specific loss functions, despite their rare usage in industrial tasks, to theoretically isolate properties for better analysis, is a standard practice. To apply covering numbers for margin bounds, Bartlett et al. (2017) utilized the ramp loss $\ell_\gamma(z) = \max(0, \min(1, 1 - z/\gamma))$ instead of cross-entropy. In Neural Tangent Kernel (NTK) theory, L_2 loss is conventionally imposed upon classification tasks to ensure analytical solutions (Du et al., 2019; Allen-Zhu et al., 2019). Furthermore, loss of form $\ell(\boldsymbol{w}; X, Y) = \max\{-Y\Phi(\boldsymbol{w}; X), 0\}$ is well-established. To isolate output confidence interference, Baldock et al. (2021) used the zero-margin hinge loss $\ell = \sum_{j \neq y} \max(0, z_j - z_y)$ (similar to ours in binary settings) to train VGGs-architectures. Kumar et al. (2021) explicitly applied $\ell(f_\theta(x), y) = \max(0, -y f_\theta(x))$ to homogeneous polynomials. Wang et al. (2022b) formally defines it as the *Perceptron Loss*, and similar structures are adopted in graph representation learning (Nakis et al., 2025).

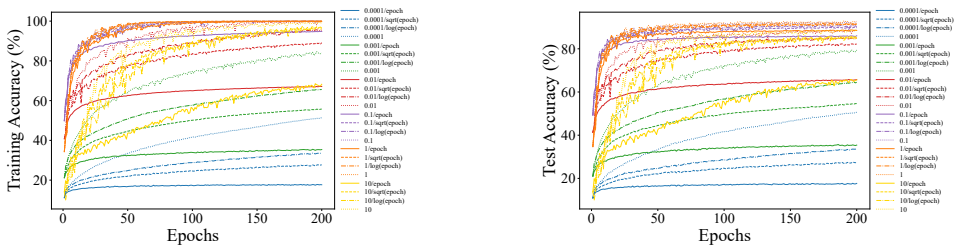
A.2 Detailed Construction and Visualization of H-Homogeneous Networks

First, we detail the underlying operations for the linear connections constructed in Proposition 2. For an unnormalized layer, we define $\mathcal{U}_{\boldsymbol{w}}(a) \triangleq \boldsymbol{w}\sigma_u(a)$. Here, a denotes the input vector (the output of the previous layer), $\boldsymbol{w}$ represents the weight matrix of the *current* layer, and $\sigma_u(\cdot)$ is an element-wise activation function. This operation proceeds by activating a followed by matrix multiplication with $\boldsymbol{w}$. In contrast, the normalized layer $\mathcal{N}_{\boldsymbol{w}}(a) \triangleq \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\sigma_n(a)$ incorporates an additional normalization step for $\boldsymbol{w}$ prior to multiplication. This can be implemented either by dividing $\boldsymbol{w}$ directly by its Frobenius norm or by normalizing each row independently. While both approaches ensure 0-homogeneity, they differ in signal propagation: the latter maintains the input's scale, whereas the former significantly attenuates the output scale relative to the layer's dimensions.

Next, We provide a visualization for the specific case with $H = 3$ to illustrate our construction. This case is of particular interest as it leads to the $\Omega(1/\sqrt{t})$ stepsize regime discussed in Corollary 7.

As shown in Figure 1, the architecture is decoupled into two segments:

- **Normalized:** The initial $h - 3$ layers utilize normalization $\mathcal{N}_{\boldsymbol{w}}$ to ensure 0-homogeneity, allowing for complex architectural features such as skip-connections or diverse activations without increasing the overall degree of the network.



(a) Training accuracy (b) Test Accuracy

Figure 2: The training and test accuracy curves for SGD with various stepsize schedulers on CIFAR-10 using ResNet 18. Models trained with stepsize $\Theta(1/\sqrt{t})$ (dotted line) converge faster and perform better than those trained with stepsize $\Theta(1/t)$ (solid line).

- **Unnormalized:** The final 3 layers are unnormalized $\mathcal{U}_{\boldsymbol{w}}$, which collectively contribute a degree of $H = 3$ to the network output.

To make this construction concrete on standard architectures, we instantiate it on VGG and ResNet. **VGG:** normalize every convolutional layer by its Frobenius norm (Section A.2 above), which makes each of them a 0-homogeneous normalized layer $\mathcal{N}_{\boldsymbol{w}}$, and leave the last 3 fully-connected ReLU layers unnormalized as $\mathcal{U}_{\boldsymbol{w}}$; the 3 unnormalized layers then yield the desired three-homogeneity. **ResNet:** a residual connection can itself serve as a normalized layer, $\mathcal{N}_{\boldsymbol{w}}(a) \triangleq a + \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\sigma_n(a)$, as already noted in the footnote to Proposition 2; normalizing all weights involved in the residual blocks and leaving the last 3 fully-connected layers unnormalized likewise gives $H = 3$.

Notably, unless otherwise specified, homogeneity in this paper refers to the entire constructed network, rather than the partial layers or activation functions.

A.3 Practical Effect of Stepsize

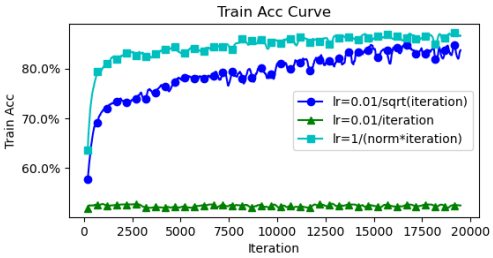
We analyze the effect of the SGD stepsize on practical optimization and generalization, as revealed in Figure 2, from two aspects: the initial coefficient and the decay rate.

A larger initial coefficient does not necessarily lead to better performance. When the initial coefficient is scaled from 0.0001 up to 1, the convergence speeds for both training and test accuracies remain comparable. However, increasing the coefficient further from 1 to 10 noticeably slows down convergence. Moreover, the training accuracy curves reveal that the configuration with a coefficient of 10 exhibits more oscillation, indicating that training becomes unstable at such a large stepsize magnitude.

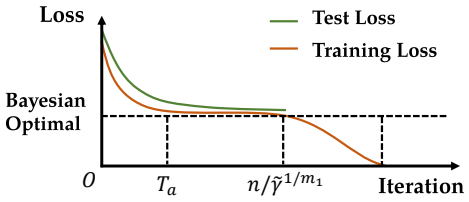
A slower decay rate does facilitate faster convergence. Across all coefficient groups, the overall trend in convergence speed is consistent: a constant stepsize $\Theta(1)$ is faster than $\Theta(1/\log(\text{epoch}))$, which in turn is faster than $\Theta(1/\sqrt{\text{epoch}})$ and $\Theta(1/\text{epoch})$. This empirical observation aligns with our theoretical contribution: relaxing the decay rate constraint on the stepsize is practically valuable, as slower-decaying stepsizes indeed yield faster convergence in real training scenarios.

B Theoretical Results

In this appendix, we organize the theoretical results as follows: Section B.1 details the notations used throughout the paper. Section B.2 provides the proof of Proposition 4. Section B.3 presents the proofs and further discussions of the main results in Theorem 5. Section B.4 provides the proof for the relaxation of the loss form as shown in Corollary 6. Section B.5 establishes the foundational analysis of the effective stepsize in Lemma 3. Section B.6 proves the optimization separation result between different stepsizes in Theorem 8. Section B.7 analyzes the weight norm iteration under both Lipschitz and non-Lipschitz conditions. Finally, Section B.8 provides the detailed proofs for the non-Lipschitz results discussed in Section 4.



(a) Normalization



(b) Loss curve illustration

Figure 3: (left) Training accuracy of SGD with different schedulers over a three-layer ReLU network. The initial learning rate is selected via grid search over $\{10^{-5}, 10^{-4}, \cdots, 10^{1}\}$. SGD with stepsize $\Theta(1/t\|\boldsymbol{w}_t\|)$ achieves similar training accuracy as SGD with stepsize $\Theta(1/\sqrt{t})$, both outperforming SGD with stepsize $\Theta(1/t)$; (right) When $T_a = o(n/\tilde{\gamma}^{1/m_1})$, the training and test loss curve does not decrease at $t \in (T_a, o(n/\tilde{\gamma}^{1/m_1}))$.

B.1 Notations

This paper uses the following standard notations for asymptotic analysis. For any positive functions $f(n)$ and $g(n)$, let $C > 0$ denote a constant independent of n:

- $f(n) = \Theta(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = C$;
- $f(n) = \mathcal{O}(g(n))$ if $\limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty$;
- $f(n) = \Omega(g(n))$ if $\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0$;
- $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$;
- $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$.

A comprehensive summary of the other notations used throughout this paper is provided in Table 2.

B.2 Proof of Proposition 4

Proof 1 (Proof of Proposition 4) *By definition, the population loss on the unit sphere is lower-bounded by the Bayesian optimal $\underline{\sigma}^2$. Thus, we have*

$$\mathcal{L}(\boldsymbol{v}_t) \geq \min_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{v}) = \underline{\sigma}^2.$$

Therefore, if the empirical loss satisfies $\hat{\mathcal{L}}(\boldsymbol{v}_t) \leq \frac{1}{2}\underline{\sigma}^2$, it immediately follows that the generalization gap is bounded by

$$|\mathcal{L}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}_t)| \geq \mathcal{L}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}_t) \geq \underline{\sigma}^2 - \frac{1}{2}\underline{\sigma}^2 = \frac{1}{2}\underline{\sigma}^2.$$

This completes the proof.

B.3 Proof of Theorem 5

This section derives the results in Theorem 5. The outline of our approach unfolds as follows:

First, we establish the relationship between the stepsize η_t and the effective stepsize $\tilde{\eta}_t$ on the unit sphere. Utilizing Lemma 3, we analyze how different orders of stepsizes induce varying orders of effective stepsizes under the homogeneous setting. Building on this, we strategically select the stepsize regime to ensure the effective stepsize remains sufficiently small to guarantee generalization. We refer to Lemma 12 for more details.

The remaining part requires deriving the algorithmic stability under such effective stepsize. Directly achieving stability proves elusive due to the presence of $\|\boldsymbol{w}_{t+1}\|$ and $\|\boldsymbol{w}_t\|$ within the updating

Table 2: Summary of Notations

Symbol	Description
Model	
$\Phi(\boldsymbol{w}; X)$	Neural network function $\Phi(\cdot;\cdot)$ with weights $\boldsymbol{w}$ and input X
$f_{\boldsymbol{w}}(\cdot)$	Equivalent notation to $\Phi(\boldsymbol{w};\cdot)$
H	Degree of positive homogeneity of the neural network Φ
$\boldsymbol{w}_t$	Weight vector at iteration t
$\boldsymbol{v}_t$	Direction of weight (normalized weight), defined as $\boldsymbol{v}_t = \boldsymbol{w}_t/\|\boldsymbol{w}_t\|$
Optimization	
S	Training dataset, a realization of $\mathcal{P}$
η_t	Stepsize (learning rate) in SGD updates
$\tilde{\eta}_t$	Effective stepsize governing dynamics on the unit sphere, $\tilde{\eta}_t = \rho_t\|\boldsymbol{w}_t\|^{H-2}\eta_t$
ρ_t	Gradient alignment ratio, defined as $\rho_t = \ell_t'(\Phi(\boldsymbol{w}_t))/\ell_t'(\Phi(\boldsymbol{v}_t))$
T	Total number of training iterations
n	Sample size of the training dataset
$[n]$	The set $\{1, 2, \cdots, n\}$
Loss and Generalization	
$\mathcal{P}$	Joint distribution of feature-response pair
$\mathcal{D}, \mathcal{A}$	Distribution of datasets and randomized algorithm
$\mathbb{E}$	Expectation of the random variable following it, taken over its subscript
$\ell(y, \hat{y})$	Individual loss function $\ell(\cdot,\cdot)$ with label y and prediction $\hat{y}$
$\ell(\boldsymbol{w}; X, Y)$	Individual loss function equivalent to $\ell(Y, \Phi(\boldsymbol{w}; X))$
$\ell_t(z)$	Equivalent to $\ell(Y_t, z)$ where z is the prediction
$\mathcal{L}(\boldsymbol{w})$	Population loss, $\mathcal{L}(\boldsymbol{w}) \triangleq \mathbb{E}_{(X,Y)\sim\mathcal{P}}\ell(\boldsymbol{w}; X, Y)$
$\hat{\mathcal{L}}(\boldsymbol{w})$	Empirical loss, $\hat{\mathcal{L}}(\boldsymbol{w}) \triangleq \frac{1}{n}\sum_{i\in[n]}\ell(\boldsymbol{w}; X_i, Y_i)$
$\mathcal{L}^s(\boldsymbol{w}), \hat{\mathcal{L}}^s(\boldsymbol{w})$	Population and empirical risks evaluated on the unit sphere, i.e., $\mathcal{L}^s(\boldsymbol{w}) \triangleq \mathcal{L}(\boldsymbol{v})$ and $\hat{\mathcal{L}}^s(\boldsymbol{w}) \triangleq \hat{\mathcal{L}}(\boldsymbol{v})$
Constants and Bounds	
$\bar{\sigma}^2$	Upper bound of the individual loss on the unit sphere
$\underline{\sigma}^2$	Bayesian optimal value of the given problem on the unit sphere, $\underline{\sigma}^2 \triangleq \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}/\|\boldsymbol{w}\|) \neq 0$
γ, β	Constants of (γ, β)-approximate smoothness for individual loss
$\tilde{\gamma}$	Additional rate multiplier of generalization bound, $\tilde{\gamma} = \max\{\gamma, 1/n\}$
L	Lipschitz constant of individual loss
m_1	Derived constant governing the generalization, $m_1 = \frac{4c_2k_1}{\underline{\sigma}^2}\left[k_1k_2\bar{\sigma}^2 + \frac{\beta}{H}\right]$, where $k_1 = k_2 = 1$ in Theorem 5
c_2	Constants specifying the stepsize

rule stipulated in (17). Consequently, bounding the ratio $\|\boldsymbol{w}_t\|/\|\boldsymbol{w}_{t+1}\|$ emerges as a requirement. This imperative stems from the upper bound imposed on the loss function by Assumption 1. We summarize the discussion in Lemma 13.

B.3.1 (EFFECTIVE) STEPSIZE

Lemma 12 ((Effective) stepsize) *Assume $\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2 = 1$ and $H > 2$ without loss of generality. Under Assumption 1 and Assumption 2:*

1. *there exist infinitely many stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \Omega(t^{\frac{H-4}{2}})$, such that the effective stepsize is of order $\Theta(1/t)$,*
2. *there exist infinitely many stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t \leq \frac{2c_2}{H\underline{\sigma}^2}(t+1)^{\frac{H-4}{2}} = \mathcal{O}(t^{\frac{H-4}{2}})$ with $c_2 \geq 1$, such that the effective stepsize is of order $\mathcal{O}(1/t)$.*
3. *there exist infinitely many stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = o(t^{\frac{H-4}{2}})$, such that the effective stepsize is of order $o(1/t)$.*

where the expectation is taken over the randomness of the dataset and the algorithm.

Lemma 12 is the core of our proof, which implies that while one may opt for a stepsize of the order $\Omega(t^{\frac{H-4}{2}})$, its manifestation as an effective stepsize upon projection onto a sphere can be of order $\Theta(1/t)$. This characteristic inherently holds the potential for establishing algorithmic stability. We remark that the existence arguments here are general since one could set $\tilde{\eta}_t = \Theta(1/t)$ with different constants. Hence, even in the context of existence arguments, there persists a degree of freedom attributed to constants.

Discussion on Stochasticity and Stepsize Regimes. We clarify the stochastic nature of the relationship $\tilde{\eta}_t = \eta_t\|\boldsymbol{w}_t\|^{H-2}$, which distinguishes our regime from standard SGD. In the SGD setting, the weight norm $\|\boldsymbol{w}_t\|$ is a random variable dependent on the sampling trajectory. Consequently, η_t and $\tilde{\eta}_t$ cannot simultaneously be deterministic unless $\|\boldsymbol{w}_t\|$ remains constant, which generally does not hold.

Three scenarios arise: (1) η_t is deterministic, rendering $\tilde{\eta}_t$ random; (2) $\tilde{\eta}_t$ is deterministic, which requires η_t to be a random variable; or (3) both are random variables. Our analysis adopts Scenario (2). Specifically, we impose a deterministic decay on the effective stepsize $\tilde{\eta}_t$ (*e.g.*, $\Theta(1/t)$), to ensure simple and stable update rule on the unit sphere.

Therefore, the *existence* established in our statements is twofold: it asserts the existence of random variables η_t capable of rendering the effective stepsize (a) deterministic, and (b) falling within the regime where it, as a non-random variable, satisfies the stated asymptotic bounds.

We further introduce the background of this approach. *First*, doing so is necessary; due to the presence of the expectation, we have to eliminate a portion of the randomness to make the analysis tractable. *Second*, although we select a fixed effective learning rate here, its parameter is variable, meaning there are infinitely many such stepsizes. In this context, our analysis is quite general, as illustrated in the footnote below. *Finally*, because the effective learning rate and the learning rate are generally positively correlated, the current analysis has the potential to be extended to the case of a fixed learning rate (where its expectation equals itself), which we leave as future work.

Concretely, Scenario (2) means η_t is *norm-adaptive*: one obtains it from a chosen effective-stepsize schedule $\tilde{\eta}_t$ by computing $\eta_t = \tilde{\eta}_t/\|\boldsymbol{w}_t\|^{H-2}$ at each iteration using the current weight norm, rather than following a fixed deterministic schedule such as $\eta_t = c/t$. This is analogous to norm-adaptive schedulers used in practice, *e.g.*, LARS (You et al., 2017), which scales the per-layer learning rate by the trust ratio $\|\boldsymbol{w}_l\|/\|\nabla\mathcal{L}(\boldsymbol{w}_l)\|$, and LAMB (You et al., 2019), which rescales the update by $\|\boldsymbol{w}_l\|/\|\boldsymbol{u}_l\|$ with $\boldsymbol{u}_l$ the Adam-style update; both adapt the learning rate to the current weight norm at each step, aligning with our stepsize regime. Extending the guarantee to a genuinely fixed deterministic schedule is left as future work (see the discussion above).

Discussion on $\log(T)$ convergence. We acknowledge the work by Soudry et al. (2018), which similarly explores a classification setting with direction convergence. Different from the $\sqrt{T}$ rate proposed in Lemma 12, the weight norm in Soudry et al. (2018) demonstrates an approximate $\log T$ growth. This distinction arises due to variations in loss forms, leading to non-conflicting conclusions.

It's worth noting that deriving an effective stepsize under their particular regime presents considerable complexity.

Proof 2 (Proof of Lemma 12) *We begin by proving the* first *statement of Lemma 12. By Lemma 21, if* $\tilde{\eta}_t \leq \frac{H\underline{\sigma}^2}{2L^2}$*, we have:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2 \leq \mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2\left(1-\frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t\right),$$

Therefore, WLOG[5] *setting* $\tilde{\eta}_t = \frac{2c_2}{H\underline{\sigma}^2(t+1)}$ *with* $c_2 \geq 1$ *leads to*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2 \leq \mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t-1}\|^2(1-1/(t+1)) \leq 1/(t+1)\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2 = 1/(t+1). \tag{4}$$

We derive the order of the stepsize given the choice of the effective stepsize. Notice that based on the order of the weight and the effective stepsize, the stepsize satisfies

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t &\overset{(i)}{=} \mathbb{E}_{\mathcal{A},\mathcal{D}}\frac{\tilde{\eta}_t}{\|\boldsymbol{w}_t\|^{H-2}}\\ &\overset{(ii)}{\geq} \frac{\tilde{\eta}_t}{[\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2]^{\frac{H-2}{2}}}\\ &\overset{(iii)}{\geq} \frac{2(t+1)^{(H-2)/2}}{H\underline{\sigma}^2(t+1)} = \frac{2}{H\underline{\sigma}^2}(t+1)^{(H-4)/2} = \Omega(t^{(H-4)/2}),\end{aligned}$$

where the equation (i) *is due to the definition of* $\tilde{\eta}_t$*, the inequality* (ii) *is from Jensen's inequality, and the inequality* (iii) *is based on* (4)*. Specifically, by taking* $f(x) = x^{-(H-2)/2}$ *which is convex when* $x > 0$ *and* $H > 2$*, we get from Jensen's inequality that* $\mathbb{E}f(x) \geq f(\mathbb{E}x)$*. Therefore, setting* $x = \|\boldsymbol{w}\|^2$ *leads to* $\mathbb{E}\|\boldsymbol{w}\|^{-(H-2)} \geq [\mathbb{E}\|\boldsymbol{w}\|^2]^{-(H-2)/2}$*.*

We next prove the second *statement of Lemma 12 by contradiction. In this case, it still holds for the weight norm that:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2 \leq \mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2\left(1-\frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t\right) \leq \mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2\exp\left(-\frac{1}{2}H\underline{\sigma}^2\sum_{k=0}^{t}\tilde{\eta}_k\right).$$

Suppose, for the sake of contradiction, that a stepsize satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t \leq \frac{2c_2}{H\underline{\sigma}^2}(t+1)^{\frac{H-4}{2}} = \mathcal{O}(t^{\frac{H-4}{2}})$ *could lead to a larger effective stepsize* $\tilde{\eta}_t > \frac{2c_2}{H\underline{\sigma}^2}\cdot\frac{1}{t+1}$*. Given this assumption on* $\tilde{\eta}_t$*, the summation in the exponent is lower bounded by:*

$$\sum_{k=0}^{t}\tilde{\eta}_k > \frac{2c_2}{H\underline{\sigma}^2}\sum_{k=0}^{t}\frac{1}{k+1} \geq \ln(k+1)|_0^{t+1} = \frac{2c_2}{H\underline{\sigma}^2}\ln(t+2).$$

Consequently, the weight norm would decay at a rate of $\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2 < \frac{1}{t+2} = \mathcal{O}(1/t)$*. This rapid decay implies that the stepsize satisfies:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_{t+1} > \frac{\frac{2c_2}{H\underline{\sigma}^2}\cdot\frac{1}{t+2}}{[1/(t+2)]^{(H-2)/2}} = \frac{2c_2}{H\underline{\sigma}^2}\cdot\frac{(t+2)^{(H-2)/2}}{t+2} = \frac{2c_2}{H\underline{\sigma}^2}\cdot(t+2)^{(H-4)/2}.$$

which contradicts the initial hypothesis that $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t \leq \frac{2c_2}{H\underline{\sigma}^2}(t+1)^{\frac{H-4}{2}}$*.*

Therefore, given the stepsize $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t \leq \frac{2c_2}{H\underline{\sigma}^2}(t+1)^{\frac{H-4}{2}}$*, the deterministic effective stepsize must satisfy* $\tilde{\eta}_t = \mathcal{O}(1/t)$*.*

[5]Setting $\tilde{\eta}_t = \frac{2c_2}{H\underline{\sigma}^2(t+1)}$ with $c_2 \geq 1$ may lead to the requirement that $2L \leq H\underline{\sigma}^2$ to satisfy the condition $\tilde{\eta}_t \leq \frac{H\underline{\sigma}^2}{2L^2}$ for all $t \geq 0$. However, this is not strictly necessary. One can always set $\tilde{\eta}_t = \frac{2c_2}{\underline{\sigma}^2(t+U)}$ for a sufficiently large U, which satisfies the requirement $\tilde{\eta}_t \leq \frac{H\underline{\sigma}^2}{2L^2}$ without influencing the asymptotic order. Similarly, when discussing the optimization in Theorem 8 and Lemma 20, we implicitly assume that the effective stepsize satisfies $\tilde{\eta}_t < \frac{1}{\beta B^2}$; the justification for this follows the exact same reasoning.

Finally, we prove the third *statement of Lemma 12. According to Lemma 18, for any given sampling trajectory, the weight norm satisfies:*

$$\|\boldsymbol{w}_{t+1}\|^2 = \|\boldsymbol{w}_t\|^2(1 - 2H\tilde{\eta}_t\ell(\boldsymbol{v}_t) + \tilde{\eta}_t^2\|\nabla\ell(\boldsymbol{v}_t)\|^2) \geq \|\boldsymbol{w}_t\|^2(1 - 2H\bar{\sigma}^2\tilde{\eta}_t).$$

Define an auxiliary sequence $r_t \triangleq t\|\boldsymbol{w}_t\|^2$. *Since* $\tilde{\eta}_t = o(1/t)$, *the iteration of* r_t *is governed by:*

$$\frac{r_{t+1}}{r_t} = \frac{t+1}{t} \cdot \frac{\|\boldsymbol{w}_{t+1}\|^2}{\|\boldsymbol{w}_t\|^2} \geq \left(1 + \frac{1}{t}\right)(1 - 2H\bar{\sigma}^2\tilde{\eta}_t) = 1 + \frac{1}{t} - 2H\bar{\sigma}^2\tilde{\eta}_t + o(\frac{1}{t^2}).$$

Then, there exists a t_0 *such that for all* $t > t_0$, *the term* $2H\bar{\sigma}^2\tilde{\eta}_t < \frac{1}{2t}$. *It holds that:*

$$\frac{r_{t+1}}{r_t} > 1 + \frac{1}{2t}. \tag{5}$$

Iterating (5) *from* t_0 *to* $T-1$, *we obtain:*

$$r_T \geq r_{t_0} \prod_{t=t_0}^{T-1}(1 + \frac{1}{2t}) \geq r_{t_0}\left(1 + \sum_{t=t_0}^{T-1}\frac{1}{2t}\right).$$

The divergence of $\sum \frac{1}{t}$ *implies* $\lim_{T\to\infty} r_T = \infty$, *which entails* $\|\boldsymbol{w}_T\|^2 = \omega(T^{-1})$. *Recall that* $\tilde{\eta}_t \triangleq \|\boldsymbol{w}_t\|^{H-2}\eta_t$ *from* (17). *Substituting the asymptotic order into this definition yields:*

$$\eta_t = \frac{\tilde{\eta}_t}{[\|\boldsymbol{w}_t\|^2]^{\frac{H-2}{2}}} = \frac{o(1/t)}{[\omega(1/t)]^{\frac{H-2}{2}}} = o(t^{\frac{H-4}{2}}).$$

Notably, this upper bound holds with t_0 *independent of sampling randomness. Taking the expectation yields* $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = o(t^{\frac{H-4}{2}})$. *This completes the proof.*

B.3.2 Generalization Gap

We next show in Lemma 13 that under the effective stepsize $\tilde{\eta}_t = \Theta(1/t)$, the generalization gap (or, algorithmic stability) can be bounded. Notably, the generalization gap is non-decreasing with respect to the effective stepsize sequence.

Lemma 13 (Generalization performance under effective stepsize) *Under the effective stepsize* $\tilde{\eta}_t = \frac{2c_2}{H\underline{\sigma}^2(t+2)}$ *with* $c_2 \geq 1$, *assuming that the function is* L*-Lipschitz and* γ*-*β *approximately smooth, it holds that*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\mathcal{L}(\boldsymbol{v}_t) - \mathcal{L}_n(\boldsymbol{v}_t) \leq (Lm_2)^{\frac{1}{m_1+1}}\left[1 + \frac{1}{m_1}\right]\frac{T^{\frac{m_1}{m_1+1}}}{n},$$

where $m_1 = \frac{4c_2}{\underline{\sigma}^2}[\bar{\sigma}^2 + \frac{\beta}{H}]$, *and* $m_2 = \frac{8c_2\bar{\sigma}^{2m_1}}{H\underline{\sigma}^2}(L + \gamma n)$.

Proof 3 (Proof of Lemma 13) *Our results are partly inspired by the results in Hardt et al. (2016). We focus on weight directions* $\boldsymbol{v}_t$ *and* $\boldsymbol{v}_t'$ *which are trained on dataset differing in at most one sample over* t *iterations. To bound the algorithmic stability, the core is to bound the difference between* $\boldsymbol{v}_{t+1}$ *and* $\boldsymbol{v}_{t+1}'$, *which is denoted by* $\Delta(\boldsymbol{v}_{t+1}, \boldsymbol{v}_{t+1}') = \|\boldsymbol{v}_{t+1} - \boldsymbol{v}_{t+1}'\|$. *To do so, we first denote the weight after one iteration by* $\bar{\boldsymbol{v}}_{t+1} = \boldsymbol{v}_t - \tilde{\eta}_t\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)$.

When choosing the same sample during the training,

$$\begin{aligned}
&\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}_{t+1}') \\
=&\|\bar{\boldsymbol{v}}_{t+1} - \bar{\boldsymbol{v}}_{t+1}'\| \\
=&\|(\boldsymbol{v}_t - \tilde{\eta}_t\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)) - (\boldsymbol{v}_t' - \tilde{\eta}_t\nabla_{\boldsymbol{v}_t'}\ell_t(\boldsymbol{v}_t'))\| \\
\leq&\|\boldsymbol{v}_t - \boldsymbol{v}_t'\| + \tilde{\eta}_t\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t'}\ell_t(\boldsymbol{v}_t')\|.
\end{aligned} \tag{6}$$

Note that the gradient difference can be bounded by Assumption 2 using a β*-smooth function* $\bar{\ell}$, *then we have that*

$$\begin{aligned}
&\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t'}\ell_t(\boldsymbol{v}_t')\| \\
=&\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t}\bar{\ell}_t(\boldsymbol{v}_t) + \nabla_{\boldsymbol{v}_t}\bar{\ell}_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t'}\bar{\ell}_t(\boldsymbol{v}_t') + \nabla_{\boldsymbol{v}_t'}\bar{\ell}_t(\boldsymbol{v}_t') - \nabla_{\boldsymbol{v}_t'}\ell_t(\boldsymbol{v}_t')\| \\
\leq&\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t}\bar{\ell}_t(\boldsymbol{v}_t)\| + \|\nabla_{\boldsymbol{v}_t}\bar{\ell}_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t'}\bar{\ell}_t(\boldsymbol{v}_t')\| + \|\nabla_{\boldsymbol{v}_t'}\bar{\ell}_t(\boldsymbol{v}_t') - \nabla_{\boldsymbol{v}_t'}\ell_t(\boldsymbol{v}_t')\| \\
\leq&2\gamma + \beta\|\boldsymbol{v}_t - \boldsymbol{v}_t'\|.
\end{aligned} \tag{7}$$

By plugging (6) *into* (7)*, we obtain*

$$\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \leq (1+\tilde{\eta}_t\beta)\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + 2\tilde{\eta}_t\gamma. \tag{8}$$

When choosing a different sample during the training, it similarly holds that

$$\begin{aligned}
&\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \\
=&\|\bar{\boldsymbol{v}}_{t+1} - \bar{\boldsymbol{v}}'_{t+1}\| \\
=&\|(\boldsymbol{v}_t - \tilde{\eta}_t\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)) - (\boldsymbol{v}'_t - \tilde{\eta}_t\nabla_{\boldsymbol{v}'_t}\ell_t(\boldsymbol{v}'_t))\| \\
\leq&\|\boldsymbol{v}_t - \boldsymbol{v}'_t\| + \tilde{\eta}_t\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}'_t}\ell_t(\boldsymbol{v}'_t)\| \\
\leq&\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + 2\tilde{\eta}_t L.
\end{aligned} \tag{9}$$

Due to the randomness of SGD, where with probability $1-1/n$ *the same sample is chosen and with probability* $1/n$ *a different sample is selected, based on* (8) *and* (9)*, it holds that*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \\
\leq&\mathbb{E}_{\mathcal{A},\mathcal{D}}(1-\frac{1}{n})\left[(1+\tilde{\eta}_t\beta)\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + 2\tilde{\eta}_t\gamma\right] + \frac{1}{n}[\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + 2\tilde{\eta}_t L] \\
=&\mathbb{E}_{\mathcal{A},\mathcal{D}}\left(1+\tilde{\eta}_t\beta(1-\frac{1}{n})\right)\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + \frac{2\tilde{\eta}_t L}{n} + (1-\frac{1}{n})2\tilde{\eta}_t\gamma \\
\leq&\mathbb{E}_{\mathcal{A},\mathcal{D}}(1+\tilde{\eta}_t\beta)\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + \frac{2\tilde{\eta}_t L}{n} + 2\tilde{\eta}_t\gamma.
\end{aligned}$$

By the results in Lemma 14 and Lemma 15, we conclude that

$$\Delta(\boldsymbol{v}_{t+1}, \boldsymbol{v}'_{t+1}) \leq \max\left\{\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|}, \frac{\|\boldsymbol{w}'_t\|}{\|\boldsymbol{w}'_{t+1}\|}\right\}\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \leq \left[1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right]\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}). \tag{10}$$

When $t \geq T_0 = 8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2} = \mathcal{O}(1)$*, taking expectations over the algorithms and dataset and plugging into* $\tilde{\eta}_t = \frac{2c_2}{H\underline{\sigma}^2(t+2)}$ *leads to*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\boldsymbol{v}_{t+1}, \boldsymbol{v}'_{t+1}) \\
\leq&\left[1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right]\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \\
\leq&\left[1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right]\left[(1+\tilde{\eta}_t\beta)\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + \frac{2\tilde{\eta}_t L}{n} + 2\tilde{\eta}_t\gamma\right] \\
=&\left[1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right]\left[(1+\frac{2c_2\beta}{H\underline{\sigma}^2(t+2)})\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + \frac{4c_2 L}{nH\underline{\sigma}^2(t+2)} + \frac{4c_2}{H\underline{\sigma}^2(t+2)}\gamma\right] \\
\leq&\left[1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2} + \frac{4c_2\beta}{H\underline{\sigma}^2(t+2)}\right]\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + \frac{8c_2 L}{nH\underline{\sigma}^2(t+2)} + \frac{8c_2\gamma}{H\underline{\sigma}^2(t+2)} \\
\leq&\exp\left(m_1\frac{1}{t}\right)\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\boldsymbol{v}_t, \boldsymbol{v}'_t) + \frac{m_2^0}{nt},
\end{aligned} \tag{11}$$

where $\tilde{\eta}_t \leq 1/\beta$*,* $m_1 = \frac{4c_2}{\underline{\sigma}^2}[\bar{\sigma}^2 + \frac{\beta}{H}]$*, and* $m_2^0 = \frac{8c_2}{H\underline{\sigma}^2}(L+\gamma n)$*.*

According to (3.10) in Hardt et al. (2016), for any given t_0*, generalization gap can be bounded by*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\mathcal{L}_T^s(\boldsymbol{w}) - \hat{\mathcal{L}}_T^s(\boldsymbol{w}) \leq \frac{t_0}{n}\bar{\sigma}^2 + L\mathbb{E}_{\mathcal{A},\mathcal{D}}(\Delta(\boldsymbol{v}_T, \boldsymbol{v}'_T) \mid \Delta(\boldsymbol{v}_{t_0}, \boldsymbol{v}'_{t_0}) = 0). \tag{12}$$

For the second term, notice that given $\Delta(\boldsymbol{v}_{t_0}, \boldsymbol{v}'_{t_0}) = 0$ and (11), it holds that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A},\mathcal{D}}\Delta(\boldsymbol{v}_T, \boldsymbol{v}'_T) \\
\leq& \sum_{t=t_0+1}^{T} \{ \prod_{k=t+1}^{T} \exp(\frac{m_1}{k})\} \frac{m_2^0}{nt} \\
=& \sum_{t=t_0+1}^{T} \exp(m_1 \sum_{k=t+1}^{T} \frac{1}{k}) \frac{m_2^0}{nt} \\
\leq& \sum_{t=t_0+1}^{T} \exp(m_1 \log(T/t)) \frac{m_2^0}{nt} \\
=& \frac{m_2^0}{n} T^{m_1} \sum_{t=t_0+1}^{T} \frac{1}{t^{m_1+1}} \\
\leq& \frac{m_2^0}{n} \frac{1}{m_1} (\frac{T}{t_0})^{m_1}.
\end{aligned}$$

Therefore, defining $t'_0 \triangleq t_0 \bar{\sigma}^2$ and $m_2 \triangleq m_2^0 \bar{\sigma}^{2m_1}$, one can bound the generalization gap as

$$\begin{aligned}
\mathbb{E}_{\mathcal{A},\mathcal{D}}\mathcal{L}_T^s(\boldsymbol{w}) - \hat{\mathcal{L}}_T^s(\boldsymbol{w}) &\leq \frac{t_0}{n}\bar{\sigma}^2 + L\frac{m_2^0}{n}\frac{1}{m_1}(\frac{T}{t_0})^{m_1} \\
&= \frac{t'_0}{n} + L\frac{m_2}{n}\frac{1}{m_1}(\frac{T}{t'_0})^{m_1}.
\end{aligned}$$

By choosing[6] $t'_0 = (Lm_2)^{\frac{1}{m_1+1}} T^{\frac{m_1}{m_1+1}}$, the generalization gap is bounded by

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\mathcal{L}_T^s(\boldsymbol{w}) - \hat{\mathcal{L}}_T^s(\boldsymbol{w}) \leq (Lm_2)^{\frac{1}{m_1+1}} \left[1 + \frac{1}{m_1}\right] \frac{T^{\frac{m_1}{m_1+1}}}{n}.$$

We finally remark that here it requires $t_0 < T$, which leads to $Lm_2 < T$. Recall that $m_2 = \frac{8c_2}{H\underline{\sigma}^2}(L + \gamma n)$ and therefore it suffices to require $\gamma = o(T/n)$.

Lemma 14 *Under the Assumptions in Lemma 13, there exists a $T_0 = 8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2} = \mathcal{O}(1)$ such that for $t > T_0$,*

$$\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|} \leq 1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}. \tag{13}$$

Proof 4 (Proof of Lemma 14) *Due to Argument 3 in Lemma 3, it holds that*

$$\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|} = [1 + \tilde{\eta}_t^2\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\tilde{\eta}_t\ell_t(\boldsymbol{v}_t)]^{-1/2}.$$

By plugging into the effective stepsize $\tilde{\eta}_t = \frac{2c_2}{H\underline{\sigma}^2(t+2)}$ and the loss upper bound in Assumption 1, it holds that

$$\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|} \leq [1 - 2H\tilde{\eta}_t\ell_t(\boldsymbol{v}_t)]^{-1/2} \leq \left[1 - 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right]^{-1/2}.$$

Furthermore, due to the fact that $(1-u)^{-1/2} \leq 1 + u$ when $u \in [0, 1/2]$,

$$\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|} \leq 1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2},$$

which is guaranteed by the condition $t > T_0 = 8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}$.

[6]Here the choice of t_0 naturally holds that $t_0 \geq T_0 = 8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2} = \mathcal{O}(1)$.

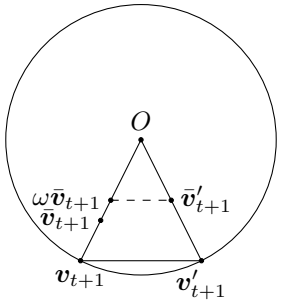
Lemma 15 *Under the Assumptions in Lemma 13, it holds that*

$$\Delta(\boldsymbol{v}_{t+1}, \boldsymbol{v}'_{t+1}) \leq \max\left\{\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|}, \frac{\|\boldsymbol{w}'_t\|}{\|\boldsymbol{w}'_{t+1}\|}\right\} \Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}). \tag{14}$$

Proof 5 (Proof of Lemma 15) *We first notice that $\boldsymbol{v}_{t+1}$ is the projection of $\bar{\boldsymbol{v}}_{t+1}$ to a standard sphere. Therefore, due to the iteration in* (17)*, we derive that*

$$\boldsymbol{v}_{t+1} = \frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|}\bar{\boldsymbol{v}}_{t+1}.$$

We next show the results in (14)*. Due to the symmetry, we assume without loss of generality that $\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|} \leq \frac{\|\boldsymbol{w}'_t\|}{\|\boldsymbol{w}'_{t+1}\|}$, which indicates that $\|\bar{\boldsymbol{v}}'_{t+1}\| \leq \|\bar{\boldsymbol{v}}_{t+1}\|$ since $\|\boldsymbol{v}_{t+1}\| = \|\boldsymbol{v}'_{t+1}\| = 1$. Therefore, it suffices to show that $\Delta(\boldsymbol{v}_{t+1}, \boldsymbol{v}'_{t+1}) \leq \frac{\|\boldsymbol{w}'_t\|}{\|\boldsymbol{w}'_{t+1}\|}\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1})$.*



We next only focus on the space spanned by vector $\boldsymbol{v}_{t+1}$ and $\boldsymbol{v}'_{t+1}$. Let $\omega = \frac{\|\bar{\boldsymbol{v}}'_{t+1}\|}{\|\bar{\boldsymbol{v}}_{t+1}\|}$ denotes a ratio, which guarantees that $\|\omega\bar{\boldsymbol{v}}_{t+1}\| = \|\bar{\boldsymbol{v}}'_{t+1}\|$. When $\|\bar{\boldsymbol{v}}'_{t+1}\| \leq \|\bar{\boldsymbol{v}}_{t+1}\|$, we have $\Delta(\omega\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \leq \Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1})$. Therefore, it holds that

$$\Delta(\boldsymbol{v}_{t+1}, \boldsymbol{v}'_{t+1}) = \frac{\|\boldsymbol{w}'_t\|}{\|\boldsymbol{w}'_{t+1}\|}\Delta(\omega\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}) \leq \frac{\|\boldsymbol{w}'_t\|}{\|\boldsymbol{w}'_{t+1}\|}\Delta(\bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}'_{t+1}),$$

where the first equation is due to the projection. This completes the proof.

B.4 Proof of Corollary 6

Proof 6 *For clarity, this proof follows the derivation of the second statement in Lemma 12. The rest follows from a similar argument. The proof proceeds by substituting specific constants with new values while reusing the analytical framework established in the proof of Theorem 5.*

Let $c'_2 = k_1 c_2$. We first show that for the identical stepsizes satisfying $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t \leq \frac{2c_2}{H\underline{\sigma}^2}(t+1)^{\frac{H-4}{2}}$, the effective stepsize in Corollary 6 is bounded by:

$$\tilde{\eta}_t \leq \frac{2c'_2}{H\underline{\sigma}^2(t+1)}. \tag{15}$$

Suppose, for the sake of contradiction, there exists some t such that $\tilde{\eta}_t > \frac{2c'_2}{H\underline{\sigma}^2(t+1)}$.

Due to assumption that $\ell_t(z) \leq z\ell'_t(z)$, Lemma 21 and therefore the equation (4) *still hold. Combining it with the assumption that $0 < \rho_t \leq k_1$ with $k_1 \geq 1$, we have:*

$$\begin{aligned}
\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_{t+1} &= \mathbb{E}_{\mathcal{A},\mathcal{D}}\frac{\tilde{\eta}_{t+1}}{\rho_{t+1}\|\boldsymbol{w}_{t+1}\|^{H-2}} \geq \frac{\tilde{\eta}_{t+1}}{\rho_{t+1}[\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2]^{(H-2)/2}} \\
&\geq \frac{1}{k_1}\cdot\frac{\tilde{\eta}_{t+1}}{[\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2]^{(H-2)/2}} > \frac{1}{k_1}\cdot\frac{\frac{2c'_2}{H\underline{\sigma}^2(t+2)}}{(1/(t+2))^{\frac{H-2}{2}}} = \frac{2(c'_2/k_1)}{H\underline{\sigma}^2}(t+2)^{\frac{H-4}{2}}.
\end{aligned}$$

Substituting $c_2'/k_1 = c_2$, we obtain $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_{t+1} > \frac{2c_2}{H\underline{\sigma}^2}(t+2)^{\frac{H-4}{2}}$, which contradicts the given condition on the stepsize. Thus, the effective stepsize must satisfy $\tilde{\eta}_t \leq \frac{2c_2'}{H\underline{\sigma}^2(t+1)}$.

Next, we bound the norm ratio $\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|}$. Combining the assumption that $z\ell_t'(z) \leq k_2\ell_t(z)$, and the homogeneity of the neural network Φ, we have:

$$\begin{aligned}
\boldsymbol{w}_t\top\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t) &= \ell_t'(\Phi(\boldsymbol{w}_t))\boldsymbol{w}_t\top\nabla_{\boldsymbol{w}_t}\Phi(\boldsymbol{w}_t) \\
&= \rho_t\ell_t'(\Phi(\boldsymbol{v}_t))\boldsymbol{w}_t\top\nabla_{\boldsymbol{w}_t}\Phi(\boldsymbol{w}_t) \\
&= \rho_t\|\boldsymbol{w}_t\|^H\ell_t'(\Phi(\boldsymbol{v}_t))\boldsymbol{v}_t\top\nabla_{\boldsymbol{v}_t}\Phi(\boldsymbol{v}_t) \\
&= \rho_t H\|\boldsymbol{w}_t\|^H\Phi(\boldsymbol{v}_t)\ell_t'(\Phi(\boldsymbol{v}_t)) \\
&\leq \rho_t k_2 H\|\boldsymbol{w}_t\|^H\ell_t(\Phi(\boldsymbol{v}_t)) \\
&\leq \rho_t k_2\bar{\sigma}^2 H\|\boldsymbol{w}_t\|^H.
\end{aligned}$$

Substituting this bound into the norm iteration $\|\boldsymbol{w}_{t+1}\|^2 = \|\boldsymbol{w}_t\|^2 + \eta_t^2\|\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t)\|^2 - 2\eta_t\boldsymbol{w}_t^\top\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t)$, we obtain:

$$\begin{aligned}
\|\boldsymbol{w}_{t+1}\|^2 &\geq \|\boldsymbol{w}_t\|^2 - 2\eta_t\left(\rho_t k_2\bar{\sigma}^2 H\|\boldsymbol{w}_t\|^H\right) \\
&= \|\boldsymbol{w}_t\|^2 - 2\left(\frac{\tilde{\eta}_t}{\rho_t\|\boldsymbol{w}_t\|^{H-2}}\right)\rho_t k_2 H\bar{\sigma}^2\|\boldsymbol{w}_t\|^H \\
&= \|\boldsymbol{w}_t\|^2\left(1 - 2k_2 H\bar{\sigma}^2\tilde{\eta}_t\right).
\end{aligned}$$

Then, we can replace the upper bound in (10) *with $1 + 4\frac{c_2'k_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}$.*

As the remaining derivation for the generalization gap in Lemma 13 remains valid, we directly apply the established results by substituting c_2 with $c_2' = k_1c_2$, and $\bar{\sigma}^2$ with $\bar{\sigma}_0^2 = k_2\bar{\sigma}^2$. Finally, updating the definitions of m_1 and m_2 with these values yields the modified constants stated in Corollary 6.

The proof of Corollary 7 can be directly derived from Theorem 5 given $H = 3$.

B.5 Proof of Lemma 3

Lemma 3 is a combination of Lemma 16, Lemma 17, and Lemma 18.

Lemma 16 (Effective stepsize) *For an H-homogeneous function ℓ_t with $H \neq 0$, the effective stepsize satisfies $\tilde{\eta}_t = \eta_t\|\boldsymbol{w}_t\|^{H-2}$, where the effective stepsize $\tilde{\eta}_t$ is defined as the stepsize in updating the direction $\boldsymbol{v}_t = \boldsymbol{w}_t/\|\boldsymbol{w}_t\|$.*

Proof 7 (Proof of Lemma 16) *For an H-homogeneous function ℓ_t with $H \neq 0$, it holds that*

$$\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t) = \|\boldsymbol{w}_t\|^{H-1}\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t), \tag{16}$$

where $\boldsymbol{v}_t = \boldsymbol{w}_t/\|\boldsymbol{w}_t\|$. This is due to the fact that $\ell_t(\boldsymbol{w}_t) = \|\boldsymbol{w}_t\|^H\ell_t(\boldsymbol{v}_t)$ and plug it into the definition of derivation. Therefore, we rewrite the update in (1) *as*

$$\boldsymbol{v}_{t+1}\|\boldsymbol{w}_{t+1}\| = \|\boldsymbol{w}_t\|(\boldsymbol{v}_t - \eta_t\|\boldsymbol{w}_t\|^{H-2}\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)). \tag{17}$$

This implies that the update on the direction $\boldsymbol{v}_t$ has effective stepsize as $\eta_t\|\boldsymbol{w}_t\|^{H-2}$.

Lemma 17 (Inner Product) *For an H-homogeneous function ℓ_t with $H \neq 0$, it holds that*

$$\boldsymbol{w}_t^\top\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t) = H\ell_t(\boldsymbol{w}_t). \tag{18}$$

Proof 8 (Proof of Lemma 17) *Note that for an H-homogeneous ℓ_t, it holds that $\ell_t(c\boldsymbol{w}_t) = c^H\ell_t(\boldsymbol{w}_t)$. By taking derivation on c, it holds that*

$$\boldsymbol{w}_t^\top\frac{\nabla\ell_t(c\boldsymbol{w}_t)}{\nabla c\boldsymbol{w}_t} = Hc^{H-1}\ell_t(\boldsymbol{w}_t).$$

Taking $c = 1$ leads to

$$\boldsymbol{w}_t^\top\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t) = H\ell_t(\boldsymbol{w}_t).$$

Lemma 18 (Norm Iterate) *For an H-homogeneous function ℓ_t with $H \neq 0$, it holds that*

$$\|\boldsymbol{w}_{t+1}\|^2 = \|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\tilde{\eta}_t\ell_t(\boldsymbol{v}_t)). \tag{19}$$

Proof 9 (Proof of Lemma 18) *We consider the iteration of $\|\boldsymbol{w}_{t+1}\|$, as follows:*

$$\begin{aligned}
&\|\boldsymbol{w}_{t+1}\|^2 \\
=&\|\boldsymbol{w}_t - \eta_t\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t)\|^2 \\
=&\|\boldsymbol{w}_t\|^2 + \eta_t^2\|\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t)\|^2 - 2\eta_t\boldsymbol{w}_t^\top\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t) \\
\overset{(i)}{=}&\|\boldsymbol{w}_t\|^2 + \eta_t^2\|\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t)\|^2 - 2H\eta_t\ell_t(\boldsymbol{w}_t) \\
\overset{(ii)}{=}&\|\boldsymbol{w}_t\|^2 + \eta_t^2\|\boldsymbol{w}_t\|^{2(H-1)}\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\eta_t\|\boldsymbol{w}_t\|^H\ell_t(\boldsymbol{v}_t) \\
=&\|\boldsymbol{w}_t\|^2(1 + \eta_t^2\|\boldsymbol{w}_t\|^{2(H-2)}\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\eta_t\|\boldsymbol{w}_t\|^{H-2}\ell_t(\boldsymbol{v}_t)) \\
\overset{(iii)}{=}&\|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\tilde{\eta}_t\ell_t(\boldsymbol{v}_t)).
\end{aligned}$$

where the equation (i) is due to Lemma 17, the equation (ii) is based on the fact that $\|\nabla_{\boldsymbol{w}_t}\ell_t(\boldsymbol{w}_t)\| = \|\boldsymbol{w}_t\|^{H-1}\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|$ and $\ell_t(\boldsymbol{w}_t) = \|\boldsymbol{w}_t\|^H\ell_t(\boldsymbol{v}_t)$, and the equation (iii) follows from the effective stepsize result in Lemma 16.

B.6 Proof of Theorem 8

This section provides the proofs of Theorem 8. This result is based on the following Lemma 19 and Lemma 20. Lemma 19 demonstrate that the effective stepsize would be $\mathcal{O}(\frac{1}{t\log t})$ given the stepsize $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \mathcal{O}(1/t)$. As a comparison, the effective stepsize used in the first statement of Lemma 12 would be $\Theta(1/t)$, given a stepsize $\Omega(1/\sqrt{t})$. Lemma 20 further derives the convergence rate given the effective stepsize. Combining Lemma 19 and Lemma 20 leads to Theorem 8.

Lemma 19 (The Effective Stepsize for Stepsize $\mathcal{O}(1/t)$) *Under the settings in Theorem 8, there exist infinitely many stepsizes satisfies $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \mathcal{O}(1/t)$, such that the effective stepsize satisfies $\tilde{\eta}_t = \mathcal{O}(\frac{1}{t\log t})$.*

Proof 10 (Proof of Lemma 19) *Firstly, notice that according to Lemma 21, it holds that*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2 \leq \mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2\left(1 - \frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t\right) \leq \mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2\exp\left(-\frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t\right),$$

By taking η_t such that $\tilde{\eta}_t = \frac{4}{H\underline{\sigma}^2((t+2)\log(t+2))}$, it holds that for $T > 0$

$$\begin{aligned}
\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_T\|^2 \leq&\exp\left(-\frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_{T-1}\right)\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{T-1}\|^2 \\
\leq&\exp\left(-\frac{1}{2}H\underline{\sigma}^2\sum_{t\in[T]}\tilde{\eta}_t\right)\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2 \\
\approx&\frac{\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2}{(\log T)^2},
\end{aligned}$$

where we use the fact that $\sum_{t\in[T]}\frac{1}{(t+2)\log(t+2)} = \Theta(\log(\log(T)))$. This leads to the learning rate

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \mathbb{E}_{\mathcal{A},\mathcal{D}}\frac{\tilde{\eta}_t}{\|\boldsymbol{w}_t\|} \geq \frac{\tilde{\eta}_t}{[\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2]^{1/2}} = \frac{4\log t}{\sqrt{\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2}H\underline{\sigma}^2(t+2)\log(t+2)} = \Omega(1/t),$$

where we use Jensen's inequality, and the assumption that $\mathbb{E}_{\mathcal{A}}\|\boldsymbol{w}_0\|^2$ is a constant.

We next prove Lemma 19 by contradiction. Assume that the stepsize $\mathbb{E}_{\mathcal{A},\mathcal{D}}\eta_t = \mathcal{O}(1/t)$ corresponds to a larger effective stepsize $\tilde{\eta}_t = \omega(\frac{1}{t\log t})$. In this case, it still holds for the weight norm that $\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_T\|^2 = \mathcal{O}(\frac{\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_0\|^2}{(\log T)^2})$. Therefore, the corresponding learning rate would become larger

according to the inequality that $\mathbb{E}_{\mathcal{A}}\eta_t \geq \frac{\tilde{\eta}_t}{[\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2]^{1/2}} = \omega(1/t)$*, which contradict with the statement that* $\eta_t = \mathcal{O}(1/t)$*. Therefore, given the stepsize* $\eta_t = \mathcal{O}(1/t)$*, the corresponding effective learning rate must be* $\tilde{\eta}_t = \mathcal{O}(\frac{1}{t\log t})$*. The proof is done.*

Lemma 20 *Under the assumptions in Theorem 8 with smoothness (parameter* β*), PL condition (parameter* μ*), strong growth condition (parameter* B*) and* $\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) \geq (1+\alpha)/\alpha\hat{\mathcal{L}}(\boldsymbol{v}^*)$*, given the effective learning rate* $\tilde{\eta}_t \leq \frac{1}{\beta B^2}$*, the convergence rate would be*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\hat{\mathcal{L}}(\boldsymbol{v}_T) - \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v}) \leq \exp\left(-\lambda\sum\tilde{\eta}_t\right)\mathbb{E}_{\mathcal{A},\mathcal{D}}(\hat{\mathcal{L}}(\boldsymbol{v}_0) - \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v})).$$

Therefore, taking $\tilde{\eta}_t = c/t$ *leads to*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\hat{\mathcal{L}}(\boldsymbol{v}_T) - \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v}) \leq \frac{1}{T^{c\lambda}}\mathbb{E}_{\mathcal{A},\mathcal{D}}(\hat{\mathcal{L}}(\boldsymbol{v}_0) - \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v})).$$

Taking $\tilde{\eta}_t = c/[t\log t]$ *leads to*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\hat{\mathcal{L}}(\boldsymbol{v}_T) - \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v}) \leq \frac{1}{[\log T]^{c\lambda}}\mathbb{E}_{\mathcal{A},\mathcal{D}}(\hat{\mathcal{L}}(\boldsymbol{v}_0) - \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v})).$$

Proof 11 (Proof of Lemma 20) *For the ease of notation, denote* $\boldsymbol{v}^*$ *as the one of the vectors which attains* $\hat{\mathcal{L}}(\boldsymbol{v}^*) = \min_{\boldsymbol{v}}\hat{\mathcal{L}}(\boldsymbol{v})$*. We first focus on the unconstrained iteration on the sphere, that is*

$$\bar{\boldsymbol{v}}_{t+1} = \boldsymbol{v}_t - \tilde{\eta}_t\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t).$$

Notice that by strong growth condition, it holds that ((Schmidt & Roux, 2013), (12)) when $\tilde{\eta}_t \leq \frac{1}{\beta B^2}$

$$\mathbb{E}_I\hat{\mathcal{L}}(\bar{\boldsymbol{v}}_{t+1}) \leq \hat{\mathcal{L}}(\boldsymbol{v}_t) - \tilde{\eta}_t\left(1 - \frac{\tilde{\eta}_t\beta B^2}{2}\right)\|\nabla_{\boldsymbol{v}_t}\hat{\mathcal{L}}(\boldsymbol{v}_t)\|^2 \leq \hat{\mathcal{L}}(\boldsymbol{v}_t) - \frac{\tilde{\eta}_t}{2}\|\nabla_{\boldsymbol{v}_t}\hat{\mathcal{L}}(\boldsymbol{v}_t)\|^2. \tag{20}$$

Besides, by PL condition and strong growth condition, for any t *and* $\boldsymbol{v}_t$*, it holds that*

$$\|\nabla_{\boldsymbol{v}_t}\hat{\mathcal{L}}(\boldsymbol{v}_t)\|^2 \geq \frac{1}{B^2}\max_i\{\|\nabla_{\boldsymbol{v}_t}\ell_i t(\boldsymbol{v}_t)\|^2\} \geq \frac{2\mu}{B^2}(\ell_i t(\boldsymbol{v}_t) - \ell_i t(\boldsymbol{v}^*)).$$

By taking expectation $\mathbb{E}_I$ *over the last iteration, it holds that (note that* $\boldsymbol{v}_t$ *is independent of the choice of the last iteration* ℓ_t*)*

$$\|\nabla_{\boldsymbol{v}_t}\hat{\mathcal{L}}(\boldsymbol{v}_t)\|^2 \geq \frac{2\mu}{B^2}(\mathbb{E}_I\ell_i(\boldsymbol{v}_t) - \mathbb{E}_I\ell_i(\boldsymbol{v}^*)) \geq \frac{2\mu}{B^2}(\hat{\mathcal{L}}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}^*)). \tag{21}$$

Combining (20) *and* (21) *leads to*

$$\mathbb{E}_I\hat{\mathcal{L}}(\bar{\boldsymbol{v}}_{t+1}) \leq \left(1 - \frac{\tilde{\eta}_t\mu}{B^2}\right)\hat{\mathcal{L}}(\boldsymbol{v}_t) + \frac{\tilde{\eta}_t\mu}{B^2}\hat{\mathcal{L}}(\boldsymbol{v}^*).$$

Next, we project $\bar{\boldsymbol{v}}_{t+1}$ *back onto the unit sphere. Analogous to Lemma 14, it holds that* $\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|} \leq (1 - 2H\bar{\sigma}^2\tilde{\eta}_t)^{-\frac{1}{2}} \leq 1 + 2H\bar{\sigma}^2\tilde{\eta}_t$*. Thus:*

$$\begin{aligned}
\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_{t+1}) - \hat{\mathcal{L}}(\boldsymbol{v}^*) &= \mathbb{E}_I\left[\frac{\|\boldsymbol{w}_t\|}{\|\boldsymbol{w}_{t+1}\|}\right]^H\hat{\mathcal{L}}(\bar{\boldsymbol{v}}_{t+1}) - \hat{\mathcal{L}}(\boldsymbol{v}^*) \\
&\leq (1 + 4H^2\bar{\sigma}^2\tilde{\eta}_t)\left[\left(1 - \frac{\tilde{\eta}_t\mu}{B^2}\right)\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) + \frac{\tilde{\eta}_t\mu}{B^2}\hat{\mathcal{L}}(\boldsymbol{v}^*)\right] - \hat{\mathcal{L}}(\boldsymbol{v}^*) \\
&= (1 + 4H^2\bar{\sigma}^2\tilde{\eta}_t)\left(1 - \frac{\tilde{\eta}_t\mu}{B^2}\right)\left(\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}^*)\right) + 4H^2\bar{\sigma}^2\tilde{\eta}_t\hat{\mathcal{L}}(\boldsymbol{v}^*).
\end{aligned}$$

By $\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) \geq (1+\alpha)/\alpha\hat{\mathcal{L}}(\boldsymbol{v}^*)$*, it holds that:*

$$\begin{aligned}
\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_{t+1}) - \hat{\mathcal{L}}(\boldsymbol{v}^*) &\leq (1 + 4H^2\bar{\sigma}^2\tilde{\eta}_t)\left(1 - \frac{\tilde{\eta}_t\mu}{B^2}\right)\left(\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}^*)\right) \\
&\quad + 4\alpha H^2\bar{\sigma}^2\tilde{\eta}_t\left(\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_t) - \hat{\mathcal{L}}(\boldsymbol{v}^*)\right) \\
&\leq \left[1 - \tilde{\eta}_t\left(\frac{\mu}{B^2} - 4H^2\bar{\sigma}^2(1+\alpha)\right)\right]\left(\mathbb{E}_I\hat{\mathcal{L}}(\boldsymbol{v}_{t+1}) - \hat{\mathcal{L}}(\boldsymbol{v}^*)\right).
\end{aligned}$$

Note that $\lambda \triangleq \frac{\mu}{B^2} - 4H^2\bar{\sigma}^2(1+\alpha) > 0$. *Taking the expectation over the algorithm and dataset, the convergence would be*

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{D}}\hat{\mathcal{L}}(\boldsymbol{v}_{t+1}) - \hat{\mathcal{L}}(\boldsymbol{v}^*) \leq& \mathbb{E}_{\mathcal{A},\mathcal{D}}(\hat{\mathcal{L}}(\boldsymbol{v}_0) - \hat{\mathcal{L}}(\boldsymbol{v}^*))\prod(1-\tilde{\eta}_t\lambda)\\ \leq& \exp(-\lambda\sum\tilde{\eta}_t)\mathbb{E}_{\mathcal{A},\mathcal{D}}(\hat{\mathcal{L}}(\boldsymbol{v}_0) - \hat{\mathcal{L}}(\boldsymbol{v}^*)).\end{aligned}$$

B.7 Technical Lemmas

Lemma 21 (Weight Norm Iteration) *Under Assumption 1 and Assumption 2, assume that* $H > 0$ *and* $\tilde{\eta}_t \leq \frac{H\underline{\sigma}^2}{2\bar{L}^2}$, *if the loss is* H*-homogeneous, it holds that*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_{t+1}\|^2 \leq \mathbb{E}_{\mathcal{A},\mathcal{D}}\|\boldsymbol{w}_t\|^2(1 - \frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t),$$

where the expectation $\mathbb{E}_{\mathcal{A}}$ *is taken over the training algorithm.*

Proof 12 (Proof of Lemma 21) *Note that due to Lemma 3 (argument 3), it holds that*

$$\begin{aligned}\mathbb{E}_I\|\boldsymbol{w}_{t+1}\|^2 =& \mathbb{E}_I\|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\tilde{\eta}_t\ell_t(\boldsymbol{v}_t))\\ =& \|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2\mathbb{E}_I\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 - 2H\tilde{\eta}_t\hat{\mathcal{L}}(\boldsymbol{v}_t)),\end{aligned}$$

where the expectation $\mathbb{E}_I$ *is taken over the choice of the last iteration. Note that we use the fact that* $\boldsymbol{v}_t$ *is independent of the choice of the last iteration (namely,* ℓ_t*), and therefore,*

$$\mathbb{E}_I\ell_t(\boldsymbol{v}_t) = \frac{1}{n}\sum_{i\in[n]}\ell_t(\boldsymbol{v}_t; X_i, Y_i) = \hat{\mathcal{L}}_t(\boldsymbol{v}_t).$$

We next plug the bound of Assumption 1, that is, $\hat{\mathcal{L}}(\boldsymbol{v}_t) \geq \frac{1}{2}\underline{\sigma}^2$, *and the bound of Assumption 2, that is,* $\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\|\nabla_{\boldsymbol{v}}\ell_t(\boldsymbol{v})\| \leq L$, *namely,*

$$\begin{aligned}\mathbb{E}_I\|\boldsymbol{w}_{t+1}\|^2 \leq& \|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2L^2 - 2H\tilde{\eta}_t\hat{\mathcal{L}}(\boldsymbol{v}_t))\\ \leq& \|\boldsymbol{w}_t\|^2(1 + \tilde{\eta}_t^2L^2 - H\tilde{\eta}_t\underline{\sigma}^2)\\ \leq& \|\boldsymbol{w}_t\|^2(1 - \tilde{\eta}_t(H\underline{\sigma}^2 - \tilde{\eta}_tL^2)).\end{aligned}$$

By the requirement that $\tilde{\eta}_t \leq \frac{H\underline{\sigma}^2}{2\bar{L}^2}$, *it holds that*

$$\mathbb{E}_I\|\boldsymbol{w}_{t+1}\|^2 \leq \|\boldsymbol{w}_t\|^2(1 - \frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t).$$

Taking expectations over the whole algorithm and the dataset finishes the proof.

Remark: Extension to the non-Lipschitz setting. *This result extends to the non-Lipschitz scenario. The analysis relies on the self-bounding property of* β*-smooth functions:*

$$\begin{aligned}\mathbb{E}_I\|\boldsymbol{w}_{t+1}\|^2 \leq& \|\boldsymbol{w}_t\|^2(1 + 2\beta\tilde{\eta}_t^2\mathbb{E}_I[\ell_t(\boldsymbol{v}_t)] - 2H\tilde{\eta}_t\hat{\mathcal{L}}(\boldsymbol{v}_t))\\ =& \|\boldsymbol{w}_t\|^2(1 + 2\beta\tilde{\eta}_t^2\hat{\mathcal{L}}(\boldsymbol{v}_t) - 2H\tilde{\eta}_t\hat{\mathcal{L}}(\boldsymbol{v}_t))\\ =& \|\boldsymbol{w}_t\|^2(1 - 2\tilde{\eta}_t\hat{\mathcal{L}}(\boldsymbol{v}_t)(H - \beta\tilde{\eta}_t))\\ \leq& \|\boldsymbol{w}_t\|^2(1 - \underline{\sigma}^2\tilde{\eta}_t(H - \beta\tilde{\eta}_t)).\end{aligned}$$

Consequently, if $\tilde{\eta}_t \leq \frac{H}{2\beta}$, *we still obtain:*

$$\mathbb{E}_I\|\boldsymbol{w}_{t+1}\|^2 \leq \|\boldsymbol{w}_t\|^2(1 - \frac{1}{2}H\underline{\sigma}^2\tilde{\eta}_t).$$

B.8 Proofs for Section 4

This section provides the proof for the results presented in Section 4. We first derive Lemma 10, and subsequently establish Theorem 11.

B.8.1 PROOF OF LEMMA 10

Proof 13 *This proof decomposes the generalization gap by utilizing the notion of* conditional on-average stability *(Definition 9).*

Let $z_i \triangleq (X_i, Y_i)$ *and define the individual loss difference on the* i*-th sample as* $\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)}) \triangleq |\ell(\boldsymbol{v}_T; z_i) - \ell(\boldsymbol{v}_T^{(i)}; z_i)|$*. Following the standard stability analysis in Bousquet & Elisseeff (2002), we have*

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] &= \mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}(\boldsymbol{v}_T) - \hat{\mathcal{L}}(\boldsymbol{v}_T)] \\ &= \mathbb{E}_{\mathcal{A},\mathcal{D}}[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{v}_T^{(i)}; z_i) - \frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{v}_T; z_i)] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)})].\end{aligned}$$

Let $\mathcal{E}_{(i)}$ *denote the event that the algorithm* $\mathcal{A}$ *does* not *select the index* i *within the first* t_0 *iterations. The probability of the complement event* $\mathcal{E}_{(i)}^c$ *(i.e., index* i *is selected at least once) is bounded by:*

$$P(\mathcal{E}_{(i)}^c) \leq \sum_{t=1}^{t_0}\frac{1}{n} = \frac{t_0}{n},$$

Analogous to the technique in (12)*, invoking the Bounded Loss assumption (Assumption 1) yields:*

$$\begin{aligned}\mathbb{E}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)})] &= P(\mathcal{E}_{(i)}^c)\mathbb{E}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)}) \mid \mathcal{E}_{(i)}^c] + P(\mathcal{E}_{(i)})\mathbb{E}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)}) \mid \mathcal{E}_{(i)}] \\ &\leq \frac{t_0}{n}\cdot\bar{\sigma}^2 + \mathbb{E}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)}) \mid \mathcal{E}_{(i)}].\end{aligned} \tag{22}$$

We recall the result derived by Lei & Ying (2020), as in Proposition 25. Notably, its derivation relies solely on the properties of the individual loss rather than the empirical loss. Consequently, under the assumption of β*-smooth, the following inequality holds:*

$$|\ell(\boldsymbol{v}_T; z) - \ell(\boldsymbol{v}_T^{(i)}; z)| \leq \frac{\beta}{\zeta}\ell(\boldsymbol{v}_T; z) + \frac{\beta+\zeta}{2}\|\boldsymbol{v}_T - \boldsymbol{v}_T^{(i)}\|^2. \tag{23}$$

where $\zeta > 0$ *is a free parameter. Conditioning* (23) *on the event* $\mathcal{E}_{(i)}$ *and taking the expectation over the randomness of the algorithm and datasets on both sides, we obtain:*

$$\mathbb{E}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)}) \mid \mathcal{E}_{(i)}] \leq \frac{\beta}{\zeta}\mathbb{E}[\ell(\boldsymbol{v}_T; z_i) \mid \mathcal{E}_{(i)}] + \frac{\beta+\zeta}{2}\mathbb{E}[\|\boldsymbol{v}_T - \boldsymbol{v}_T^{(i)}\|^2 \mid \mathcal{E}_{(i)}].$$

The first term $\mathbb{E}[\ell(\boldsymbol{v}_T; z_i) \mid \mathcal{E}_{(i)}]$ *is uniformly bounded by* $\bar{\sigma}^2$ *under Assumption 1, while the second term corresponds exactly to the definition of the conditional on-average stability. Substituting this back into* (22) *yields:*

$$\mathbb{E}[\delta(\boldsymbol{v}_T, \boldsymbol{v}_T^{(i)})] \leq \frac{t_0}{n}\bar{\sigma}^2 + \frac{\beta}{\zeta}\bar{\sigma}^2 + \frac{\beta+\zeta}{2}\mathbb{E}[\|\boldsymbol{v}_T - \boldsymbol{v}_T^{(i)}\|^2 \mid \mathcal{E}_{(i)}].$$

Finally, by setting a uniform t_0 *for all indices* i *and averaging the individual bounds over* $i = 1, \ldots, n$*, the generalization gap satisfies:*

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] &\leq \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{t_0}{n} + \frac{\beta}{\zeta}\right)\bar{\sigma}^2 + \frac{\beta+\zeta}{2}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\|\boldsymbol{v}_T - \boldsymbol{v}_T^{(i)}\|^2 \mid \mathcal{E}_{(i)}]\right] \\ &= \left(\frac{t_0}{n} + \frac{\beta}{\zeta}\right)\bar{\sigma}^2 + \frac{\beta+\zeta}{2}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\|\boldsymbol{v}_T - \boldsymbol{v}_T^{(i)}\|^2 \mid \mathcal{E}_{(i)}]}_{\epsilon^2_{avg,T|t_0}}.\end{aligned}$$

This concludes the proof.

B.8.2 Proof of Theorem 11

We begin by establishing a recursive bound for stability in the following lemma.

Lemma 22 *Suppose the assumptions of Theorem 11 hold. Using the effective stepsize $\tilde{\eta}_t$ defined in Lemma 13, for any iteration $t > t_0$, the following recurrence holds:*

$$\mathbb{E}_I\|\boldsymbol{v}_{t+1} - \boldsymbol{v}_{t+1}^{(i)}\|^2 \leq \exp\left(\frac{m_1'}{t+2} + \frac{1}{T}\right)\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + m_2'\left(1+\frac{T}{n}\right)\frac{1}{n}\frac{1}{(t+2)^2},$$

where $m_1' = 8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2} + \frac{4c_2\beta}{H\underline{\sigma}^2}$ and $m_2' = \frac{32ec_2^2\beta\bar{\sigma}^2}{H^2\underline{\sigma}^4}$.

Proof 14 (Proof of Lemma 22) *We analyze the update at step $t > t_0$. Let i_t denote the sample index selected by SGD at step t.*

With probability $1 - 1/n$, the algorithm selects an index $i_t \neq i$ (i.e., the identical sample is used). By the β-smoothness of the loss function, the update satisfies:

$$\begin{aligned}\|\bar{\boldsymbol{v}}_{t+1} - \bar{\boldsymbol{v}}_{t+1}^{(i)}\|^2 &= \|(\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}) - \tilde{\eta}_t[\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)) - \nabla_{\boldsymbol{v}_t^{(i)}}\ell_t(\boldsymbol{v}_t^{(i)})]\|^2 \\ &\leq [\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\| + \tilde{\eta}_t\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)) - \nabla_{\boldsymbol{v}_t^{(i)}}\ell_t(\boldsymbol{v}_t^{(i)})\|]^2 \\ &\leq (1+\beta\tilde{\eta}_t)^2\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2.\end{aligned} \tag{24}$$

With probability $1/n$, the algorithm selects the index $i_t = i$ (i.e, the samples differ). Using the inequality $\|a+b\|^2 \leq (1+p)\|a\|^2 + (1+1/p)\|b\|^2$ with $p > 0$, the self-bounding property $\|\nabla\ell(\boldsymbol{v})\|^2 \leq 2\beta\ell(\boldsymbol{v})$, and the Bounded Loss assumption, we have

$$\begin{aligned}\|\bar{\boldsymbol{v}}_{t+1} - \bar{\boldsymbol{v}}_{t+1}^{(i)}\| &= \|(\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}) - \tilde{\eta}_t[\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)) - \nabla_{\boldsymbol{v}_t^{(i)}}\ell_t(\boldsymbol{v}_t^{(i)})]\|^2 \\ &\leq (1+p)\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + \tilde{\eta}_t^2(1+1/p)\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t) - \nabla_{\boldsymbol{v}_t^{(i)}}\ell_t(\boldsymbol{v}_t^{(i)})\|^2 \\ &\leq (1+p)\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + 2\tilde{\eta}_t^2(1+1/p)[\|\nabla_{\boldsymbol{v}_t}\ell_t(\boldsymbol{v}_t)\|^2 + \|\nabla_{\boldsymbol{v}_t^{(i)}}\ell_t(\boldsymbol{v}_t^{(i)})\|^2] \\ &\leq (1+p)\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + 4\beta(1+1/p)\tilde{\eta}_t^2[\ell(\boldsymbol{v}_t; z_i) + \ell(\boldsymbol{v}_t^{(i)}; z_i')] \\ &\leq (1+p)\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + 8\beta(1+1/p)\tilde{\eta}_t^2\bar{\sigma}^2.\end{aligned} \tag{25}$$

Taking the expectation over the choice of i_t, we combine (24) *and* (25)*:*

$$\mathbb{E}_I\|\bar{\boldsymbol{v}}_{t+1} - \bar{\boldsymbol{v}}_{t+1}^{(i)}\|^2 \leq \left[\left(1-\frac{1}{n}\right)(1+\beta\tilde{\eta}_t)^2 + \frac{1}{n}(1+p)\right]\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + \frac{8\beta}{n}\left(1+\frac{1}{p}\right)\tilde{\eta}_t^2\bar{\sigma}^2. \tag{26}$$

Utilizing the norm ratio bound in (10)*, we have*

$$\mathbb{E}_I\|\boldsymbol{v}_{t+1} - \boldsymbol{v}_{t+1}^{(i)}\|^2 \leq (1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2})^2\mathbb{E}_I\|\bar{\boldsymbol{v}}_{t+1} - \bar{\boldsymbol{v}}_{t+1}^{(i)}\|^2. \tag{27}$$

Setting $p = n/T$ and substituting (26) *into* (27)*, we obtain:*

$$\begin{aligned}&\mathbb{E}_I\|\boldsymbol{v}_{t+1} - \boldsymbol{v}_{t+1}^{(i)}\|^2 \\ &\leq (1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2})^2\left\{\left[\left(1-\frac{1}{n}\right)(1+\beta\tilde{\eta}_t)^2 + \frac{1}{n}(1+p)\right]\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 + \frac{8\beta}{n}\left(1+\frac{1}{p}\right)\tilde{\eta}_t^2\bar{\sigma}^2\right\} \\ &= \left\{\left(1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)^2\left[\left(1-\frac{1}{n}\right)(1+\beta\tilde{\eta}_t)^2 + \frac{1}{n}\left(1+\frac{n}{T}\right)\right]\right\}\|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|^2 \\ &\quad + \left(1 + 4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)^2\frac{8\beta}{n}\left(1+\frac{1}{p}\right)\tilde{\eta}_t^2\bar{\sigma}^2.\end{aligned} \tag{28}$$

For the first term in (28)*, we have:*

$$\begin{aligned}
&\left(1+4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)^2\left[\left(1-\frac{1}{n}\right)(1+\beta\tilde{\eta}_t)^2+\frac{1}{n}\left(1+\frac{n}{T}\right)\right]\\
&\leq\left(1+4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)^2\left[(1+\beta\tilde{\eta}_t)^2+\frac{1}{T}\right]\\
&\leq\left(1+4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)^2(1+\beta\tilde{\eta}_t)^2\left(1+\frac{1}{T}\right)\\
&\leq\exp\left(8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)\exp(2\beta\tilde{\eta}_t)\exp\left(\frac{1}{T}\right)\\
&=\exp\left(\frac{1}{T}+8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}+2\beta\frac{2c_2}{H\underline{\sigma}^2(t+2)}\right)\\
&=\exp\left(\frac{1}{T}+\left[8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}+\frac{4c_2\beta}{H\underline{\sigma}^2}\right]\frac{1}{t+2}\right)=\exp\left(\frac{1}{T}+\frac{m_1'}{t+2}\right).
\end{aligned}$$

For the second term in (28)*, we have:*

$$\begin{aligned}
&\left(1+4\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)^2\frac{8\beta}{n}\left(1+\frac{1}{p}\right)\tilde{\eta}_t^2\bar{\sigma}^2\\
&\leq\exp\left(8\frac{c_2\bar{\sigma}^2}{\underline{\sigma}^2}\frac{1}{t+2}\right)\frac{8\beta}{n}\left(1+\frac{T}{n}\right)\left(\frac{2c_2}{H\underline{\sigma}^2(t+2)}\right)^2\bar{\sigma}^2\\
&\leq e\cdot\frac{8\beta}{n}\left(1+\frac{T}{n}\right)\frac{4c_2^2}{H^2\underline{\sigma}^4(t+2)^2}\bar{\sigma}^2\\
&=\left[\frac{32ec_2^2\beta\bar{\sigma}^2}{H^2\underline{\sigma}^4}\right]\left(1+\frac{T}{n}\right)\frac{1}{(t+2)^2}\frac{1}{n}\\
&=m_2'\left(1+\frac{T}{n}\right)\frac{1}{n}\frac{1}{(t+2)^2}.
\end{aligned}$$

Consequently, (28) *is equivalent to*

$$\mathbb{E}_I\|\boldsymbol{v}_{t+1}-\boldsymbol{v}_{t+1}^{(i)}\|^2\leq\exp\left(\frac{m_1'}{t+2}+\frac{1}{T}\right)\|\boldsymbol{v}_t-\boldsymbol{v}_t^{(i)}\|^2+m_2'\left(1+\frac{T}{n}\right)\frac{1}{n}\frac{1}{(t+2)^2}. \tag{29}$$

This completes the proof of Lemma 22.

We then proceed to derive the generalization bound in Theorem 11. We first bound the conditional on-average stability, and then optimize the resulting generalization gap.

Proof 15 (Proof of Theorem 11) *Taking expectations of* (29) *over the randomness of the algorithm and datasets, conditioned on $\mathcal{E}_{(i)}$, we obtain:*

$$\begin{aligned}
\mathbb{E}_{\mathcal{A},\mathcal{D}}[\|\boldsymbol{v}_{t+1}-\boldsymbol{v}_{t+1}^{(i)}\|^2\mid\mathcal{E}_{(i)}]\leq\exp\left(\frac{m_1'}{t+2}+\frac{1}{T}\right)\mathbb{E}_{\mathcal{A},\mathcal{D}}[\|\boldsymbol{v}_t-\boldsymbol{v}_t^{(i)}\|^2\mid\mathcal{E}_{(i)}]\\
+m_2'\left(1+\frac{T}{n}\right)\frac{1}{n}\frac{1}{(t+2)^2}.
\end{aligned}$$

Averaging over all samples $1,\ldots,n$, and defining $\Delta_t\triangleq\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\mathcal{A},\mathcal{D}}[\|\boldsymbol{v}_t-\boldsymbol{v}_t^{(i)}\|^2\mid\mathcal{E}_{(i)}]$, we have:

$$\Delta_{t+1}\leq\exp\left(\frac{m_1'}{t+2}+\frac{1}{T}\right)\Delta_t+m_2'\left(1+\frac{T}{n}\right)\frac{1}{n}\frac{1}{(t+2)^2}.$$

Let $K_{n,T}\triangleq\frac{e}{m_1'+1}(1+\frac{T}{n})\frac{1}{n}m_2'$, where e is the base of the natural logarithm. Then:

$$\Delta_{t+1}\leq\exp\left(\frac{m_1'}{t+2}+\frac{1}{T}\right)\Delta_t+\frac{m_1'+1}{e}K_{n,T}\frac{1}{(t+2)^2}.$$

Unrolling the recurrence from $t = t_0$ to $T-1$ yields:

$$\begin{aligned}\Delta_T &\le \frac{m_1'+1}{e} K_{n,T} \sum_{t=t_0}^{T-1} \frac{1}{(t+2)^2} \prod_{k=t+1}^{T-1} \exp\left(\frac{m_1'}{k+2} + \frac{1}{T}\right) \\ &= \frac{m_1'+1}{e} K_{n,T} \sum_{t=t_0}^{T-1} \frac{1}{(t+2)^2} \exp \sum_{k=t+1}^{T-1} \left(\frac{m_1'}{k+2} + \frac{1}{T}\right) \\ &= \frac{m_1'+1}{e} K_{n,T} \sum_{t=t_0}^{T-1} \frac{1}{(t+2)^2} \exp\left(m_1' \sum_{k=t+1}^{T-1} \frac{1}{k+2} + \frac{T-t-1}{T}\right) \\ &\le (m_1'+1) K_{n,T} \sum_{t=t_0}^{T-1} \frac{1}{(t+2)^2} \exp\left(m_1' \sum_{k=t+1}^{T-1} \frac{1}{k+2}\right).\end{aligned}$$

Using the inequality $\sum_{k=t+1}^{T-1} \frac{1}{k+2} \le \ln(\frac{T}{t+1})$ and $\sum_{t=t_0}^{T-1}(t+1)^{-(2+m_1')} \le \int_{t_0}^{\infty} x^{-(2+m_1')} dx = \frac{1}{1+m_1'} t_0^{-(m_1'+1)}$, we obtain:

$$\Delta_T \le (m_1'+1) K_{n,T} T^{m_1'} \sum_{t=t_0}^{T-1} (t+1)^{-(2+m_1')} \le K_{n,T} T^{m_1'} t_0^{-(m_1'+1)}. \tag{30}$$

Substituting (30) *into* (24) *gives*

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] &\le \left(\frac{t_0}{n} + \frac{\beta}{\zeta}\right)\bar{\sigma}^2 + \frac{\beta+\zeta}{2}\Delta_T \\ &\le \left(\frac{t_0}{n} + \frac{\beta}{\zeta}\right)\bar{\sigma}^2 + \frac{\beta+\zeta}{2} K_{n,T} T^{m_1'} t_0^{-(m_1'+1)} \\ &= \frac{K_{n,T} T^{m_1'}}{2}(\zeta+\beta)\cdot t_0^{-(m_1'+1)} + \beta\bar{\sigma}^2 \frac{1}{\zeta} + \frac{\bar{\sigma}^2}{n} t_0.\end{aligned}$$

Define $g(\zeta, t_0) \triangleq \frac{K_{n,T} T^{m_1'}}{2}(\zeta+\beta)\cdot t_0^{-(m_1'+1)} + \beta\bar{\sigma}^2\frac{1}{\zeta} + \frac{\bar{\sigma}^2}{n} t_0$. Observing that the generalization gap is independent of ζ and t_0, we minimize g to obtain:

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}^s(\boldsymbol{w}_T) - \hat{\mathcal{L}}^s(\boldsymbol{w}_T)] &\le \min_{t_0,\zeta} g(t_0,\zeta) \\ &= (m_1'+3)/(m_1'+1)^{\frac{m_1'+2}{m_1'+3}} \cdot (\beta e m_2'/2)^{\frac{1}{m_1'+3}} \bar{\sigma}^{\frac{2m_1'+4}{m_1'+3}} \\ &\quad \cdot n^{-\frac{m_1'+2}{m_1'+3}} \left(1 + \frac{T}{n}\right)^{\frac{1}{m_1'+3}} T^{\frac{m_1'}{m_1'+3}} \\ &= \mathcal{O}(n^{-\frac{m_1'+2}{m_1'+3}} \left(1 + \frac{T}{n}\right)^{\frac{1}{m_1'+3}} T^{\frac{m_1'}{m_1'+3}}).\end{aligned}$$

This concludes the proof of Theorem 11.

B.9 Review of Several Stability Metrics

B.9.1 Review of Algorithmic Stability

Algorithmic stability is a popular technique in generalization theory, offering the following favorable properties:

Proposition 23 (Algorithmic stability and generalization)*The arguments are based on Bousquet & Elisseeff (2002); Shalev-Shwartz et al. (2010); Hardt et al. (2016). Assume that the algorithm is ϵ_T-stable, meaning that for all datasets $S = \{(X_1, Y_1), \cdots, (X_i, Y_i), \cdots, (X_n, Y_n)\}$ and $S' = \{(X_1, Y_1), \cdots, (X_i', Y_i'), \cdots, (X_n, Y_n)\}$ differing by at most one sample, it holds that*

$$\sup_{(X,Y)} \mathbb{E}_{\mathcal{A},\mathcal{D}}[\ell(\boldsymbol{w}_T; X, Y) - \ell(\boldsymbol{w}_T'; X, Y)] \le \epsilon_T,$$

where $\boldsymbol{w}_T$ and $\boldsymbol{w}'_T$ are trained on datasets S and S' over T iterations, respectively, and the expectation is taken over both the training set randomness and the algorithmic randomness. Then the generalization gap can be bounded via algorithmic stability:

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}\mathcal{L}(\boldsymbol{w}_T) - \hat{\mathcal{L}}(\boldsymbol{w}_T) \leq \epsilon_T.$$

Furthermore, it holds that for SGD with replacement using stepsize η_t for each iteration t:

1. *For **convex** loss functions, provided they are L-Lipschitz and β-smooth, and the stepsize satisfies $\eta_T \leq 2/\beta$, it holds that $\epsilon_T \leq \frac{2L^2}{n}\sum_{t=1}^{T}\eta_t$.*
2. *For **non-convex** loss functions, provided they are L-Lipschitz and β-smooth, and the stepsize is non-increasing with $\eta_t \leq c/t$ for a given constant c, it holds that $\epsilon_T \leq \frac{1+1/\beta c}{n-1}(2cL^2)^{\frac{1}{\beta c+1}}T^{\frac{\beta c}{\beta c+1}}$.*

B.9.2 Review of On-Average Model Stability

In this section, we provide the necessary background on the *on-average model stability* introduced by Lei & Ying (2020). Unlike uniform stability, which requires an upper bound on the loss difference for any pair of sample in the distribution, on-average model stability measures the average distance of weights trained via the same algorithm on two datasets differing at most one sample.

Definition 24 (ℓ_2 On-Average Model Stability) *Let $S = \{z_1, \ldots, z_n\}$ and $S^{(i)} = \{z_1, \ldots, z'_i, \ldots, z_n\}$ be two datasets differing only in the i-th sample. A randomized algorithm $\mathcal{A}$ is ϵ_{avg}-stable in ℓ_2 on-average if:*

$$\mathbb{E}_{S,\tilde{S},\mathcal{A}}\left[\frac{1}{n}\sum_{i=1}^{n}\|\mathcal{A}(S) - \mathcal{A}(S^{(i)})\|_2^2\right] \leq \epsilon_{avg}^2.$$

Based on this definition, the generalization gap can be bounded as follows:

Proposition 25 (Generalization via Model Stability) *If for any z, the loss $\ell(\boldsymbol{w}; z)$ is non-negative and β-smooth, then for any free parameter $\zeta > 0$:*

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathcal{L}(\boldsymbol{w}_T) - \hat{\mathcal{L}}(\boldsymbol{w}_T)] \leq \frac{\beta}{\zeta}\mathbb{E}_{\mathcal{A},\mathcal{D}}[\hat{\mathcal{L}}(\boldsymbol{w}_T)] + \frac{\beta+\zeta}{2}\epsilon_{avg}^2,$$

where ϵ_{avg}^2 denotes the on-average model stability of the algorithm's output $\boldsymbol{w}_T$.