BEYOND LIPSCHITZ: SHARP GENERALIZATION AND EXCESS RISK BOUNDS FOR FULL-BATCH GD

Konstantinos E. Nikolakakis*, **Farzin Haddadpour, Amin Karbasi**[1] **& Dionysios S. Kalogerias**
Department of Electrical Engineering
Yale University, Yale University & Google Research[1]
{first.last}@yale.edu

ABSTRACT

We provide sharp path-dependent generalization and excess risk guarantees for the full-batch Gradient Descent (GD) algorithm on smooth losses (possibly non-Lipschitz, possibly nonconvex). At the heart of our analysis is an upper bound on the generalization error, which implies that *average output stability* and a *bounded expected optimization error at termination* lead to generalization. This result shows that a small generalization error occurs along the optimization path, and allows us to bypass Lipschitz or sub-Gaussian assumptions on the loss prevalent in previous works. For nonconvex, convex, and strongly convex losses, we show the explicit dependence of the generalization error in terms of the accumulated path-dependent optimization error, terminal optimization error, number of samples, and number of iterations. For nonconvex smooth losses, we prove that full-batch GD efficiently generalizes close to any stationary point at termination, and recovers the generalization error guarantees of stochastic algorithms with fewer assumptions. For smooth convex losses, we show that the generalization error is tighter than existing bounds for SGD (up to one order of error magnitude). Consequently the excess risk matches that of SGD for quadratically less iterations. Lastly, for strongly convex smooth losses, we show that full-batch GD achieves essentially the same excess risk rate as compared with the state of the art on SGD, but with an exponentially smaller number of iterations (logarithmic in the dataset size).

1 INTRODUCTION

Gradient based learning (Lecun et al., 1998) is a well established topic with a large body of literature on algorithmic generalization and optimization errors. For general smooth convex losses, optimization error guarantees have long been well-known (Nesterov, 1998). Similarly, Absil et al. (2005) and Lee et al. (2016) have showed convergence of Gradient Descent (GD) to minimizers and local minima for smooth nonconvex functions. More recently, Chatterjee (2022), Liu et al. (2022) and Allen-Zhu et al. (2019) established global convergence of GD for deep neural networks under appropriate conditions.

Generalization error analysis of stochastic training algorithms has recently gained increased attention. Hardt et al. (2016) showed uniform stability final-iterate bounds for vanilla Stochastic Gradient Descent (SGD). More recent works have developed alternative generalization error bounds with probabilistic guarantees (Feldman & Vondrak, 2018; 2019; Madden et al., 2020; Klochkov & Zhivotovskiy, 2021) and data-dependent variants (Kuzborskij & Lampert, 2018), or under weaker assumptions such as strongly quasi-convex (Gower et al., 2019), non-smooth convex (Feldman, 2016; Bassily et al., 2020; Lei & Ying, 2020b; Lei et al., 2021a), and pairwise losses (Lei et al., 2021b; 2020). In the nonconvex case, Zhou et al. (2021b) provide bounds that involve the on-average variance of the stochastic gradients. Generalization performance of other algorithmic variants lately gain further attention, including SGD with early momentum (Ramezani-Kebrya et al., 2021), randomized coordinate descent (Wang et al., 2021c), look-ahead approaches (Zhou et al., 2021a), noise injection methods (Xing et al., 2021), and stochastic gradient Langevin dynamics (Pensia et al., 2018; Mou et al., 2018; Li et al., 2020; Negrea et al., 2019; Zhang et al., 2021; Farghly & Rebeschini, 2021; Wang et al., 2021a;b).

*Lead & corresponding author

Excess Risk Upper Bounds: GD vs SGD				
Algorithm	Iterations	Interpolation	Bound	β-Smooth Loss
GD **(this work)** $\eta_t = 1/2\beta$	$T = \sqrt{n}$	No	$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	Convex
SGD $\eta_t = 1/\sqrt{T}$ (Lei & Ying, 2020b)	$T = n$	No	$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	Convex
GD **(this work)** $\eta_t = 1/2\beta$	$T = n$	Yes	$\mathcal{O}\left(\frac{1}{n}\right)$	Convex
SGD $\eta_t = 1/2\beta$ (Lei & Ying, 2020b)	$T = n$	Yes	$\mathcal{O}\left(\frac{1}{n}\right)$	Convex
GD **(this work)** $\eta_t = 2/(\beta+\gamma)$	$T = \Theta(\log n)$	No	$\mathcal{O}\left(\frac{\sqrt{\log(n)}}{n}\right)$	γ-Strongly Convex (Objective)
SGD $\eta_t = 2/(t+t_0)\gamma$ (Lei & Ying, 2020b)	$T = \Theta(n)$	No	$\mathcal{O}\left(\frac{1}{n}\right)$	γ-Strongly Convex (Objective)

Table 1: Comparison of the excess risk bounds for the full-batch GD and SGD algorithms by Lei & Ying (2020b, Corollary 5 & Theorem 11). We denote by n the number of samples, T the total number of iterations, η_t the step size at time t, and by $\epsilon_{\mathbf{c}} \triangleq \mathbb{E}[R_S(W_S^*)]$ the interpolation error.

Even though many previous works consider stochastic training algorithms and some even suggest that stochasticity may be necessary (Hardt et al., 2016; Charles & Papailiopoulos, 2018) for good generalization, recent empirical studies have demonstrated that deterministic algorithms can indeed generalize well; see, e.g., (Hoffer et al., 2017; Geiping et al., 2022). In fact, Hoffer et al. (2017) showed empirically that for large enough number of iterations full-batch GD generalizes comparably to SGD. Similarly, Geiping et al. (2022) experimentally showed that strong generalization behavior is still observed in the absence of stochastic sampling. Such interesting empirical evidence reasonably raise the following question: "Are there problem classes for which deterministic training generalizes more efficiently than stochastic training?"

While prior works provide extensive analysis of the generalization error and excess risk of stochastic gradient methods, tight and path-dependent generalization error and excess risk guarantees in non-stochastic training (for general smooth losses) remain unexplored. Our main purpose in this work is to theoretically establish that full-batch GD indeed generalizes efficiently for general smooth losses. While SGD appears to generalize better than full-batch GD for non-smooth and Lipschitz convex losses (Bassily et al., 2020; Amir et al., 2021), non-smoothness seems to be problematic for efficient algorithmic generalization. In fact, tightness analysis on non-smooth losses (Bassily et al., 2020) shows that the generalization error bounds become vacuous for standard step-size choices. Our work shows that for general smooth losses, full-batch GD achieves either tighter stability and excess error rates (convex case), or equivalent rates (compared to SGD in the strongly convex setting) but with significantly shorter training horizon (strongly-convex objective).

2 Related Work and Contributions

Let n denote the number of available training samples (examples). Recent results (Lei & Tang, 2021; Zhou et al., 2022) on SGD provided bounds of order $\mathcal{O}(1/\sqrt{n})$ for Lipschitz and smooth nonconvex losses. Neu et al. (2021) also provided generalization bounds of order $\mathcal{O}(\eta T/\sqrt{n})$, with $T = \sqrt{n}$ and step-size $\eta = 1/T$ to recover the rate $\mathcal{O}(1/\sqrt{n})$. In contrast, we show that full-batch GD generalizes efficiently for appropriate choices of decreasing learning rate that guarantees faster convergence and smaller generalization error, simultaneously. Additionally, the generalization error involves an intrinsic dependence on the set of the stationary points and the initial point. Specifically,

Full-Batch Gradient Descent		
Step Size	Excess Risk	Loss
$\eta_t \leq c/\beta t, \forall c < 1$	$C\frac{T^c\sqrt{\log(T)+1}}{n} + \epsilon_{\text{opt}}$	Nonconvex
$\eta_t = 1/2\beta$	$C\left(\frac{T\epsilon_{\mathbf{c}}+1}{n} + \frac{1}{T}\right)$	Convex
$\eta_t = 1/2\beta$, $T = \sqrt{n}$	$C\frac{\epsilon_{\mathbf{c}}+1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)$	Convex
$\eta_t = 2/(\beta+\gamma)$	$C\left(\sqrt{\epsilon_{\mathbf{c}}}\frac{\sqrt{T}}{n} + \exp\left(\frac{-4T\gamma}{\beta+\gamma}\right)\right) + \mathcal{O}\left(\frac{1}{n^2}\right)$	γ-Strongly Convex
$\eta_t = 2/(\beta+\gamma)$ $T = \frac{(\beta+\gamma)\log n}{2\gamma}$	$C\sqrt{\epsilon_{\mathbf{c}}}\frac{\sqrt{\log(n)}}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)$	γ-Strongly Convex

Table 2: A list of excess risk bounds for the full-batch GD up to some constant factor $C > 0$. We denote the number of samples by n. "ϵ_{opt}" denotes the optimization error $\epsilon_{\text{opt}} \triangleq \mathbb{E}[R_S(A(S)) - R_S^*]$, T is the total number of iterations and $\epsilon_{\mathbf{c}} \triangleq \mathbb{E}[R_S(W_S^*)]$ is the model capacity (interpolation) error.

we show that full-batch GD with the decreasing learning rate choice of $\eta_t = 1/2\beta t$ achieves tighter bounds of the order $\mathcal{O}(\sqrt{T\log(T)}/n)$ (since $\sqrt{T\log(T)}/n \leq 1/\sqrt{n}$) for any $T \leq n/\log(n)$. In fact, $\mathcal{O}(\sqrt{T\log(T)}/n)$ essentially matches the rates in prior works (Hardt et al., 2016) for smooth and Lipschitz (and often bounded) loss, however we assume only smoothness at the expense of the $\sqrt{\log(T)}$ term. Further, for convex losses we show that full-batch GD attains tighter generalization error and excess risk bounds than those of SGD in prior works (Lei & Ying, 2020b), or similar rates in comparison with prior works that consider additional assumptions (Lipschitz or sub-Gaussian loss) (Hardt et al., 2016; Lugosi & Neu, 2022; Kozachkov et al., 2022). In fact, for convex losses and for a fixed step-size $\eta_t = 1/2\beta$, we show generalization error bounds of order $\mathcal{O}(T/n)$ for non-Lipschitz losses, while SGD bounds in prior work are of order $\mathcal{O}(\sqrt{T/n})$ (Lei & Ying, 2020b). As a consequence, full-batch GD attains improved generalization error rates by one order of error magnitude and appears to be more stable in the non-Lipschitz case, however tightness guarantees for non-Lipschitz losses remains an open problem.

Our results also establish that full-batch GD provably achieves efficient excess error rates through fewer number of iterations, as compared with the state-of-the-art excess error guarantees for SGD. Specifically, for convex losses with limited model capacity (non-interpolation), we show that with constant step size and $T = \sqrt{n}$, the excess risk is of the order $\mathcal{O}(1/\sqrt{n})$, while the SGD algorithm requires $T = n$ to achieve excess risk of the order $\mathcal{O}(1/\sqrt{n})$ (Lei & Ying, 2020b, Corollary 5.a).

For γ-strongly convex objectives, our analysis for full-batch GD relies on a *leave-one-out* γ_{loo}-strong convexity of the objective instead of the full loss function being strongly-convex. This property relaxes strong convexity, while it provides stability and generalization error guarantees that recover the convex loss setting when $\gamma_{loo} \to 0$. Prior work (Lei & Ying, 2020b, Section 6, Stability with Relaxed Strong Convexity) requires a Lipschitz loss, while the corresponding bound becomes infinity when $\gamma \to 0$, in contrast to the leave-one-out approach. Further, prior guarantees on SGD (Lei & Ying, 2020b, Theorem 11 and Theorem 12) often achieve the same rate of $\mathcal{O}(1/n)$, however with $T = \Theta(n)$ iterations (and a Lipschitz loss), in contrast with our full-batch GD bound that requires only $T = \Theta(\log n)$ iterations (at the expense of a $\sqrt{\log(n)}$ term)[1]. Finally, our approach does not require a projection step (in contrast to Hardt et al. (2016); Lei & Ying (2020b)) in the update rule and consequently avoids dependencies on possibly large Lipschitz constants.

[1]SGD naturally requires less computation than GD. However, the directional step of GD can be evaluated in parallel. As a consequence, for a strongly-convex objective GD would be more efficient than SGD (in terms of running time) if some parallel computation is available.

In summary, we show that for smooth nonconvex, convex and strongly convex losses, full-batch GD generalizes, which provides an explanation of its good empirical performance in practice (Hoffer et al., 2017; Geiping et al., 2022). We refer the reader to Table 1 for an overview and comparison of our excess risk bounds and those of prior work (on SGD). A more detailed presentation of the bounds appears in Table 2 (see also Appendix A, Table 3 and Table 4).

3 Problem Statement

Let $f(w, z)$ be the loss at the point $w \in \mathbb{R}^d$ for some example $z \in \mathcal{Z}$. Given a dataset $S \triangleq \{z_i\}_{i=1}^n$ of i.i.d samples z_i from an unknown distribution $\mathcal{D}$, our goal is to find the parameters w^* of a learning model such that $w^* \in \arg\min_w R(w)$, where $R(w) \triangleq \mathbb{E}_{Z\sim\mathcal{D}}[f(w, Z)]$ and $R^* \triangleq R(w^*)$. Since the distribution $\mathcal{D}$ is not known, we consider the empirical risk

$$R_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(w; z_i). \tag{1}$$

The corresponding empirical risk minimization (ERM) problem is to find $W_S^* \in \arg\min_w R_S(w)$ (assuming minimizers on data exist for simplicity) and we define $R_S^* \triangleq R_S(W_S^*)$. For a deterministic algorithm A with input S and output $A(S)$, the excess risk ϵ_{excess} is bounded by the sum of the generalization error ϵ_{gen} and the optimization error ϵ_{opt} (Hardt et al., 2016, Lemma 5.1), (Dentcheva & Lin, 2022)

$$\epsilon_{\text{excess}} \triangleq \mathbb{E}[R(A(S))] - R^* \leq \underbrace{\mathbb{E}[R(A(S)) - R_S(A(S))]}_{\epsilon_{\text{gen}}} + \underbrace{\mathbb{E}[R_S(A(S))] - \mathbb{E}[R_S(W_S^*)]}_{\epsilon_{\text{opt}}}. \tag{2}$$

For the rest of the paper we assume that the loss is smooth and non-negative. These are the only globally required assumptions on the loss function.

Assumption 1 (β-Smooth Loss) *The gradient of the loss function is β-Lipschitz*

$$\|\nabla_w f(w, z) - \nabla_u f(u, z)\|_2 \leq \beta\|w - u\|_2, \quad \forall z \in \mathcal{Z}. \tag{3}$$

Additionally, we define the interpolation error that will also appear in our results.

Definition 1 (Model Capacity/Interpolation Error) *Define* $\epsilon_{\mathbf{c}} \triangleq \mathbb{E}[R_S(W_S^*)]$.

In general $\epsilon_{\mathbf{c}} \geq 0$ (non-negative loss). If the model has sufficiently large capacity, then for almost every $S \in \mathbb{Z}^n$, it is true that $R_S(W_S^*) = 0$. Equivalently, it holds that $\epsilon_{\mathbf{c}} = 0$. In the next section we provide a general theorem for the generalization error that holds for any symmetric deterministic algorithm (e.g. full-batch gradient descent) and any smooth loss under memorization of the data-set.

4 Symmetric Algorithm and Smooth Loss

Consider the i.i.d random variables $z_1, z_2, \ldots, z_n, z_1', z_2', \ldots, z_n'$, with respect to an unknown distribution $\mathcal{D}$, the sets $S \triangleq (z_1, z_2, \ldots, z_n)$ and $S^{(i)} \triangleq (z_1, z_2, \ldots, z_i', \ldots, z_n)$ that differ at the ith random element. Recall that an algorithm is symmetric if the output remains unchanged under permutations of the input vector. Then (Bousquet & Elisseeff, 2002, Lemma 7) shows that for any $i \in \{1, \ldots, n\}$ and any symmetric deterministic algorithm A the generalization error is $\epsilon_{\text{gen}} = \mathbb{E}_{S^{(i)}, z_i}[f(A(S^{(i)}); z_i) - f(A(S); z_i)]$. Identically, we write $\epsilon_{\text{gen}} = \mathbb{E}[f(A(S^{(i)}); z_i) - f(A(S); z_i)]$, where the expectation is over the random variables $z_1, \ldots, z_n, z_1', \ldots, z_n'$ for the rest of the paper. We define the model parameters $W_t, W_t^{(i)}$ evaluated at time t with corresponding inputs S and $S^{(i)}$. For brevity, we also provide the next definition.

Definition 2 *We define the* expected output stability *as* $\epsilon_{\text{stab}(A)} \triangleq \mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]$ *and the* expected optimization error *as* $\epsilon_{\text{opt}} \triangleq \mathbb{E}[R_S(A(S)) - R_S(W_S^*)]$.

We continue by providing an upper bound that connects the generalization error with the expected output stability and the expected optimization error at the final iterate of the algorithm.

Theorem 3 (Generalization Error) *Let $f(\cdot\,;z)$ be non-negative β-smooth loss for any $z \in \mathcal{Z}$. For any symmetric deterministic algorithm $A(\cdot)$ the generalization error is bounded as*

$$|\epsilon_{\text{gen}}| \leq 2\sqrt{2\beta(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{stab}(A)}} + 2\beta\epsilon_{\text{stab}(A)}, \tag{4}$$

where $\epsilon_{\text{stab}(A)} \triangleq \mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]$. In the limited model capacity case it is true that $\epsilon_{\mathbf{c}}$ is positive (and independent of n and T) and $|\epsilon_{\text{gen}}| = \mathcal{O}(\sqrt{\epsilon_{\text{stab}(A)}})$.

We provide the proof of Theorem 3 in Appendix B.1. The generalization error bound in equation 4 holds for any symmetric algorithm and smooth loss. Theorem 3 consist the tightest variant of (Lei & Ying, 2020b, Theorem 2, b)) and shows that the expected output stability and a small expected optimization error at termination sufficiently provide an upper bound on the generalization error for smooth (possibly non-Lipschitz) losses. Further, the optimization error term ϵ_{opt} is always bounded and goes to zero (with specific known rates) in the cases of (strongly) convex losses. Under the interpolation condition the generalization error bound satisfies tighter bounds.

Corollary 4 (Generalization under Memorization) *If memorization of the training set is feasible under sufficiently large model capacity, then $\epsilon_{\mathbf{c}} = 0$ and consequently $|\epsilon_{\text{gen}}| \leq 2\sqrt{2\beta\epsilon_{\text{opt}}\epsilon_{\text{stab}(A)}} + 2\beta\epsilon_{\text{stab}(A)}$ and $|\epsilon_{\text{gen}}| = \mathcal{O}(\max\{\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{stab}(A)}}, \epsilon_{\text{stab}(A)}\})$.*

For a small number of iterations T the above error rate is equivalent with Theorem 3. For sufficiently large T the optimization error rate matches the expected output stability and provides a tighter rate (with respect to that of Theorem 3) of $|\epsilon_{\text{gen}}| = \mathcal{O}(\epsilon_{\text{stab}(A)})$.

Remark. As a byproduct of Theorem 3, one can show generalization and excess risk bounds for a uniformly μ-PL objective (Karimi et al., 2016) defined as $\mathbb{E}[\|\nabla R_S(w)\|_2^2] \geq 2\mu\mathbb{E}[R_S(w) - R_S^*]$ for all $w \in \mathbb{R}^d$. Let $\pi_S \triangleq \pi(A(S))$ be the projection of the point $A(S)$ to the set of the minimizers of $R_S(\cdot)$. Further, define the constant $\tilde{c} \triangleq \mathbb{E}[R_S(\pi_S) + R(\pi_S)]$. Then a bound on the excess risk is (the proof appears in Appendix D)

$$\epsilon_{\text{excess}} \leq \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}}} + \frac{8\sqrt{2\beta\epsilon_{\text{opt}}\epsilon_{\mathbf{c}}}}{\sqrt{\mu}} + \frac{16\tilde{c}\beta^2}{n^2\mu^2} + \frac{45\beta}{\mu}\epsilon_{\text{opt}}. \tag{5}$$

We note that a closely related bound has been shown in (Lei & Ying, 2020a). In fact, (Lei & Ying, 2020a, Therem 7) requires simultaneously the interpolation error to be zero ($\epsilon_{\mathbf{c}} = 0$) and an additional assumption, namely the inequality $\beta \leq n\mu/4$, to hold. However, if $\beta \leq n\mu/4$ and $\epsilon_{\mathbf{c}} = 0$ (interpolation assumption), then (Lei & Ying, 2020a, inequality (B.13), Proof of Theorem 1) implies that ($\mathbb{E}[R(\pi_S)] \leq 3\mathbb{E}[R_S(\pi_S)]$ and) the expected *population* risk at π_S is zero, i.e., $\mathbb{E}[R(\pi_S)] = 0$. Such a situation is apparently trivial since the population risk is zero at the empirical minimizer $\pi_S \in \arg\min R_S(\cdot)$.[2] On the other hand, if $\beta > n\mu/4$, these bounds become vacuous. A PL condition is interesting under the interpolation regime and since the PL is not uniform (with respect to the data-set) in practice (Liu et al., 2022), it is reasonable to consider similar bounds to that of 5 as trivial.

5 Full-Batch GD

In this section, we derive generalization error and excess risk bounds for the full-batch GD algorithm. We start by providing the definition of the expected path error ϵ_{path}, in addition to the optimization error ϵ_{opt}. These quantities will prominently appear in our analysis and results.

Definition 5 (Path Error) *For any β-smooth (possibly) nonconvex loss, learning rate η_t, and for any $i \in \{1, \ldots, n\}$, we define the* expected path error *as*

$$\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2]. \tag{6}$$

[2] In general, this occurs when $n \to \infty$, and the generalization error is zero. As a consequence, the excess risk becomes equals to the optimization error (see also (Lei & Ying, 2020a, Therem 7)), and the analysis becomes not interesting from generalization error prospective.

The ϵ_{path} term expresses the path-dependent quantity that appears in the generalization bounds in our results[3]. Additionally, as we show, the generalization error also depends on the average optimization error ϵ_{opt} (Theorem 3). A consequence of the dependence on ϵ_{opt}, is that full-batch GD generalizes when it reaches the neighborhoods of the loss minima. Essentially, the expected path error and optimization error replace bounds in prior works (Hardt et al., 2016; Kozachkov et al., 2022) that require a Lipschitz loss assumption to upper bound the gradients and substitute the Lipschitz constant with tighter quantities. Later we show the dependence of the expected output stability term in Theorem 3 with respect to the expected path error. Then through explicit rates for both ϵ_{path} and ϵ_{opt} we characterize the generalization error and excess risk.

5.1 Nonconvex Loss

We proceed with the average output stability and generalization error bounds for nonconvex smooth losses. Through a stability error bound, the next result connects Theorem 3 with the expected path error and the corresponding learning rate. Then we use that expression to derive generalization error bounds for the full-batch GD in the case of nonconvex losses.

Theorem 6 (Stability Error — Nonconvex Loss) *Assume that the (possibly) nonconvex loss $f(\cdot, z)$ is β-smooth for all $z \in \mathcal{Z}$. Consider the full-batch GD where T denotes the total number of iterates and η_t denotes the learning rate, for all $t \leq T+1$. Then for the outputs of the algorithm $W_{T+1} \equiv A(S)$, $W_{T+1}^{(i)} \equiv A(S^{(i)})$ it is true that*

$$\epsilon_{\text{stab}(A)} \leq \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2 . \tag{7}$$

The expected output stability in Theorem 6 is bounded by the product of the expected path error (Definition 5), a sum-product term ($\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2$) that only depends on the step-size and the term $4/n^2$ that provides the dependence on the sample complexity. In light of Theorem 3, and Theorem 6, we derive the generalization error of full-batch GD for smooth nonconvex losses.

Theorem 7 (Generalization Error — Nonconvex Loss) *Assume that the loss $f(\cdot, z)$ is β-smooth for all $z \in \mathcal{Z}$. Consider the full-batch GD where T denotes the total number of iterates, and the learning rate is chosen as $\eta_t \leq C/t \leq 1/\beta$, for all $t \leq T+1$. Let $\epsilon \triangleq \beta C < 1$ and $\bar{C}(\epsilon, T) \triangleq \min\{\epsilon + 1/2, \epsilon \log(eT)\}$. Then the generalization error of full-batch GD is bounded by*

$$\begin{aligned} |\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{2}}{n} \sqrt{(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{path}}} (eT)^{\epsilon} \bar{C}^{\frac{1}{2}}(\epsilon, T) + 8\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon}\bar{C}(\epsilon, T) \\ &\leq \frac{4\sqrt{3}(eT)^{\epsilon}}{n} \sqrt{(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{path}}} + 12\frac{(eT)^{2\epsilon}}{n^2}\epsilon_{\text{path}}. \end{aligned} \tag{8}$$

Additionally, by the definition of the expected path and optimization error, and from the descent direction of algorithm, we evaluate upper bounds on the terms ϵ_{path} and ϵ_{opt} and derive the next bound as a byproduct of Theorem 7.

Corollary 8 *The generalization error of full-batch GD in Theorem 7 can be further bounded as*

$$|\epsilon_{\text{gen}}| \leq \left(\frac{8\sqrt{3}}{n} \sqrt{\log(eT)} (eT)^{\epsilon} + \frac{48}{n^2} \log(eT)(eT)^{2\epsilon} \right) \mathbb{E}[R_S(W_1)]. \tag{9}$$

The inequality (8) in Theorem 7 shows the explicit dependence of the generalization error bound on the path-dependent error ϵ_{path} and the optimization error ϵ_{opt}. Note that during the training process the path-dependent error increases, and the optimization error decreases. Both terms ϵ_{path} and ϵ_{opt} may be upper bounded, to find the simplified (but potentially looser) bound appeared in Corollary 8. We prove Theorem 6, Theorem 7 and Corollary 8 in Appendix C. Finally, the generalization error in Corollary 8 matches bounds in prior work, including information theoretic bounds for the SGLD algorithm (Wang et al., 2021b, Corollary 1) (with fixed step-size), while our results do not require the sub-Gaussian loss assumption and show that similar generalization is achievable through deterministic training.

[3]Recall that the initial point W_1 may be chosen arbitrarily and uniformly over the dataset.

Remark. (Dependence on Stationary Points) Let W_1 be an arbitrary initial point (independent of S). Under mild assumptions (provided in (Lee et al., 2016)) GD convergences to (local) minimizers. Let W^*_{S,W_1} be the stationary point such $\lim_{T\uparrow\infty} A(S) \to W^*_{S,W_1}$. Then through the smoothness of the loss, we derive an alternative form of the generalization error bound in Theorem 3 that expresses the dependence of the generalization error with respect to the quality of the set of stationary points, i.e.,

$$|\epsilon_{\text{gen}}| \leq 4\sqrt{\beta\left(\beta\mathbb{E}[\|A(S) - W^*_{S,W_1}\|_2^2] + \mathbb{E}[R_S(W^*_{S,W_1})]\right)\epsilon_{\text{stab}(A)}} + 2\beta\epsilon_{\text{stab}(A)} \tag{10}$$

Inequality (10) provides a detailed bound that depends on the expected loss at the stationary point and the expected distance of the output from the stationary point, namely $\mathbb{E}[\|A(S) - W^*_{S,W_1}\|_2^2]$.

5.2 Convex Loss

In this section, we provide generalization error guarantees for GD on smooth convex losses. Starting from the stability of the output of the algorithm, we show that the dependence on the learning rate is weaker than that of the nonconvex case. That dependence and the fast convergence to the minimum guarantee tighter generalization error bounds than the general case of nonconvex losses in Section 5.1. The generalization error and the corresponding optimization error bounds provide an excess risk bound through the error decomposition (2). We refer the reader to Table 2 for a summary of the excess risk guarantees. We continue by providing the stability bound for convex losses.

Theorem 9 (Stability Error — Convex Loss) *Assume that the convex loss $f(\cdot, z)$ is β-smooth for all $z \in \mathcal{Z}$. Consider the full-batch GD where T denotes the total number of iterates and $\eta_t \leq 1/2\beta$ learning rate, for all $t \leq T+1$. Then for outputs of the algorithm $W_{T+1} \equiv A(S), W^{(i)}_{T+1} \equiv A(S^{(i)})$ it is true that*

$$\epsilon_{\text{stab}(A)} \leq \frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t \leq \frac{32\beta\sum_{t=1}^{T}\eta_t}{n^2}\left(\mathbb{E}[\|W_1 - W^*_S\|_2^2] + \epsilon_{\mathbf{c}}\sum_{t=1}^{T}\eta_t\right) \tag{11}$$

In the convex case, the expected output stability (inequality 11) is bounded by the product of the expected path error, the number of samples term $2/n^2$ and the accumulated learning rate. The inequality (11) gives $\epsilon_{\text{stab}(A)} = \mathcal{O}((\sum_{t=1}^{T}\eta_t/n)^2)$ and through Theorem 3 we find $|\epsilon_{\text{gen}}| = \mathcal{O}(\sum_{t=1}^{T}\eta_t/n)$. In contrast, stability guarantees for the SGD and non-Lipschitz losses in prior work (Lei & Ying, 2020b, Theorem 3, (4.4)) give $\epsilon_{\text{stab}(A)} = \mathcal{O}\left(\sum_{t=1}^{T}\eta_t^2/n\right)$ and $|\epsilon_{\text{gen}}| = \mathcal{O}(\sqrt{\sum_{t=1}^{T}\eta_t^2/n})$. As a consequence, GD guarantees are tighter than existing bounds of the SGD for non-Lipschitz losses, a variety of learning rates and $T \leq n$. For instance, for fixed $\eta_t = 1/\sqrt{T}$, the generalization error bound of GD is $|\epsilon_{\text{gen}}| = \mathcal{O}(\sqrt{T}/n)$ which is tighter than the corresponding bound of SGD, namely $|\epsilon_{\text{gen}}| = \mathcal{O}(1/\sqrt{n})$. Further, GD applies for much larger learning rates ($\eta_t = 1/\beta$), which provide not only tighter generalization error bound but also tighter excess risk guarantees than SGD as we later show. By combining Theorem 3 and Theorem 9, we show the next generalization error bound.

Theorem 10 (Generalization Error — Convex Loss) *Let the loss function $f(\cdot, z)$ be convex and β-smooth for all $z \in \mathcal{Z}$. Consider the full-batch GD where T denotes the total number of iterates. We chose the learning rate such that $\eta_t \leq 1/2\beta$, for all $t \leq T+1$. Then the generalization error of full-batch GD is bounded by*

$$|\epsilon_{\text{gen}}| \leq \frac{4\sqrt{2\beta\left(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}}\right)\epsilon_{\text{path}}}}{n}\sqrt{\sum_{t=1}^{T}\eta_t} + 8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t. \tag{12}$$

We provide the proof of Theorem 9 and Theorem 10 in Appendix E. Similar to the nonconvex case (Theorem 7), the bound in Theorem 10 shows the explicit dependence of the generalization error on the number of samples n, the path-dependent term ϵ_{path}, and the optimization error ϵ_{opt}, as well as the effect of the accumulated learning rate. From the inequality (12), we can proceed by deriving exact bounds on the optimization error and the accumulated learning rate, to find explicit expressions of the generalization error bound. Through Theorem 9, Theorem 10 (and Lemma 20 in Appendix E), we derive explicit generalization error bounds for certain choices of the learning rate. In fact, we consider the standard choice $\eta_t = 1/2\beta$ in the next result.

Theorem 11 (Generalization/Excess Error — Convex Loss) *Let the loss function $f(\cdot, z)$ be convex and β-smooth for all $z \in \mathcal{Z}$. If $\eta_t = 1/2\beta$ for all $t \in \{1, \ldots, T\}$, then*

$$|\epsilon_{\text{gen}}| \leq 8 \left(\frac{1}{n} + \frac{2T}{n^2} \right) \left(3\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\epsilon_{\mathbf{c}} \right), \tag{13}$$

and

$$\epsilon_{\text{excess}} \leq 8 \left(\frac{1}{n} + \frac{2T}{n^2} \right) \left(3\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\epsilon_{\mathbf{c}} \right) + \frac{3\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2]}{T}. \tag{14}$$

As a consequence, for $T = \sqrt{n}$ iterations the GD algorithm achieves $\epsilon_{\text{excess}} = \mathcal{O}(1/\sqrt{n})$. In contrast, SGD requires $T = n$ number of iterations to achieve $\epsilon_{\text{excess}} = \mathcal{O}(1/\sqrt{n})$ (Lei & Ying, 2020b, Corollary 5, a)). However, if $\epsilon_{\mathbf{c}} = 0$, then both algorithms have the same excess risk rate of $\mathcal{O}(1/n)$ through longer training with $T = n$ iterations. Finally, observe that the term $\mathbb{E}[\|W_1 - W_S^*]\|_2^2$ should be $\mathcal{O}(1)$ and independent of the parameters of interest (for instance n) to derive the aforementioned rates.

5.3 Strongly-Convex Objective

One common approach to enforce strong-convexity is through explicit regularization. In such a case both the objective $R_S(\cdot)$ the individual losses $f(\cdot; z)$ are strongly-convex. In other practical scenarios, the objective is often strongly-convex but the individual losses are not (Ma et al., 2018, Section 3). In this section, we show stability and generalization error guarantees that include the above cases by assuming a γ-strongly convex objective. We also show a property of full-batch GD that requires only a *leave-one-out* variant of the objective to be strongly-convex. If the objective $R_S(\cdot)$ is γ-strongly convex and the loss $f(\cdot; z)$ is β-smooth, then the leave-one-out function $R_{S^{-i}}(w) \triangleq \sum_{j=1, j\neq i}^{n} f(w; z_j)/n$ is γ_{loo}-strongly convex for all $i \in \{1, \ldots, n\}$ for some $\gamma_{loo} \leq \gamma$. Although γ_{loo} is slightly smaller than γ ($\gamma_{loo} = \max\{\gamma - \beta/n, 0\}$), our results reduce to the convex loss generalization and stability bounds when $\gamma_{loo} \to 0$. Further, the faster convergence also provides tighter bounds for the excess risk (see Table 2).

Theorem 12 (Stability Error — Strongly Convex Loss) *Assume that the loss $f(\cdot, z)$ is β-smooth for all $z \in \mathcal{Z}$ and that $R_S(\cdot)$ is γ-strongly convex. Consider the full-batch GD where T denotes the total number of iterates and $\eta_t \leq 2/(\beta + \gamma)$ denotes the learning rate, for all $t \leq T$. Then for outputs of the algorithm $W_{T+1} \equiv A(S)$, $W_{T+1}^{(i)} \equiv A(S^{(i)})$ it is true that*

$$\epsilon_{\text{stab}(A)} \leq \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1 - \eta_j \gamma_{loo}). \tag{15}$$

Specifically, if $\eta_t = 2/(\beta + \gamma)$, then

$$\epsilon_{\text{stab}(A)} \leq \frac{4\epsilon_{\text{path}}}{n^2} \min\left\{ \frac{1}{\gamma_{loo}}, \frac{2T}{\beta} \right\}. \tag{16}$$

By comparing the stability guarantee of Theorem 9 with Theorem 12, we observe that the learning rate dependent term (sum-product) is smaller than that of the convex case. While the dependence on expected path error (ϵ_{path}) is identical, we show (Appendix F) that the ϵ_{path} term is smaller in the strongly convex case. Additionally, Theorem 12 recovers the stability bounds of the convex loss case, when $\gamma_{loo} \to 0$ (and possibly $\gamma \to 0$). Similarly to the nonconvex and convex loss cases, Theorem 3 and the stability error bound in Theorem 12 provide the generalization error bound for strongly convex losses.

Theorem 13 (Generalization Error — Strongly Convex Loss) *Let the loss function $f(\cdot, z)$ β-smooth for all $z \in \mathcal{Z}$ and the objective $R_S(\cdot)$ be γ-strongly convex. Consider the full-batch GD where T denotes the total number of iterates. Let us set the learning rate to $\eta_t \leq 2/(\beta + \gamma)$, for all $t \leq T$. Then the generalization error of full-batch GD is bounded by*

$$|\epsilon_{\text{gen}}| \leq \frac{4\sqrt{(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{path}}}}{n} \sqrt{2\beta \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1 - \eta_j \gamma_{loo}) + 8\beta \frac{\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1 - \eta_j \gamma_{loo})}.$$

We prove Theorem 12 and Theorem 13 in Appendix F. Recall that the sum-product term in the inequality of Theorem 13 is smaller than the summation of the learning rates in Theorem 10. This fact together with the tighter optimization error bound provide a smaller excess risk than those of the convex losses. Similar to the convex loss setting, we use known optimization error guarantees of full-batch GD for strongly convex losses to derive explicit expressions of the generalization and excess risk bounds. By combining Theorem 13 and optimization and path error bounds (Lemma 22, Lemma 23 in Appendix F, and Lemma 15 in Appendix B.2), we derive our generalization error bound for fixed step size as follows in the next result.

Theorem 14 (Generalization/Excess — Strongly Convex Loss) *Let the objective function $R_S(\cdot)$ be γ-strongly convex and β-smooth by choosing some β-smooth loss $f(\cdot, z)$), not necessarily (strongly) convex for all $z \in \mathcal{Z}$. Define $m(\gamma_{loo}, T) \triangleq \beta T \min\{\beta/\gamma_{loo}, 2T\}/(\beta+\gamma)$ and $M(W_1) \triangleq \max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}$, and set the learning rate to $\eta_t = 2/(\beta+\gamma)$. Then the generalization error of the full-batch GD at the last iteration satisfies the inequality*

$$|\epsilon_{\text{gen}}| \leq \frac{8\sqrt{6}}{n}\left(\sqrt{M(W_1)} + \left(\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right) + \frac{4\sqrt{3}}{n}\sqrt{m(\gamma_{loo}, T)}\right)M(W_1)\right)\sqrt{m(\gamma_{loo}, T)}.$$

Additionally the optimization error (Lemma 23 in Appendix F) and the inequality (2) give the following excess risk

$$\epsilon_{\text{excess}} \leq \frac{8\sqrt{6}}{n}\left[\sqrt{\Delta_T} + \left(\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right) + \frac{4\sqrt{3}}{n}\Delta_T\right)\right] + \Lambda\exp\left(\frac{-4T\gamma}{\beta+\gamma}\right), \tag{17}$$

where $\Delta_T \triangleq \beta T M(W_1)\min\{\beta/\gamma_{loo}, 2T\}/(\beta+\gamma)$ and $\Lambda \triangleq \beta\mathbb{E}[\|W_1 - W_S^*\|_2^2]/2$.

Theorem 13 and Theorem 14 also recover the convex setting when $\gamma \to 0$ or $\gamma_{loo} \to 0$. Additionally, for $\gamma > 0$ and by setting the number of iterations as $T = (\beta/\gamma + 1)\log(n)/2$ and by defining the sequence $\mathrm{m}_{n,\gamma_{loo}} \triangleq \beta\min\{\beta/\gamma_{loo}, (\beta/\gamma+1)\log n\}/2\gamma$, the last inequality gives

$$|\epsilon_{\text{gen}}| \leq \frac{8\sqrt{6\log n}}{n}\left(\sqrt{\mathrm{M}(W_1)} + \frac{1+4\sqrt{3\mathrm{m}_{n,\gamma_{loo}}}}{n}\mathrm{M}(W_1)\right)\sqrt{\mathrm{m}_{n,\gamma_{loo}}}. \tag{18}$$

Finally, for $T = (\beta/\gamma + 1)\log(n)/2$ iterations it is true that

$$\epsilon_{\text{excess}} \leq \frac{8\sqrt{6\log n}}{n}\left(\sqrt{\Gamma_n} + \frac{1+4\sqrt{3}}{n}\Gamma_n\right) + \mathcal{O}\left(\frac{1}{n^2}\right), \tag{19}$$

where $\Gamma_n \triangleq \beta\mathrm{M}(W_1)\min\{\beta/\gamma_{loo}, (\beta/\gamma+1)\log n\}/2\gamma$ and as a consequence the excess risk is of the order $\mathcal{O}\left(\sqrt{\log(n)/n}\right)$. As a comparison, the SGD algorithm (Lei & Ying, 2020b, Theorem 12) requires $T = n$ number of iterations to achieve an excess risk of the order $\mathcal{O}(1/n)$, while full-batch GD achieves essentially the same rate with $T = (\beta/\gamma + 1)\log(n)/2$ iterations.

6 Conclusion

In this paper we developed generalization error and excess risk guarantees for deterministic training on smooth losses via the the full-batch GD algorithm. At the heart of our analysis is a sufficient condition for generalization, implying that, for every symmetric algorithm, average algorithmic output stability and a small expected optimization error at termination ensure generalization. By exploiting this sufficient condition, we explicitly characterized the generalization error in terms of the number of samples, the learning rate, the number of iterations, a path-dependent quantity and the optimization error at termination, further exploring the generalization ability of full-batch GD for different types of loss functions. More specifically, we derived explicit rates on the generalization error and excess risk for nonconvex, convex and strongly convex smooth (possibly non-Lipschitz) losses/objectives. Our theoretical results shed light on recent empirical observations indicating that full-batch gradient descent generalizes efficiently and that stochastic training procedures might not be necessary and in certain cases may even lead to higher generalization errors and excess risks.

REFERENCES

P. A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. doi: 10.1137/040605266. URL https://doi.org/10.1137/040605266.

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*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25033–25043. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d27b95cac4c27feb850aaa4070cc4675-Paper.pdf.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4381–4391. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Paper.pdf.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. URL https://www.jmlr.org/papers/v2/bousquet02a.html.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 745–754. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/charles18a.html.

Sourav Chatterjee. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022. URL https://arxiv.org/abs/2203.16462.

Darinka Dentcheva and Yang Lin. Bias reduction in sample-based optimization. *SIAM Journal on Optimization*, 32(1):130–151, 2022. doi: 10.1137/20M1326428. URL https://doi.org/10.1137/20M1326428.

Tyler Farghly and Patrick Rebeschini. Time-independent generalization bounds for SGLD in non-convex settings. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19836–19846. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/a4ee59dd868ba016ed2de90d330acb6a-Paper.pdf.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/8c01a75941549a705cf7275e41b21f0d-Paper.pdf.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/05a624166c8eb8273b8464e8d9cb5bd9-Paper.pdf.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1270–1279. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/feldman19a.html.

Jonas Geiping, Micah Goldblum, Phil Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ZBESeIUB5k.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5200–5209. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/qian19b.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/hardt16.html.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a5e0ff62be0b08456fc7f1e88812af3d-Paper.pdf.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016. URL https://doi.org/10.1007/978-3-319-46128-1_50.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate o(1/n). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5065–5076. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/286674e3082feb7e5afb92777e48821f-Paper.pdf.

Leo Kozachkov, Patrick M Wensing, and Jean-Jacques Slotine. Generalization in supervised learning through Riemannian contraction. *arXiv preprint arXiv:2201.06656*, 2022. URL https://arxiv.org/abs/2201.06656.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2815–2824. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kuzborskij18a.html.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/lee16.html.

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. doi: 10.1109/TPAMI.2021.3068154.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2020a. URL https://iclr.cc/virtual/2021/poster/3141.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International*

Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 5809–5819. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/lei20c.html.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21236–21246. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f3173935ed8ac4bf073c1bcd63171f8a-Paper.pdf.

Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *The Journal of Machine Learning Research*, 22(25):1–41, 2021a. URL http://jmlr.org/papers/v22/19-716.html.

Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of sgd for pairwise learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21216–21228. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper/2021/file/b1301141feffabac455e1f90a7de2054-Paper.pdf.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxxtgHKPS.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.009. URL https://www.sciencedirect.com/science/article/pii/S106352032100110X. Special Issue on Harmonic Analysis and Machine Learning.

Gabor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3524–3546. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/lugosi22a.html.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3325–3334. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ma18a.html.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv e-prints*, pp. arXiv–2006, 2020. URL https://arxiv.org/abs/2006.05610.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/mou18a.html.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf.

Yu Nesterov. Introductory lectures on convex programming, 1998. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.855&rep=rep1&type=pdf.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3526–3545. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/neu21a.html.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550, 2018. doi: 10.1109/ISIT.2018.8437571.

Ali Ramezani-Kebrya, Ashish Khisti, and Ben Liang. On the generalization of stochastic gradient descent with momentum. *arXiv preprint arXiv:2102.13653*, 2021. URL url={https://arxiv.org/abs/1809.04564}.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

Bohan Wang, Huishuai Zhang, Jieyu Zhang, Qi Meng, Wei Chen, and Tie-Yan Liu. Optimizing information-theoretical generalization bound via anisotropic noise of SGLD. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26080–26090. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/db2b4182156b2f1f817860ac9f409ad7-Paper.pdf.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of SGLD using properties of gaussian channels. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24222–24234. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper/2021/file/cb77649f5d53798edfa0ff40dae46322-Paper.pdf.

Puyu Wang, Liang Wu, and Yunwen Lei. Stability and generalization for randomized coordinate descent. *arXiv preprint arXiv:2108.07414*, 2021c. URL https://arxiv.org/abs/2108.07414.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26523–26535. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/df1f1d20ee86704251795841e6a9405a-Paper.pdf.

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*, 2021. URL https://par.nsf.gov/biblio/10366270.

Pan Zhou, Hanshu Yan, Xiaotong Yuan, Jiashi Feng, and Shuicheng Yan. Towards understanding why lookahead generalizes better than SGD and beyond. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27290–27304. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf.

Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of SGD in nonconvex optimization. *Machine Learning*, pp. 1–31, 2021b. URL https://link.springer.com/article/10.1007/s10994-021-06056-w.

Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of SGD in nonconvex optimization. *Machine Learning*, 111(1):345–375, 2022. URL https://doi.org/10.1007/s10994-021-06056-w.

Full-Batch Gradient Descent		
Step Size	Generalization Error	Loss
$\eta_t \leq C/\beta t, \forall C < 1$	$\frac{4e\sqrt{3}}{n} T^C \sqrt{(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{path}}} + \frac{12e^2}{n^2} T^{2C} \epsilon_{\text{path}}$	NC
$\eta_t \leq C/\beta t, \forall C < 1$	$48\left(\frac{\sqrt{\log(eT)}(eT)^C}{n} + \frac{\log(eT)(eT)^{2C}}{n^2}\right)\mathbb{E}[R_S(W_1)]$	NC
$\eta_t = 1/2\beta$	$8\left(\frac{1}{n} + \frac{2T}{n^2}\right)\left(3\beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2] + T\epsilon_{\mathbf{c}}\right)$	C
$\eta_t = 2/(\beta+\gamma)$	$\frac{8\sqrt{6}}{n}\left[\sqrt{\Delta_T} + \left(\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right) + \frac{4\sqrt{3}}{n}\Delta_T\right)\right]$	γ-SC

Table 3: A list of the generalization error bounds for the full-batch GD. We denote the number of samples by n. W_1 is the initial point of the algorithm, and W_S^* is a point in the set of minimizers of the objective. Also, "ϵ_{path}" denotes the expected path error $\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T} \eta_t \mathbb{E}[\lVert \nabla f(W_t, z_i)\rVert_2^2]$, "$\epsilon_{\text{opt}}$" denotes the optimization error $\epsilon_{\text{opt}} \triangleq \mathbb{E}[R_S(A(S)) - R_S^*]$, T is the total number of iterations and we define the model capacity (interpolation) error risk as $\epsilon_{\mathbf{c}} \triangleq \mathbb{E}[R_S(W_S^*)]$. Lastly, we define the constant $\mathrm{M}(W_1) \triangleq \max\left\{\beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2], \mathbb{E}[R_S(W_S^*)]\right\}$ and the terms $\Gamma_n \triangleq \beta \mathrm{M}(W_1) \min\left\{\beta/\gamma_{loo}, (\beta/\gamma + 1)\log n\right\}/2\gamma$, $\Delta_T \triangleq \beta T M(W_1) \min\left\{\beta/\gamma_{loo}, 2T\right\}/(\beta+\gamma)$. Lastly, "NC", "C" and "γ-SC" correspond to nonconvex, convex and γ-strongly convex objective, respectively.

A Summary of The Results

Herein, we present a summary of the generalization and excess risk bounds. The detailed expressions of the generalization and excess risk bounds appear in Table 3 and 4.

B Proofs

We provide the proofs of the results in these sections. We start by proving Theorem 3 and the bounds on the sum-product terms that appear in the stability error bounds, and then we continue with stability and generalization error guarantees, that we prove in parallel. We derive the excess risk bounds by applying the decomposition of the inequality (2).

B.1 Proof of Theorem 3

It is true that for any $i, j \in \{1, \ldots, n\}$

$$\mathbb{E}[f(A(S); z_i)] = \mathbb{E}[f(A(S); z_j)] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[f(A(S); z_k)] = \mathbb{E}[R_S(A(S))]. \tag{20}$$

We show equation 20 through the symmetry of the algorithm (at each iteration) and the fact that $\{z_i\}_{i=1}^n$ are identically distributed as follows. The random variables $\{z_i\}_{i=1}^n$ remain exchangeable.[4] The β-smooth property of $f(\cdot\,; z)$ for all $z \in \mathcal{Z}$ gives

$$f(A(S^{(i)}); z) - f(A(S); z) \leq \langle A(S^{(i)}) - A(S), \nabla f(A(S); z)\rangle + \frac{\beta\lVert A(S^{(i)}) - A(S)\rVert_2^2}{2}. \tag{21}$$

[4] $\mathbb{P}(z_1 = c_1, z_2 = c_2, \ldots, z_i = c_i, \ldots, z_j = c_j, \ldots, z_n = c_n, A(S) = \mathbf{w}) = \mathbb{P}(z_1 = c_1, z_2 = c_2, \ldots, z_i = c_j, \ldots, z_j = c_i, \ldots, z_n = c_n, A(S) = \mathbf{w})$ for any choice of the values $c_1, c_2, \ldots, c_n, \mathbf{w}$ and for any $i, j \in \{1, \ldots, n\}$.

Full-Batch Gradient Descent		
Step Size	Excess Risk	Loss
$\eta_t \leq C/\beta t$, $\forall C < 1$	$48\left(\frac{\sqrt{\log(eT)}(eT)^C}{n} + \frac{\log(eT)(eT)^{2C}}{n^2}\right)\mathbb{E}[R_S(W_1)] + \epsilon_{\text{opt}}$	NC
$\eta_t = 1/2\beta$	$\left(\frac{8}{n} + \frac{16T}{n^2}\right)\left(3\beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2] + T\epsilon_{\mathbf{c}}\right) + \frac{3\beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2]}{T}$	C
$\eta_t = 1/2\beta$, $T = \sqrt{n}$	$8\frac{\epsilon_{\mathbf{c}} + 3\beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2]}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)$	C
$\eta_t = 2/(\beta+\gamma)$	$\frac{8\sqrt{3}}{n}\left[\sqrt{\Delta_T} + \left(\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right) + \frac{4\sqrt{3}}{n}\Delta_T\right)\right] + \Lambda\exp\left(\frac{-4T\gamma}{\beta+\gamma}\right)$	γ-SC
$\eta_t = 2/(\beta+\gamma)$ $T = \frac{(\beta+\gamma)\log n}{2\gamma}$	$\frac{8\sqrt{3\log n}}{n}\left(\sqrt{\Gamma_n} + \frac{1+4\sqrt{3}}{n}\Gamma_n\right) + \mathcal{O}\left(\frac{1}{n^2}\right)$	γ-SC

Table 4: A list of excess risk bounds for the full-batch GD. We denote the number of samples by n. W_1 is the initial point of the algorithm, and W_S^* is a point in the set of minimizers of the objective. Also, "ϵ_{path}" denotes the expected path error $\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T}\eta_t\mathbb{E}[\lVert\nabla f(W_t, z_i)\rVert_2^2]$, "$\epsilon_{\text{opt}}$" denotes the optimization error $\epsilon_{\text{opt}} \triangleq \mathbb{E}[R_S(A(S)) - R_S^*]$, T is the total number of iterations and we define the model capacity (interpolation) error risk as $\epsilon_{\mathbf{c}} \triangleq \mathbb{E}[R_S(W_S^*)]$. Lastly, we define the constants $\Lambda \triangleq \beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2]/2$, $\mathrm{M}(W_1) \triangleq \max\left\{\beta\mathbb{E}[\lVert W_1 - W_S^*\rVert_2^2], \mathbb{E}[R_S(W_S^*)]\right\}$ and the terms $\Gamma_n \triangleq \beta\mathrm{M}(W_1)\min\left\{\beta/\gamma_{loo}, (\beta/\gamma+1)\log n\right\}/2\gamma$, $\Delta_T \triangleq \beta T M(W_1)\min\left\{\beta/\gamma_{loo}, 2T\right\}/(\beta+\gamma)$. Lastly, "NC", "C" and "γ-SC" correspond to nonconvex, convex and γ-strongly convex objective, respectively.

The expression $\epsilon_{\text{gen}} = \mathbb{E}[f(A(S^{(i)}); z_i) - f(A(S); z_i)]$ and the inequality (21) give

$$\epsilon_{\text{gen}} \leq \mathbb{E}\left[\langle A(S^{(i)}) - A(S), \nabla f(A(S); z_i)\rangle + \frac{\beta\lVert A(S^{(i)}) - A(S)\rVert_2^2}{2}\right] \tag{22}$$

We find an upper bound for the expectation of the inner product of the inequality (22) by applying Cauchy-Schwartz inequality as

$$\begin{aligned}
&\mathbb{E}\left[\langle A(S^{(i)}) - A(S), \nabla f(A(S); z_i)\rangle\right] \\
&\leq \mathbb{E}\left[\lVert A(S^{(i)}) - A(S)\rVert_2 \lVert\nabla f(A(S); z_i)\rVert_2\right] \qquad (23)\\
&\leq \sqrt{\epsilon_{\text{stab}(A)}\mathbb{E}\left[\lVert\nabla f(A(S); z_i)]\rVert_2^2\right]}, \qquad (24)
\end{aligned}$$

here we use the inequalities $\langle a, b\rangle \leq \lVert a\rVert_2\lVert b\rVert_2$ and $\mathbb{E}^2[XY] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$ to derive the bounds in 23 and 24 respectively. By combining the inequalities 22 and 24 we find that for any $i \in \{1, \ldots, n\}$ it is true that

$$\epsilon_{\text{gen}} \leq \sqrt{\epsilon_{\text{stab}(A)}\mathbb{E}\left[\lVert\nabla f(A(S); z_i)\rVert_2^2\right]} + \frac{\beta}{2}\epsilon_{\text{stab}(A)}. \tag{25}$$

To find an upper bound for the $|\epsilon_{\text{gen}}|$, we also need an upper bound for negative of ϵ_{gen}, namely $\mathbb{E}[f(A(S); z_i) - f(A(S^{(i)}); z_i)] = -\epsilon_{\text{gen}}$. Note that by the same argument

$$-\epsilon_{\text{gen}} \leq \sqrt{\mathbb{E}\left[\lVert A(S) - A(S^{(i)})\rVert_2^2\right]\mathbb{E}\left[\lVert\nabla f(A(S^{(i)}); z_i)]\rVert_2^2\right]} + \frac{\beta}{2}\mathbb{E}[\lVert A(S) - A(S^{(i)})\rVert_2^2]. \tag{26}$$

Then we find an upper bound on $\mathbb{E}[\lVert\nabla f(A(S^{(i)}); z_i)]\rVert_2^2]$ as follows

$$\mathbb{E}\left[\lVert\nabla f(A(S^{(i)}); z_i)]\rVert_2^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\|\nabla f(A(S^{(i)}); z_i) - \nabla f(A(S); z_i) + \nabla f(A(S); z_i)\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f(A(S^{(i)}); z_i) - \nabla f(A(S); z_i)]\|_2^2 + \|\nabla f(A(S); z_i)\|_2^2\right] \\
&\leq 2\beta^2\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2] + 2\mathbb{E}[\|\nabla f(A(S); z_i)]\|_2^2]. \qquad (27)
\end{aligned}
$$

The inequality 27 holds because of the β-smoothness of the loss. Additionally,

$$
\begin{aligned}
&\sqrt{2\beta^2\mathbb{E}^2\left[\|A(S) - A(S^{(i)})\|_2^2\right] + 2\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right]\mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} \\
&\leq \sqrt{2\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right]\mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} + \sqrt{2}\beta\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]. \qquad (28)
\end{aligned}
$$

We combine the inequalities 26, 27 and 28 to find

$$
-\epsilon_{\text{gen}} \leq \sqrt{2\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right]\mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} + 2\beta\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]. \qquad (29)
$$

Finally, through the inequalities 25 and 29 we find

$$
|\epsilon_{\text{gen}}| \leq \sqrt{2\epsilon_{\text{stab}(A)}\mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} + 2\beta\epsilon_{\text{stab}(A)} \qquad (30)
$$

We use the self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ (Srebro et al., 2010, Lemma 3.1), to show

$$
\|\nabla f(A(S); z_i)\|_2^2 \leq 4\beta f(A(S); z_i). \qquad (31)
$$

The last display, Assumption 1 and equation 20 give

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2] \leq 4\beta\mathbb{E}[f(A(S); z_i)] &= 4\beta\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(A(S); z_i)] \\
&= 4\beta\mathbb{E}[R_S(A(S))] \qquad (32) \\
&= 4\beta\left(\mathbb{E}[R_S(A(S))] - \mathbb{E}[R_S(W_S^*)] + \mathbb{E}[R_S(W_S^*)]\right) \\
&= 4\beta\left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)]\right). \qquad (33)
\end{aligned}
$$

We combine the inequalities 30, 33 and the Definition 1 to find

$$
|\epsilon_{\text{gen}}| \leq 2\sqrt{2\beta\left(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}}\right)\epsilon_{\text{stab}(A)}} + 2\beta\epsilon_{\text{stab}(A)}. \qquad (34)
$$

The last inequality gives the bound on the generalization error and completes the proof. □

B.2 Sum Product Terms in the Stability Bounds

Herein we show a lemma for the sum product terms associated with learning rate in Theorem 6 and Theorem 12. Then we will apply that lemma to derive the corresponding stability error bounds.

Lemma 15 *The following are true:*

- *If $\eta_t = C \leq 2/(\beta + \gamma)$, then*

$$
\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1 - \eta_j\gamma) = \frac{1 - (1 - C\gamma)^T}{\gamma}, \qquad (35)
$$

- *If $\eta_t = C/t \leq 2/(\beta + \gamma)$, for some $C \geq 2/\gamma$ for $t \geq 1 + \lceil\frac{\beta}{\gamma}\rceil$ and $\eta_t = C'/t \leq 2/(\beta + \gamma)$ for some $C' < 2/(\gamma + \beta)$ for $t \leq \lceil\frac{\beta}{\gamma}\rceil$, then*

$$
\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1 - \frac{\eta_j\gamma}{2}\right) \leq C\log\left(e^2\lceil\beta/\gamma\rceil\right). \qquad (36)
$$

- *If $\eta_t \leq C/t < 2/\beta$, then*

$$
\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1 + \beta\eta_j)^2 \leq Ce^{2C\beta}T^{2C\beta}\min\left\{1 + \frac{1}{2C\beta}, \log(eT)\right\}. \qquad (37)
$$

Proof.

- If $\eta_t = C \leq 2/(\beta+\gamma)$ then

$$\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1-\eta_j\gamma) = C\sum_{t=1}^{T}(1-C\gamma)^{T-t} = C\,(1-C\gamma)^T \sum_{t=1}^{T}(1-C\gamma)^{-t}$$
$$= C\frac{1-(1-C\gamma)^T}{C\gamma} = \frac{1-(1-C\gamma)^T}{\gamma},$$

- If $\eta_t = C/t \leq 2/(\beta+\gamma)$, for some $C \geq 2/\gamma$ for $t \geq 1+\lceil\frac{\beta}{\gamma}\rceil$ and $\eta_t = C'/t \leq 2/(\beta+\gamma)$ for some $C' < 2/(\gamma+\beta)$ for $t \leq \lceil\frac{\beta}{\gamma}\rceil$ then

$$\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1-\frac{\eta_j\gamma}{2}\right)$$
$$\leq \sum_{t=1}^{\lceil\frac{\beta}{\gamma}\rceil} \frac{C'}{t} \prod_{j=t+1}^{T}\left(1-\frac{C'\gamma}{2j}\right) + \sum_{t=1+\lceil\frac{\beta}{\gamma}\rceil}^{T} \frac{C}{t}\prod_{j=t+1}^{T}\left(1-\frac{1}{j}\right)$$
$$= \sum_{t=1}^{\lceil\frac{\beta}{\gamma}\rceil} \frac{C'}{t} \prod_{j=t+1}^{T}\left(1-\frac{C'\gamma}{2j}\right) + \sum_{t=1+\lceil\frac{\beta}{\gamma}\rceil}^{T} \frac{C}{t}\frac{t}{T} \leq \sum_{t=1}^{\lceil\frac{\beta}{\gamma}\rceil}\frac{C'}{t} + C\frac{\left[T-\lceil\frac{\beta}{\gamma}\rceil\right]_+}{T}$$
$$\leq C\left(1+\log(\lceil\beta/\gamma\rceil) + \left[1-\lceil\frac{\beta}{T\gamma}\rceil\right]_+\right) \leq C\log\left(e^2\lceil\beta/\gamma\rceil\right).$$

- If $\eta_t \leq C/t \leq 2/\beta$, then

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2 &= \sum_{t=1}^{T}\frac{C}{t}\prod_{j=t+1}^{T}\left(1+\beta\frac{C}{j}\right)^2 \\
&\leq \sum_{t=1}^{T}\frac{C}{t}\prod_{j=t+1}^{T}\exp\left(2\beta\frac{C}{j}\right) \\
&= \sum_{t=1}^{T}\frac{C}{t}\exp\left(2\beta\sum_{j=t+1}^{T}\frac{C}{j}\right) \\
&\leq \sum_{t=1}^{T}\frac{C}{t}\exp\left(2C\beta\left(\log(T)+1-\log(t+1)\right)\right) \\
&= Ce^{2C\beta}T^{2C\beta}\sum_{t=1}^{T}\frac{1}{t}\frac{1}{(t+1)^{2C\beta}} \\
&\leq Ce^{2C\beta}T^{2C\beta}\sum_{t=1}^{T}\frac{1}{t}\frac{1}{(t+1)^{2C\beta}} \\
&\leq Ce^{2C\beta}T^{2C\beta}\sum_{t=1}^{T}\frac{1}{t^{1+2C\beta}} \qquad (38)\\
&= Ce^{2C\beta}T^{2C\beta}\left(1+\sum_{t=2}^{T}\frac{1}{t^{1+2C\beta}}\right) \\
&\leq Ce^{2C\beta}T^{2C\beta}\left(1+\int_1^T\frac{1}{x^{1+2C\beta}}dx\right)
\end{aligned}$$

$$
\begin{aligned}
&= Ce^{2C\beta}T^{2C\beta}\left(1+\frac{1}{2C\beta}\left(1-T^{-2C\beta}\right)\right)\\
&= Ce^{2C\beta}T^{2C\beta}\left(1+\frac{1}{2C\beta}\right)-C\frac{e^{2C\beta}}{2\beta}\\
&\leq Ce^{2C\beta}T^{2C\beta}\left(1+\frac{1}{2C\beta}\right), \qquad (39)
\end{aligned}
$$

additionally $\sum_{t=1}^{T} 1/t \leq \log(eT)$, thus the term in the inequality 38 may be upper bounded by $Ce^{2C\beta}T^{2C\beta}\log(eT)$ for any $T \in \mathbb{N}$, and we conclude that

$$
\sum_{t=1}^{T}\eta_t \prod_{j=t+1}^{T}(1+\beta\eta_j)^2 \leq Ce^{2C\beta}T^{2C\beta}\min\left\{1+\frac{1}{2C\beta}, \log(eT)\right\}. \qquad (40)
$$

The last inequality completes the proof. □

In the next section we prove the stability and generalization error bounds for nonconvex losses.

C Nonconvex Loss: Proof of Theorem 6 & Theorem 7

Let $z_1, z_2, \ldots, z_i, \ldots, z_n, z_i'$ be i.i.d. random variables, define $S \triangleq (z_1, z_2, \ldots, z_i, \ldots, z_n)$ and $S^{(i)} \triangleq (z_1, z_2, \ldots, z_i', \ldots, z_n)$, $W_1 = W_1'$. The updates for any $t \geq 1$ are

$$
W_{t+1} = W_t - \frac{\eta_t}{n}\sum_{j=1}^{n}\nabla f(W_t, z_j), \qquad (41)
$$

$$
W_{t+1}^{(i)} = W_t^{(i)} - \frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\nabla f(W_t^{(i)}, z_j) - \frac{\eta_t}{n}\nabla f(W_t^{(i)}, z_i'). \qquad (42)
$$

Then for any $t \geq 1$, we derive the stability recursion as

$$
\begin{aligned}
&\|W_{t+1} - W_{t+1}^{(i)}\|_2\\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n}\left\|\sum_{j=1, j\neq i}^{n}\left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\right)\right\|_2\\
&\qquad + \frac{\eta_t}{n}\|\nabla f(W_t, z_i) - \nabla f(W_t^{(i)}, z_i')\|_2\\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\|\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\|_2\\
&\qquad + \frac{\eta_t}{n}\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right)\\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t(n-1)}{n}\beta\|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n}\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right) \qquad (43)\\
&= \left(1+\frac{n-1}{n}\beta\eta_t\right)\|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n}\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right), \qquad (44)
\end{aligned}
$$

inequality 43 comes from the smoothness of the loss. Then by solving the recursion we find

$$
\begin{aligned}
&\|W_{T+1} - W_{T+1}^{(i)}\|_2\\
&\leq \frac{1}{n}\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right)\prod_{j=t+1}^{T}\left(1+\frac{n-1}{n}\beta\eta_j\right)\\
&\leq \frac{1}{n}\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right)\prod_{j=t+1}^{T}(1+\beta\eta_j)
\end{aligned}
$$

$$\leq \frac{1}{n}\sqrt{\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2+\|\nabla f(W_t^{(i)},z_i')\|_2\right)^2\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2}$$
$$\leq \frac{\sqrt{2}}{n}\sqrt{\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2^2+\|\nabla f(W_t^{(i)},z_i')\|_2^2\right)\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2}.$$

The last display gives

$$\|W_{T+1}-W_{T+1}^{(i)}\|_2^2 \leq \frac{2}{n^2}\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2^2+\|\nabla f(W_t^{(i)},z_i')\|_2^2\right)\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2,$$

and by taking the expectation we find

$$\begin{aligned}
&\mathbb{E}[\|W_{T+1}-W_{T+1}^{(i)}\|_2^2]\\
&\leq \frac{2}{n^2}\sum_{t=1}^{T}\eta_t\left(\mathbb{E}[\|\nabla f(W_t,z_i)\|_2^2]+\mathbb{E}[\|\nabla f(W_t^{(i)},z_i')\|_2^2]\right)\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2\\
&\leq \frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2. \qquad (45)
\end{aligned}$$

We evaluate the summation of the products in the inequality 45. Lemma 15 under the choice of decreasing learning rate $\eta_t \leq C/t \leq 2/\beta$ shows that

$$\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2 \leq Ce^{2C\beta}T^{2C\beta}\min\left\{1+\frac{1}{2C\beta},\log(eT)\right\}. \qquad (46)$$

Through the inequalities 45, 46 and Theorem 3, we derive the bound on the generalization error as

$$\begin{aligned}
&|\epsilon_{\text{gen}}|\\
&\leq 2\sqrt{2\beta(\epsilon_{\text{opt}}+\epsilon_{\mathbf{c}})\epsilon_{\text{stab}(A)}}+2\beta\epsilon_{\text{stab}(A)}\\
&\leq \frac{4}{n}\sqrt{2\beta(\epsilon_{\text{opt}}+\epsilon_{\mathbf{c}})\epsilon_{\text{path}}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1+\beta\eta_j)^2\\
&\leq \frac{4}{n}\sqrt{2C\beta(\epsilon_{\text{opt}}+\epsilon_{\mathbf{c}})\epsilon_{\text{path}}}e^{C\beta}T^{C\beta}\min\left\{1+\frac{1}{2C\beta},\log(eT)\right\}^{\frac{1}{2}}\\
&\quad+8C\beta\frac{\epsilon_{\text{path}}}{n^2}e^{2C\beta}T^{2C\beta}\min\left\{1+\frac{1}{2C\beta},\log(eT)\right\}
\end{aligned}$$

Under the choice $\eta_t \leq C/t < 1/\beta$ for all t, we choose $C < 1/\beta$, further we define $\epsilon \triangleq \beta C < 1$, and $\bar{C}(\epsilon,T) \triangleq \min\{\epsilon+1/2, \epsilon\log(eT)\}$ to get

$$\begin{aligned}
|\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{2}}{n}\sqrt{(\epsilon_{\text{opt}}+\epsilon_{\mathbf{c}})\epsilon_{\text{path}}}(eT)^{\epsilon}\bar{C}^{\frac{1}{2}}(\epsilon,T)+8\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon}\bar{C}(\epsilon,T)\\
&\leq \frac{4\sqrt{3}}{n}\sqrt{(\epsilon_{\text{opt}}+\epsilon_{\mathbf{c}})\epsilon_{\text{path}}}(eT)^{\epsilon}+12\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon}. \qquad (47)
\end{aligned}$$

The last inequality provide the generalization error bound and completes the proof. □

Next we derive upper bounds on expected path error ϵ_{path} and optimization error ϵ_{opt}, to show an alternative expression of the generalization error inequality 47. We continue by proving the proof of Corollary 8.

C.1 Proof of Corollary 8.

The self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ (Srebro et al., 2010, Lemma 3.1) gives $\|\nabla f(W_t, z_i)\|_2^2 \leq 4\beta f(W_t, z_i)$. By taking expectation, and through the Assumption 1 and the equation 20 we find

$$\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \leq 4\beta \mathbb{E}[f(W_t, z_i)] = 4\beta \mathbb{E}[R_S(W_t)]. \tag{48}$$

The definition of ϵ_{path} (Definition 5), and the decreasing learning rate ($\eta_t = C/t < 1/\beta t$) give

$$\begin{aligned}
\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] &\leq 4\beta \sum_{t=1}^{T} \eta_t \mathbb{E}[R_S(W_t)] \\
&\leq 4\beta \mathbb{E}[R_S(W_1)] \sum_{t=1}^{T} \eta_t \qquad (49) \\
&< 4\mathbb{E}[R_S(W_1)] \sum_{t=1}^{T} \frac{1}{t} \\
&\leq 4\mathbb{E}[R_S(W_1)] \log(eT), \qquad (50)
\end{aligned}$$

and the inequality 49 holds since the learning rate $\eta_t < 2/\beta$ guarantees descent at each iteration. Similarly, $\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}} \leq \mathbb{E}[R_S(W_1)]$. The last inequality together with the inequalities 50 and 47 give

$$\begin{aligned}
|\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{3}}{n} \sqrt{(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{path}}} (eT)^{\epsilon} + 12 \frac{\epsilon_{\text{path}}}{n^2} (eT)^{2\epsilon} \\
&\leq \left(\frac{8\sqrt{3}}{n} \sqrt{\log(eT)} (eT)^{\epsilon} + \frac{48}{n^2} \log(eT) (eT)^{2\epsilon} \right) \mathbb{E}[R_S(W_1)].
\end{aligned}$$

The last inequality provides the bound of the corollary.

D PL Objective

Herein we provide the proofs of the results associated with the PL condition on the objective. We start by proving an upper bound on the average output stability. Then by combining Lemma 16 and Theorem 3 we derive generalization error bounds for symmetric algorithms and smooth losses, as well as the generalization error bound of the full-batch GD under the PL condition. A similar proof technique of the next lemma also appears in prior work by Lei et al. (Lei & Ying, 2020a, Proof of Lemma B.2).

Lemma 16 *Let the loss function $f(\cdot; z)$ be non-negative, nonconvex and β-smooth for all $z \in \mathcal{Z}$. Further, let the objective be μ-PL, $\mathbb{E}[\|\nabla R_S(w)\|_2^2] \geq 2\mu \mathbb{E}[R_S(w) - R_S^*]$ for all $w \in \mathbb{R}^d$. Then for any algorithm it is true that*

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \leq \frac{16}{\mu} \epsilon_{\text{opt}} + \frac{8\beta}{n^2 \mu^2} \left(\mathbb{E}\left[R_S(\pi_S)\right] + \mathbb{E}[R(\pi_S)] \right). \tag{51}$$

Proof. Define the projection $\pi_{S^{(i)}} \triangleq \pi(A(S^{(i)}))$ of the point $A(S^{(i)})$ to the set of the minimizers of $R_{S^{(i)}}(\cdot)$, and the similarly the projection $\pi_S \triangleq \pi(A(S))$ of the point $A(S)$ to the set of the minimizers of $R_S(\cdot)$. Then

$$\begin{aligned}
&\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \\
&\leq 4\mathbb{E}[\|A(S^{(i)}) - \pi_{S^{(i)}}\|_2^2] + 4\mathbb{E}[\|A(S) - \pi_S\|_2^2] + 2\mathbb{E}[\|\pi_{S^{(i)}} - \pi_S\|_2^2] \\
&\leq \frac{8}{\mu} \mathbb{E}[R_{S^{(i)}}(A(S^{(i)})) - R_{S^{(i)}}^*] + \frac{8}{\mu} \mathbb{E}[R_S(A(S)) - R_S^*] + 2\mathbb{E}[\|\pi_{S^{(i)}} - \pi_S\|_2^2] \qquad (52) \\
&= \frac{16}{\mu} \epsilon_{\text{opt}} + 2\mathbb{E}[\|\pi_{S^{(i)}} - \pi_S\|_2^2] \\
&\leq \frac{16}{\mu} \epsilon_{\text{opt}} + \frac{4}{\mu} \left(\mathbb{E}[R_S(\pi_{S^{(i)}})] - \mathbb{E}[R_S(\pi_S)] \right), \qquad (53)
\end{aligned}$$

the inequalities 52 and 53 come from the quadratic growth (Karimi et al., 2016). Recall that, the PL condition on the objective gives

$$\frac{1}{2\mu}\mathbb{E}[\|\nabla R_S(\pi_{S^{(i)}})\|_2^2] \geq \mathbb{E}[R_S(\pi_{S^{(i)}}) - R_S(\pi_S)]. \tag{54}$$

We combine the inequalities 53 and 54 to find

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{2}{\mu^2}\mathbb{E}[\|\nabla R_S(\pi_{S^{(i)}})\|_2^2]. \tag{55}$$

Also, it is true that

$$\begin{aligned}
\|\nabla R_S(\pi_{S^{(i)}})\|_2^2 &= \|\nabla R_{S^{(i)}}(\pi_{S^{(i)}}) - \frac{1}{n}\nabla f(\pi_{S^{(i)}}; z_i') + \frac{1}{n}\nabla f(\pi_{S^{(i)}}; z_i)\|_2^2 \\
&= \frac{2}{n^2}\|\nabla f(\pi_{S^{(i)}}; z_i')\|_2^2 + \frac{2}{n^2}\|\nabla f(\pi_{S^{(i)}}; z_i)\|_2^2 && (56)\\
&\leq \frac{4\beta}{n^2} f(\pi_{S^{(i)}}; z_i') + \frac{4\beta}{n^2} f(\pi_{S^{(i)}}; z_i), && (57)
\end{aligned}$$

equation 56 holds because $\nabla R_{S^{(i)}}(\pi_{S^{(i)}}) = 0$, the inequality 57 holds for nonnegative losses (Srebro et al., 2010, (Lemma 3.1). Through inequality57 we find,

$$\begin{aligned}
\mathbb{E}[\|\nabla R_S(\pi_{S^{(i)}})\|_2^2] &\leq \frac{4\beta}{n^2}\mathbb{E}[f(\pi_{S^{(i)}}; z_i')] + \frac{4\beta}{n^2}\mathbb{E}[f(\pi_{S^{(i)}}; z_i)] && (58)\\
&= \frac{4\beta}{n^2}\mathbb{E}[f(\pi_S; z_i)] + \frac{4\beta}{n^2}\mathbb{E}[f(\pi_S; z_i')], && (59)
\end{aligned}$$

and the last equality holds because z_i, z_i' are exchangeable. We combine the inequalities 55 and 59 to find

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] &\leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(\pi_S; z_i)] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(\pi_S; z_i')\right) && (60)\\
&= \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\mathbb{E}\left[R_S(\pi_S)\right] + \mathbb{E}[R(\pi_S)]\right). && (61)
\end{aligned}$$

Since $\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] = \mathbb{E}[\|A(S^{(j)}) - A(S)\|_2^2]$ for any $i, j \in \{1, \ldots, n\}$, we conclude that for any $i \in \{1, \ldots, n\}$

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\mathbb{E}\left[R_S(\pi_S)\right] + \mathbb{E}[R(\pi_S)]\right). \tag{62}$$

The last inequality provides the bound on the expected stability and completes the proof. □

Corollary 17 *Let $\pi_S \triangleq \pi(A(S))$ be the projection of the point $A(S)$ to the set of the minimizers of $R_S(\cdot)$. Further, define the constant $\tilde{c} \triangleq \mathbb{E}[R_S(\pi_S) + R(\pi_S)]$. For any symmetric algorithm, non-negative β-smooth loss function $f(\cdot; z)$ for all $z \in \mathcal{Z}$, μ-PL objective and $\mathbb{E}[R_S^*] = 0$, it is true that*

$$|\epsilon_{\text{gen}}| \leq \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} + \frac{44\beta}{\mu}\epsilon_{\text{opt}}. \tag{63}$$

Further, define the constant $c \triangleq 44\max\{\mathbb{E}[R_S(\pi_S) + R(\pi_S)], \mathbb{E}[R_S(W_1) - R_S^*]\}$. Then the generalization error of the full-batch GD with step-size choice $\eta_t = 1/\beta$ and T total number of iterations is bounded as follows

$$|\epsilon_{\text{gen}}| \leq \frac{c\beta}{\mu}\frac{\left(1 - \frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{c\beta^2}{n^2\mu^2} + \frac{c\beta}{\mu}\left(1 - \frac{\mu}{\beta}\right)^T. \tag{64}$$

Proof. We define the constant $\tilde{c} \triangleq \mathbb{E}[R_S(\pi_S) + R(\pi_S)]$ apply Theorem 3 and Lemma 16 to find

$$\begin{aligned}
&|\epsilon_{\text{gen}}| \\
&\leq 2\sqrt{2\beta(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})\epsilon_{\text{stab}(A)}} + 2\beta\epsilon_{\text{stab}(A)} \\
&\leq \left(\frac{8}{\sqrt{\mu}}\sqrt{\epsilon_{\text{opt}}} + \frac{4\sqrt{2\beta\tilde{c}}}{n\mu}\right)\sqrt{2\beta(\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}})} + \frac{32\beta}{\mu}\epsilon_{\text{opt}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} \\
&\leq \frac{8\sqrt{2\beta\epsilon_{\text{opt}}}}{\sqrt{\mu}}\sqrt{\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}}} + \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}}} + \frac{32\beta}{\mu}\epsilon_{\text{opt}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} \\
&\leq \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}} + \epsilon_{\mathbf{c}}} + \frac{8\sqrt{2\beta\epsilon_{\text{opt}}\epsilon_{\mathbf{c}}}}{\sqrt{\mu}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} + \frac{44\beta}{\mu}\epsilon_{\text{opt}}. \qquad (65)
\end{aligned}$$

The last inequality completes the proof. □

E Convex Loss: Proof of Theorem 9 and Theorem 10.

We start by proving the non-expansive property of the stability iterates for the case of β-smooth convex loss. Then we continue with the proof of the stability generalization error.

Lemma 18 *Let the gradient of the loss be β-Lipschitz for all $z \in \mathcal{Z}$. If the loss function is convex and $\eta_t < 2/\beta$, then for any $t \leq T+1$ the updates $W_t, W_t^{(i)}$ satisfy the next inequality*

$$\left\| W_t - W_t^{(i)} - \frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\right)\right\|_2^2 \leq \|W_t - W_t^{(i)}\|_2^2. \qquad (66)$$

Proof. By the definition of β-Lipschitz gradients and triangle inequality, it is true that

$$\|\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\|_2 \leq \beta\|W_t - W_t^{(i)}\|_2 \implies \qquad (67)$$

$$\|\sum_{j\in\mathcal{J}}\nabla f(W_t, z_j) - \sum_{j\in\mathcal{J}}\nabla f(W_t^{(i)}, z_j)\|_2 \leq \beta|\mathcal{J}|\|W_t - W_t^{(i)}\|_2. \qquad (68)$$

Since the function $h(W) \triangleq \sum_{j\in\mathcal{J}}\nabla f(W, z_j)$ is convex and the gradient of $h(w)$ is $\beta|\mathcal{J}|$-Lipschitz, it follows that (co-coersivity of the gradient)

$$\sum_{j\in\mathcal{J}}\langle\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)}\rangle \qquad (69)$$

$$\geq \frac{1}{\beta|\mathcal{J}|}\|\sum_{j\in\mathcal{J}}\nabla f(W_t, z_j) - \sum_{j\in\mathcal{J}}\nabla f(W_t^{(i)}, z_j)\|_2^2. \qquad (70)$$

Then prove the inequality 66 as follows

$$\begin{aligned}
&\left\| W_t - W_t^{(i)} - \frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\right)\right\|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\langle\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)}\rangle \qquad (71) \\
&\quad + \frac{\eta_t^2}{n^2}\|\sum_{j=1, j\neq i}^{n}\left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\right)\|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{n}\langle\sum_{j=1, j\neq i}^{n}\nabla f(W_t, z_j) - \sum_{j=1, j\neq i}^{n}\nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)}\rangle \\
&\quad + \frac{\eta_t^2}{n^2}\|\sum_{j=1, j\neq i}^{n}\nabla f(W_t, z_j) - \sum_{j=1, j\neq i}^{n}\nabla f(W_t^{(i)}, z_j)\|_2^2
\end{aligned}$$

$$\leq \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{\beta(n-1)n}\|\sum_{j=1,j\neq i}^{n} \nabla f(W_t, z_j) - \sum_{j=1,j\neq i}^{n} \nabla f(W_t^{(i)}, z_j)\|_2^2$$
$$+ \frac{\eta_t^2}{n^2}\|\sum_{j=1,j\neq i}^{n} \nabla f(W_t, z_j) - \sum_{j=1,j\neq i}^{n} \nabla f(W_t^{(i)}, z_j)\|_2^2 \tag{72}$$
$$= \|W_t - W_t^{(i)}\|_2^2 + \frac{\eta_t}{n}\left(\frac{\eta_t}{n} - \frac{2}{\beta(n-1)}\right)\left\|\sum_{j=1,j\neq i}^{n}\left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\right)\right\|_2^2$$
$$\leq \|W_t - W_t^{(i)}\|_2^2, \tag{73}$$

equation 71 holds from the expansion of the squared norm, 72 comes from the inequality 70. The inequality 73 holds under the choice $\eta_t < 2/\beta$ and completes the proof. □

Lemma 19 (Accumulated Path Error - Convex Loss) *Let the loss function $f(\cdot; z)$ be convex and β-smooth and $\eta_t \leq 1/2\beta$. Then the expected path-error of the full-batch GD after T iterations is bounded as*

$$\epsilon_{\text{path}} \leq 4\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + 8\beta\mathbb{E}[R_S(W_S^*)]\sum_{t=1}^{T}\eta_t \tag{74}$$

Proof. The self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ (Srebro et al., 2010, Lemma 3.1) gives $\|\nabla f(W_t, z_i)\|_2^2 \leq 4\beta f(W_t, z_i)$. By taking expectation, and through the equation 20 we find

$$\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \leq 4\beta\mathbb{E}[f(W_t, z_i)] = 4\beta\mathbb{E}[R_S(W_t)]. \tag{75}$$

Similarly to the approach by Lei & Ying (2020b, Appendix A, Lemma 2), we use the convexity and the assumption $\eta_t \leq 1/2\beta$ to find

$$\begin{aligned}
\|W_{t+1} - W_S^*\|_2^2 &= \|W_t - \eta_t\nabla R_S(W_t) - W_S^*\|_2^2 \\
&= \|W_t - W_S^*\|_2^2 + \eta_t^2\|\nabla R_S(W_t)\|_2^2 + 2\eta_t\langle W_S^* - W_t, \nabla R_S(W_t)\rangle \\
&\leq \|W_t - W_S^*\|_2^2 + \eta_t^2\|\nabla R_S(W_t)\|_2^2 + 2\eta_t\left(R_S(W_S^*) - R_S(W_t)\right) \\
&\leq \|W_t - W_S^*\|_2^2 + 2\beta\eta_t^2 R_S(W_t) + 2\eta_t\left(R_S(W_S^*) - R_S(W_t)\right) \\
&\leq \|W_t - W_S^*\|_2^2 + 2\eta_t R_S(W_S^*) - \eta_t R_S(W_t).
\end{aligned}$$

The last gives

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t R_S(W_t) &\leq \sum_{t=1}^{T}\|W_t - W_S^*\|_2^2 - \sum_{t=1}^{T}\|W_{t+1} - W_S^*\|_2^2 + 2\sum_{t=1}^{T}\eta_t R_S(W_S^*) \\
&\leq \|W_1 - W_S^*\|_2^2 + 2\sum_{t=1}^{T}\eta_t R_S(W_S^*).
\end{aligned} \tag{76}$$

The definition of ϵ_{path} (Definition 5), the inequalities 75, 76 and the choice of the learning rate ($\eta_t \leq 1/2\beta$) give

$$\begin{aligned}
\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] &\leq 4\beta\sum_{t=1}^{T}\eta_t\mathbb{E}[R_S(W_t)] \\
&\leq 4\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + 8\beta\sum_{t=1}^{T}\eta_t\mathbb{E}[R_S(W_S^*)].
\end{aligned} \tag{77}$$

The last inequality provides the bound on the ϵ_{path}. □

The standard choice of $\eta_t \leq 1/\beta$ gives the next known bound on the optimization error.

Lemma 20 (Optimization Error - Convex Loss (Nesterov, 1998)) *If $f(\cdot; z)$ is a convex and β-smooth function and $\eta_t \leq 1/\beta$, then*

$$\epsilon_{\text{opt}} = \mathbb{E}[R_S(A(S)) - R_S(W_S^*)] \leq \frac{\mathbb{E}[\|W_1 - W_S^*\|_2^2]}{\sum_{t=1}^{T}\eta_t\left(1 - \frac{\beta\eta_t}{2}\right)}. \tag{78}$$

E.1 Proof of Theorem 9 and Theorem 10

Let $z_1, z_2, \ldots, z_i, \ldots, z_n, z_i'$ be i.i.d. random variables, define $S \triangleq (z_1, z_2, \ldots, z_i, \ldots, z_n)$ and $S^{(i)} \triangleq (z_1, z_2, \ldots, z_i', \ldots, z_n)$, $W_1 = W_1'$. The updates for any $t \geq 1$ are

$$W_{t+1} = W_t - \frac{\eta_t}{n} \sum_{j=1}^{n} \nabla f(W_t, z_j), \tag{79}$$

$$W_{t+1}^{(i)} = W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) - \frac{\eta_t}{n} \nabla f(W_t^{(i)}, z_i'). \tag{80}$$

Then for any $t \geq 1$

$$\begin{aligned}
&\|W_{t+1} - W_{t+1}^{(i)}\|_2 \\
&\leq \left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \right) \right\|_2 \\
&\quad + \frac{\eta_t}{n} \|\nabla f(W_t, z_i) - \nabla f(W_t^{(i)}, z_i')\|_2 \\
&\leq \sqrt{\left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \right) \right\|_2^2} \\
&\quad + \frac{\eta_t}{n} \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right) \\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right).
\end{aligned} \tag{81}$$

The inequality 81 comes from Lemma 18. Then by solving the recursion, we find

$$\|W_{T+1} - W_{T+1}^{(i)}\|_2 \leq \frac{1}{n} \sum_{t=1}^{T} \eta_t \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right)$$

thus

$$\begin{aligned}
\|W_{T+1} - W_{T+1}^{(i)}\|_2^2 &\leq \frac{1}{n^2} \left(\sum_{t=1}^{T} \eta_t \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right) \right)^2 \\
&\leq \frac{2}{n^2} \sum_{t=1}^{T} \eta_t \left(\|\nabla f(W_t, z_i)\|_2^2 + \|\nabla f(W_t^{(i)}, z_i')\|_2^2 \right) \sum_{t=1}^{T} \eta_t.
\end{aligned} \tag{82}$$

Inequality 82 gives that for any $i \in \{1, \ldots, n\}$

$$\begin{aligned}
\mathbb{E}[\|W_{T+1} - W_{T+1}^{(i)}\|_2^2] &\leq \frac{2}{n^2} \sum_{t=1}^{T} \eta_t \left(\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] + \mathbb{E}[\|\nabla f(W_t^{(i)}, z_i')\|_2^2] \right) \sum_{t=1}^{T} \eta_t \\
&= \frac{4}{n^2} \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \sum_{t=1}^{T} \eta_t = \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t.
\end{aligned} \tag{83}$$

Recall that $W_{T+1} \equiv A(S)$ and $W_{T+1}^{(i)} \equiv A(S^{(i)})$. Theorem 3 and the inequality 83 give

$$\begin{aligned}
|\epsilon_{\text{gen}}| &\leq 2\sqrt{2\beta \left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)] \right) \epsilon_{\text{stab}(A)}} + 2\beta \epsilon_{\text{stab}(A)} \\
&\leq 2\sqrt{2\beta \left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)] \right) \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t} + 2\beta \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \\
&= \frac{4\sqrt{\left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)] \right) \epsilon_{\text{path}}}}{n} \sqrt{2\beta \sum_{t=1}^{T} \eta_t} + 8\beta \frac{\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t.
\end{aligned} \tag{84}$$

Under the choice of constant learning rate $\eta_t = 1/2\beta$, Lemma 19 together with the inequality 83 give $\epsilon_{\text{path}} \leq 4\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + 8\beta\mathbb{E}[R_S(W_S^*)]\sum_{t=1}^T \eta_t$, and $\epsilon_{\text{opt}} \leq 3\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2]/T$. Thus

$$\epsilon_{\text{stab}(A)} \leq \frac{32}{n^2}\left(\frac{\beta}{2}\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \beta\mathbb{E}[R_S(W_S^*)]\sum_{t=1}^T \eta_t\right)\sum_{t=1}^T \eta_t \tag{85}$$

$$= \frac{32}{n^2}\left(\frac{\beta}{2}\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \frac{1}{2}\mathbb{E}[R_S(W_S^*)]T\right)\frac{T}{2\beta} \tag{86}$$

$$= \frac{8T}{n^2}\left(\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \mathbb{E}[R_S(W_S^*)]\frac{T}{\beta}\right). \tag{87}$$

The inequality 84 and Lemma 20 give

$$\begin{aligned}
|\epsilon_{\text{gen}}| &\leq 2\sqrt{2\beta\left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)]\right)\epsilon_{\text{stab}(A)}} + 2\beta\epsilon_{\text{stab}(A)} \\
&\leq 2\sqrt{2\beta\left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)]\right)\frac{8T}{n^2}\left(\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \mathbb{E}[R_S(W_S^*)]\frac{T}{\beta}\right)} \\
&\quad + 2\beta\frac{8T}{n^2}\left(\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \mathbb{E}[R_S(W_S^*)]\frac{T}{\beta}\right) \\
&\leq \frac{8\sqrt{T}}{n}\sqrt{\left(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)]\right)\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right)} \\
&\quad + \frac{16T}{n^2}\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right) \\
&\leq \frac{8\sqrt{T}}{n}\sqrt{\left(\frac{3\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2]}{T} + \mathbb{E}[R_S(W_S^*)]\right)\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right)} \\
&\quad + \frac{16T}{n^2}\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right) \\
&= \frac{8}{n}\sqrt{\left(3\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right)\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right)} \\
&\quad + \frac{16T}{n^2}\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right) \\
&\leq \frac{8}{n}\left(3\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right) + \frac{16T}{n^2}\left(\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right) \\
&\leq 8\left(\frac{1}{n} + \frac{2T}{n^2}\right)\left(3\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2] + T\mathbb{E}[R_S(W_S^*)]\right).
\end{aligned}$$

The last inequality completes the proof. □

F STRONGLY-CONVEX OBJECTIVE: PROOF OF THEOREM 12 AND THEOREM 13

Similarly to the convex case, first we provide the contractive property of the stability recursion in the strongly convex loss case. Then we prove the stability and generalization error bounds.

Lemma 21 *Let the objective function be γ-strongly convex ($\gamma > 0$) and the leave-one-out objective function be γ_{loo}-strongly convex for some $\gamma_{loo} \geq 0$. If the loss function is convex β-smooth for all $z \in \mathcal{Z}$ and $\eta_t \leq 2/(\beta + \gamma)$, then for any $t \leq T + 1$ the updates $W_t, W_t^{(i)}$ satisfy the inequality*

$$\left\|W_t - W_t^{(i)} - \eta_t\left(\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\right)\right\|_2^2 \leq (1 - \eta_t\gamma_{loo})\|W_t - W_t^{(i)}\|_2^2.$$

Proof. The function $R_{S^{-i}}(\cdot)$ is also β-smooth for all $z \in \mathcal{Z}$ and the strong convexity gives

$$\langle \nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)}), W_t - W_t^{(i)}\rangle$$

$$\geq \frac{\beta\gamma_{loo}}{\beta+\gamma_{loo}}\|W_t - W_t^{(i)}\|_2^2 + \frac{1}{(\beta+\gamma_{loo})}\|\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\|_2^2 \tag{88}$$

We expand the squared norm as follows

$$\begin{aligned}
&\left\|W_t - W_t^{(i)} - \eta_t\left(\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\right)\right\|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\eta_t\langle \nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)}), W_t - W_t^{(i)}\rangle \\
&\quad + \eta_t^2\|\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\|_2^2 \\
&\leq \|W_t - W_t^{(i)}\|_2^2 + \eta_t^2\|\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\|_2^2 \\
&\quad - 2\eta_t\left(\frac{\beta\gamma_{loo}}{\beta+\gamma_{loo}}\|W_t - W_t^{(i)}\|_2^2 + \frac{1}{(\beta+\gamma_{loo})}\|\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\|_2^2\right) \qquad (89) \\
&= \left(1 - 2\eta_t\frac{\beta\gamma_{loo}}{\beta+\gamma_{loo}}\right)\|W_t - W_t^{(i)}\|_2^2 \\
&\quad + \eta_t\left(\eta_t - \frac{2}{\beta+\gamma_{loo}}\right)\|\nabla R_{S^{-i}}(W_t) - \nabla R_{S^{-i}}(W_t^{(i)})\|_2^2 \\
&\leq \left(1 - 2\eta_t\frac{\beta\gamma_{loo}}{\beta+\gamma_{loo}}\right)\|W_t - W_t^{(i)}\|_2^2. \qquad (90)
\end{aligned}$$

We apply the inequality 88 to derive 89. The inequality 90 holds since $\eta_t \leq 2/(\beta+\gamma)$ and $\beta \geq \gamma > \gamma_{loo}$. Also

$$2\eta_t\frac{\beta\gamma_{loo}}{\beta+\gamma_{loo}} \geq 2\eta_t\frac{\beta\gamma_{\text{loo}}}{2\beta} = \eta_t\gamma_{\text{loo}}. \tag{91}$$

Through the inequalities 90 and 91 to derive the bound of the lemma. □

Lemma 22 (Accumulated Path Error - Strongly Convex Loss) *Let the objective function $R_S(\cdot)$ be γ-strongly convex and β-smooth. Define $\Gamma(\gamma, T) \triangleq (1 - \exp(\frac{-4T\gamma}{\beta+\gamma})/(\exp(\frac{-4\gamma}{\beta+\gamma}) - 1)$. If $\eta_t = 2/(\beta+\gamma)$, then the expected path-error of the full-batch GD after T iterations are bounded as*

$$\epsilon_{\text{path}} \leq \frac{4\beta^2}{\beta+\gamma}\Gamma(\gamma, T)\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)], \tag{92}$$

Proof. The self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ (Srebro et al., 2010, Lemma 3.1) gives $\|\nabla f(W_t, z_i)\|_2^2 \leq 4\beta f(W_t, z_i)$. By taking expectation, and through the Assumption 1 and equation 20, we find

$$\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \leq 4\beta\mathbb{E}[f(W_t, z_i)] = 4\beta\mathbb{E}[R_S(W_t)] = 4\beta\mathbb{E}[R_S(W_t) - R_S(W_S^*) + R_S(W_S^*)]. \tag{93}$$

Further, Lemma 23 and the choice of constant learning rate $\eta = 2/(\beta+\gamma)$ give

$$\mathbb{E}[R_S(W_t) - R_S(W_S^*)] \leq \frac{\beta}{2}\exp\left(\frac{-4t}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{94}$$

The definition of ϵ_{path} (Definition 5), the inequalities 93 and 94 and the constant learning rate ($\eta_t = 2/(\beta+\gamma)$) give

$$\begin{aligned}
\epsilon_{\text{path}} &\triangleq \sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \qquad (95) \\
&\leq 4\beta\sum_{t=1}^{T}\eta_t\mathbb{E}[R_S(W_t) - R_S(W_S^*) + R_S(W_S^*)] \\
&\leq 4\beta\sum_{t=1}^{T}\frac{2}{\beta+\gamma}\frac{\beta}{2}\exp\left(\frac{-4t}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{4\beta^2}{\beta+\gamma}\mathbb{E}[\|W_1 - W_S^*\|_2^2]\sum_{t=1}^{T}\exp\left(\frac{-4t}{\frac{\beta}{\gamma}+1}\right) + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)] \\
&= \frac{4\beta^2}{\beta+\gamma}\mathbb{E}[\|W_1 - W_S^*\|_2^2]\underbrace{\exp\left(\frac{-4}{\frac{\beta}{\gamma}+1}\right)\frac{1-\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)}{1-\exp\left(\frac{-4}{\frac{\beta}{\gamma}+1}\right)}}_{\Gamma(\gamma,T)} + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)] \\
&= \frac{4\beta^2}{\beta+\gamma}\Gamma(\gamma,T)\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)]. \qquad (96)
\end{aligned}
$$

The last inequality provides the bound on the ϵ_{path}. Further, we can show that

$$
\Gamma(\gamma,T) \leq \min\left\{\frac{1}{e^{\frac{4\gamma}{\beta+\gamma}}-1}, T\right\} \qquad (97)
$$

to simplify the expression in the inequality 96. □

Lemma 23 ((Nesterov, 1998, Theorem 2.1.14)) *If $f(\cdot;z)$ is a γ-strongly convex and β-smooth function and $\eta_t = 2/(\beta+\gamma)$, then*

$$
\epsilon_{\text{opt}} \leq \frac{\beta}{2}\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \qquad (98)
$$

Alternatively, if $\eta_t = c/t$, then

$$
\epsilon_{\text{opt}} \leq \frac{\beta}{2}T^{-\frac{2c\beta\gamma}{\beta+\gamma}}\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \qquad (99)
$$

F.1 Proof of Theorem 12 and Theorem 13

Let $z_1, z_2, \ldots, z_i, \ldots, z_n, z_i'$ be i.i.d. random variables, define $S \triangleq (z_1, z_2, \ldots, z_i, \ldots, z_n)$ and $S^{(i)} \triangleq (z_1, z_2, \ldots, z_i', \ldots, z_n)$, $W_1 = W_1'$. The updates for any $t \geq 1$ are

$$
W_{t+1} = W_t - \frac{\eta_t}{n}\sum_{j=1}^{n}\nabla f(W_t, z_j), \qquad (100)
$$

$$
W_{t+1}^{(i)} = W_t^{(i)} - \frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\nabla f(W_t^{(i)}, z_j) - \frac{\eta_t}{n}\nabla f(W_t^{(i)}, z_i'). \qquad (101)
$$

Then similarly to the inequality 81 we get

$$
\begin{aligned}
&\|W_{t+1} - W_{t+1}^{(i)}\|_2 \\
&\leq \left\|W_t - W_t^{(i)} - \frac{\eta_t}{n}\sum_{j=1, j\neq i}^{n}\left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\right)\right\|_2 \\
&\quad + \frac{\eta_t}{n}\|\nabla f(W_t, z_i) - \nabla f(W_t^{(i)}, z_i')\|_2 \\
&\leq \sqrt{\left\|W_t - W_t^{(i)} - \eta_t\left(R_{S^{-i}}(W_t) - R_{S^{-i}}(W_t^{(i)})\right)\right\|_2^2} \\
&\quad + \frac{\eta_t}{n}\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right) \\
&\leq (1-\eta_t\gamma_{loo})^{\frac{1}{2}}\|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n}\left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2\right) \qquad (102)
\end{aligned}
$$

and we apply Lemma 21 to derive the bound in 102. Then by solving the recursion we find

$$
\|W_{T+1} - W_{T+1}^{(i)}\|_2
$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2+\|\nabla f(W_t^{(i)},z_i')\|_2\right)\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})^{\frac{1}{2}}\\
&\leq \frac{1}{n}\sqrt{\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2+\|\nabla f(W_t^{(i)},z_i')\|_2\right)^2\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})}\\
&\leq \frac{1}{n}\sqrt{2\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2^2+\|\nabla f(W_t^{(i)},z_i')\|_2^2\right)\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})}.
\end{aligned}
$$

The last inequality provides the stability bound

$$
\begin{aligned}
&\|W_{T+1}-W_{T+1}^{(i)}\|_2^2\\
&\leq \frac{2}{n^2}\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2^2+\|\nabla f(W_t^{(i)},z_i')\|_2^2\right)\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})\,. \qquad (103)
\end{aligned}
$$

Inequality 103 gives that for any $i\in\{1,\dots,n\}$

$$
\begin{aligned}
&\mathbb{E}[\|W_{T+1}-W_{T+1}^{(i)}\|_2^2]\\
&\leq \frac{2}{n^2}\sum_{t=1}^{T}\eta_t\left(\mathbb{E}[\|\nabla f(W_t,z_i)\|_2^2]+\mathbb{E}[\|\nabla f(W_t^{(i)},z_i')\|_2^2]\right)\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})\\
&= \frac{4}{n^2}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(W_t,z_i)\|_2^2]\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})\\
&= \frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_j\gamma_{loo})\,. \qquad (104)
\end{aligned}
$$

Recall that $W_{T+1}\equiv A(S)$ and $W_{T+1}^{(i)}\equiv A(S^{(i)})$. Due to space limitation, we define $\Omega(\eta_t,\gamma_{loo})\triangleq\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\eta_t\gamma_{\text{loo}}^2/\beta\right)$ Theorem 3 and the inequality 104 give

$$
\begin{aligned}
|\epsilon_{\text{gen}}| &\leq 2\sqrt{2\beta(\epsilon_{\text{opt}}+\mathbb{E}[R_S(W_S^*)])\epsilon_{\text{stab}(A)}}+2\beta\epsilon_{\text{stab}(A)} \qquad (105)\\
&\leq 2\sqrt{2\beta(\epsilon_{\text{opt}}+\mathbb{E}[R_S(W_S^*)])\frac{4\epsilon_{\text{path}}}{n^2}\Omega(\eta_t,\gamma_{loo})}+2\beta\frac{4\epsilon_{\text{path}}}{n^2}\Omega(\eta_t,\gamma_{loo})\\
&= \frac{4\sqrt{(\epsilon_{\text{opt}}+\mathbb{E}[R_S(W_S^*)])\epsilon_{\text{path}}}}{n}\sqrt{2\beta\Omega(\eta_t,\gamma_{loo})}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\Omega(\eta_t,\gamma_{loo}).
\end{aligned}
$$

Under the choice of $\eta_t=C=\frac{2}{\beta+\gamma}<\frac{2}{\beta}$, the inequality 104, Lemmata 15 and 22 and give

$$
\begin{aligned}
\mathbb{E}[\|A(S)-A(S^{(i)})\|_2^2] &\leq \frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\eta_t\gamma_{\text{loo}})\\
&= \frac{4\epsilon_{\text{path}}}{n^2}\underbrace{\frac{1-\left(1-\frac{2\gamma_{loo}}{\beta+\gamma}\right)^T}{\gamma_{loo}}}_{\Lambda(\gamma_{loo},T)}\\
&\leq \frac{4\epsilon_{\text{path}}}{n^2}\Lambda(\gamma_{loo},T)\leq\frac{4\epsilon_{\text{path}}}{n^2}\min\left\{\frac{1}{\gamma_{loo}},\frac{2T}{\beta}\right\} \qquad (106)
\end{aligned}
$$

and the last inequality holds since $\Lambda(\gamma_{loo},T)\leq 1/\gamma_{loo}$ for any T and the monotonicity of $\Lambda(\gamma_{loo},T)$ gives $\Lambda(\gamma_{loo},T)\leq 2T/\beta$ for any pair $\gamma_{loo}\leq\gamma$. Though the inequality 105, we find the generalization error bound

$$
|\epsilon_{\text{gen}}|\leq\frac{4\sqrt{(\epsilon_{\text{opt}}+\mathbb{E}[R_S(W_S^*)])\epsilon_{\text{path}}}}{n}\sqrt{2\beta\Omega(\eta_t,\gamma_{loo})}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\Omega(\eta_t,\gamma_{loo})
$$

$$
\begin{aligned}
&\leq \frac{4\sqrt{(\epsilon_{\text{opt}} + \mathbb{E}[R_S(W_S^*)])\epsilon_{\text{path}}}}{n}\sqrt{2\beta\Lambda(\gamma_{loo}, T)} + 8\beta\frac{\epsilon_{\text{path}}}{n^2}\Lambda(\gamma_{loo}, T) \\
&\leq \frac{4}{n}\sqrt{\left(\frac{\beta}{2}\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \mathbb{E}[R_S(W_S^*)]\right)} \\
&\quad \times \sqrt{\left(\frac{4\beta^2}{\beta+\gamma}\Gamma(\gamma, T)\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)]\right)}\sqrt{2\beta\Lambda(\gamma_{loo}, T)} \\
&\quad + 8\beta\frac{\frac{4\beta^2}{\beta+\gamma}\Gamma(\gamma, T)\mathbb{E}[\|W_1 - W_S^*\|_2^2] + \frac{8\beta T}{\beta+\gamma}\mathbb{E}[R_S(W_S^*)]}{n^2}\Lambda(\gamma_{loo}, T) \\
&\leq \frac{4}{n}\sqrt{\left(\frac{4\beta}{\beta+\gamma}\Gamma(\gamma, T) + \frac{8\beta T}{\beta+\gamma}\right)\max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}} \\
&\quad \times \sqrt{2\beta\Lambda(\gamma_{loo}, T)} + \left(\frac{4}{n}\sqrt{\frac{1}{2}\exp\left(\frac{-4T\gamma}{\beta+\gamma}\right)\left(\frac{4\beta}{\beta+\gamma}\Gamma(\gamma, T) + \frac{8\beta T}{\beta+\gamma}\right)}\sqrt{2\beta\Lambda(\gamma_{loo}, T)}\right. \\
&\quad \left. + 8\frac{\frac{4\beta}{\beta+\gamma}\Gamma(\gamma, T)] + \frac{8\beta T}{\beta+\gamma}}{n^2}\beta\Lambda(\gamma_{loo}, T)\right)\max\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\} \\
&= \frac{4}{n}\sqrt{\left(\frac{4\beta}{\beta+\gamma}\Gamma(\gamma, T) + \frac{8\beta T}{\beta+\gamma}\right)\max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}} \\
&\quad \times \sqrt{2\beta\Lambda(\gamma_{loo}, T)} + \left(\frac{4}{n}\sqrt{\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)\left(\frac{4\beta}{\beta+\gamma}\Gamma(\gamma, T) + \frac{8\beta T}{\beta+\gamma}\right)}\sqrt{\beta\Lambda(\gamma_{loo}, T)}\right. \\
&\quad \left. + 8\frac{\frac{4\beta}{\beta+\gamma}\Gamma(\gamma, T) + \frac{8\beta T}{\beta+\gamma}}{n^2}\beta\Lambda(\gamma_{loo}, T)\right)\max\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\}. \qquad (107)
\end{aligned}
$$

We proceed by applying the upper bounds of $\Gamma(\gamma, T), \Lambda(\gamma_{loo}, T)$ as appear in the inequalities 97 and 106 respectively and equation 107 gives

$$
\begin{aligned}
&|\epsilon_{\text{gen}}| \\
&\leq \frac{4}{n}\sqrt{\left(\frac{4\beta}{\beta+\gamma}\min\left\{\frac{1}{e^{\frac{4\gamma}{\beta+\gamma}}-1}, T\right\} + \frac{8\beta T}{\beta+\gamma}\right)} \\
&\quad \times \sqrt{\max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}2\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}} \\
&\quad + \left(\frac{4}{n}\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right)\sqrt{\left(\frac{4\beta}{\beta+\gamma}\min\left\{\frac{1}{e^{\frac{4\gamma}{\beta+\gamma}}-1}, T\right\} + \frac{8\beta T}{\beta+\gamma}\right)2\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}}\right. \\
&\quad \left. + 8\frac{\frac{4\beta}{\beta+\gamma}\min\left\{\frac{1}{e^{\frac{4\gamma}{\beta+\gamma}}-1}, T\right\} + \frac{8\beta T}{\beta+\gamma}}{n^2}\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}\right)\max\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\} \\
&\leq \frac{4}{n}\sqrt{\left(\frac{4\beta T}{\beta+\gamma} + \frac{8\beta T}{\beta+\gamma}\right)\max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}2\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}} \\
&\quad + \left(\frac{4}{n}\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right)\sqrt{\left(\frac{4\beta T}{\beta+\gamma} + \frac{8\beta T}{\beta+\gamma}\right)2\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}}\right. \\
&\quad \left. + 8\frac{\frac{4\beta T}{\beta+\gamma} + \frac{8\beta T}{\beta+\gamma}}{n^2}\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}\right)\max\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{8\sqrt{3}}{n}\sqrt{\frac{\beta T}{\beta+\gamma}\max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\} 2\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}} \\
&+ \left(\frac{8\sqrt{3}}{n}\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right)\sqrt{\frac{\beta T}{\beta+\gamma}2\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}}\right. \\
&\quad \left. + 96\frac{\frac{\beta T}{\beta+\gamma}}{n^2}\min\left\{\frac{\beta}{\gamma_{loo}}, 2T\right\}\right)\max\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\}. \qquad (108)
\end{aligned}
$$

To simplify the last display, we define the terms $\mathrm{m}(\gamma_{loo}, T) \triangleq \beta T \min\left\{\beta/\gamma_{loo}, 2T\right\}/(\beta+\gamma)$ and $\mathrm{M}(W_1) \triangleq \max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}$

$$
\begin{aligned}
&|\epsilon_{\text{gen}}| \\
&\leq \frac{8\sqrt{6}}{n}\sqrt{\mathrm{m}(\gamma_{loo}, T)M(W_1)} \\
&\quad + \left(\frac{8\sqrt{6}}{n}\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right)\sqrt{\mathrm{m}(\gamma_{loo}, T)} + \frac{96}{n^2}\mathrm{m}(\gamma_{loo}, T)\right)\mathrm{M}(W_1) \\
&= \frac{8\sqrt{6}}{n}\left(\sqrt{M(W_1)} + \left(\exp\left(\frac{-2T\gamma}{\beta+\gamma}\right) + \frac{4\sqrt{3}}{n}\sqrt{\mathrm{m}(\gamma_{loo}, T)}\right)\mathrm{M}(W_1)\right)\sqrt{\mathrm{m}(\gamma_{loo}, T)}
\end{aligned}
$$

Choose $T = \log(n)(\beta+\gamma)/2\gamma$ and define $\mathrm{m}_{n,\gamma_{loo}} \triangleq \frac{\beta}{2\gamma}\min\left\{\frac{\beta}{\gamma_{loo}}, \frac{\beta+\gamma}{\gamma}\log n\right\}$, then the inequality 108 gives

$$
\begin{aligned}
&|\epsilon_{\text{gen}}| \\
&\leq \frac{8\sqrt{6\log n}}{n}\sqrt{\frac{\beta}{2\gamma}\max\left\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\right\}\min\left\{\frac{\beta}{\gamma_{loo}}, \frac{\beta+\gamma}{\gamma}\log n\right\}} \\
&\quad + \left(\frac{8\sqrt{6}}{n^2}\sqrt{\frac{\beta\log n}{2\gamma}\min\left\{\frac{\beta}{\gamma_{loo}}, \frac{\beta+\gamma}{\gamma}\log(n)\right\}}\right. \\
&\quad \left. + \frac{48\beta}{\gamma}\frac{\log n}{n^2}\min\left\{\frac{\beta}{\gamma_{loo}}, \frac{\beta+\gamma}{\gamma}\log(n)\right\}\right)\max\{\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2], \mathbb{E}[R_S(W_S^*)]\} \\
&= \frac{8\sqrt{6\log n}}{n}\left(\sqrt{\mathrm{m}_{n,\gamma_{loo}}\mathrm{M}(W_1)} + \left(\frac{1}{n}\sqrt{\mathrm{m}_{n,\gamma_{loo}}} + \frac{4\sqrt{3}}{n}\mathrm{m}_{n,\gamma_{loo}}\right)\mathrm{M}(W_1)\right) \\
&= \frac{8\sqrt{6\log n}}{n}\left(\sqrt{\mathrm{M}(W_1)} + \frac{1 + 4\sqrt{3\mathrm{m}_{n,\gamma_{loo}}}}{n}\mathrm{M}(W_1)\right)\sqrt{\mathrm{m}_{n,\gamma_{loo}}}. \qquad (109)
\end{aligned}
$$

The last inequality completes the proof. □