Beyond Lipschitz: Sharp Generalization and Excess Risk Bounds for Full-Batch GD

Konstantinos E. Nikolakakis
konstantinos.nikolakakis@yale.edu

Farzin Haddadpour
farzin.haddadpour@yale.edu

Amin Karbasi
amin.karbasi@yale.edu

Dionysios S. Kalogerias
dionysis.kalogerias@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), under an interpolation regime. At the heart of our analysis is a new generalization error bound for deterministic symmetric algorithms, which implies that *average output stability* and a *bounded expected optimization error at termination* lead to generalization. This result shows that 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, Polyak-Lojasiewicz (PL), 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, under the proper choice of a decreasing step size. Further, if the loss is nonconvex but the objective is PL, we derive quadratically vanishing bounds on the generalization error and the corresponding excess risk, for a choice of a large constant step size. For (resp. strongly-) convex smooth losses, we prove that full-batch GD also generalizes for large constant step sizes, and achieves (resp. quadratically) small excess risk while training fast. In all cases, we close the generalization error gap, by showing matching generalization and optimization error rates. Our full-batch GD generalization error and excess risk bounds are strictly tighter than existing bounds for (stochastic) GD, when the loss is smooth (but possibly non-Lipschitz).

Keywords: Full-Batch GD, Generalization Error, Smooth Nonconvex/Convex Optimization

1 Introduction

Gradient based learning [1] is a well established topic with a large body of literature that includes detailed analysis on the algorithmic generalization and optimization errors. For general smooth losses, optimization error guarantees are known [2]. Specifically, Absil et al. [3] and Lee et al. [4] showed the convergence of Gradient Descent (GD) to minimizers and local minima for smooth nonconvex functions. Recently, Chatterjee [5] and Liu et al. [6] showed the convergence of GD for deep neural networks. While prior works provide extensive optimization analysis, tight and path-dependent generalization error and excess risk guarantees in non-stochastic training (for general smooth losses) remain unexplored.

Generalization error analysis of stochastic training algorithms has gained increased attention. In recent years Hardt et al. [7] first showed uniform stability final-iterate bounds for vanilla Stochastic Gradient Descent (SGD). More recent works develop alternative generalization error bounds based on high-probability analysis [8–11] and data-dependent variants [12], or under weaker assumptions

such as as strongly quasi-convex [13], non-smooth convex [14–17], and pairwise losses [18, 19]. In the nonconvex case, [20] provide bounds that involve on-average variance of the stochastic gradients. Generalization performance of other algorithmic variants lately gain further attention, including SGD with early momentum [21], randomized coordinate descent [22], look-ahead approaches [23], noise injection methods [24], and stochastic gradient Langevin dynamics [25–32].

Even though many of previous works consider stochastic training algorithms and some even suggest that stochasticity may be necessary [7, 33] for good generalization, recent empirical studies observe that deterministic algorithms indeed generalize [34, 35]. In fact, Hoffer et al. [34] showed empirically that for large enough number of iterations full-batch GD generalizes comparably to SGD. Similarly, Geiping et al. [35] experimentally showed that strong generalization behavior is still observed in the absence of stochastic sampling. Indeed, one of the main contributions of our work is to prove that full-batch GD generalizes efficiently under the assumption of general smooth losses. In the regime of non-smooth and Lipschitz convex losses, SGD appears to generalizes better than the full-batch GD [15, 36]. In contrast, we show that for general smooth and (possibly) nonconvex losses, full-batch GD outperforms state-of-the-art generalization rates of the SGD.

1.1 Contributions

In this work, we consider the classic full-batch GD under the interpolation regime and provide the first generalization error guarantees for smooth and (possibly) non-Lipschitz nonconvex losses. In fact, we prove generalization error bounds that are tighter than the corresponding state-of-the-art SGD bounds in prior works, effectively showing that stochastic training is not necessary for generalization. In particular, the generalization bounds herein are tighter than existing state-of-the-art on average [37, 38], (or PAC [39]) SGD bounds (for general smooth losses), as well as full-batch GD bounds for Lipschitz losses [40, 41] that appear in prior works. A summary of our contribution is as follows:

- We show a new bound on the generalization error for symmetric algorithms [42] and smooth losses (Theorem 3) under the interpolation regime. Essentially, we show that bounded average algorithmic stability of the output and bounded optimization error suffice to ensure generalization. This result allows to derive tighter bounds than those in prior works.
- We provide sharp generalization error guarantees that characterize the dependence of the generalization error with respect to the number of samples, the number of iterations, a path-dependent term and the optimization error. The explicit expression of the generalization error through the aforementioned quantities allows us to derive tight bounds for different types of (non)convexity.
- For nonconvex losses, we show that full-batch GD generalizes efficiently for appropriate choices of decreasing learning rate. For PL, convex and and strongly convex losses, we show that full-batch GD attains efficient generalization and excess risk performance for large constant step-sizes. Specifically for convex losses, we show that full-batch GD generalizes efficiently closes the generalization error gap through memorization and while training longer. For nonconvex losses with PL objective, as well as strongly convex losses and large constant step-sizes, the generalization error decreases exponentially with respect to the number of iterations.

1.2 Related Work

Let n denote the number of available samples (examples). Recent results [39, 43] on SGD provide bounds of the order $\mathcal{O}(1/\sqrt{n})$ for smooth nonconvex losses. Similarly, Neu et al. [37] also provide generalization bounds of the order $\mathcal{O}(1/\sqrt{n})$, with $T = \sqrt{n}$ and step-size $\eta = 1/T$. In contrast, we

Full-Batch Gradient Descent		
Step Size	Generalization Error	Loss
$\eta_t \leq C/\beta t,\ \forall C < 1$	$\frac{4e\sqrt{3}}{n} T^C \sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}} + \frac{12e^2}{n^2} T^{2C} \epsilon_{\text{path}}$	Nonconvex
$\eta_t \leq C/\beta t,\ \forall C < 1$	$48\left(\frac{\sqrt{\log(eT)}(eT)^{\epsilon}}{n} + \frac{\log(eT)(eT)^{2\epsilon}}{n^2}\right)\mathbb{E}[R_S(W_1)]$	Nonconvex
$\eta_t = 1/\beta$	$\frac{c\beta}{\mu}\frac{\left(1-\frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{c\beta^2}{n^2\mu^2} + \frac{c\beta}{\mu}\left(1-\frac{\mu}{\beta}\right)^{T}$	μ-PL (Objective)
$\eta_t = 1/\beta$, $T = 2\log(n)/\log\left(\frac{\beta}{\beta-\mu}\right)$	$\frac{1}{n^2}\frac{c\beta}{\mu}\left(2+\frac{\beta}{\mu}\right)$	μ-PL (Objective)
$\eta_t = 1/\beta$	$16\beta\left(\frac{\sqrt{2\log(eT)}}{n} + 4\frac{T\log(eT)}{n^2}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2$	Convex
$\eta_t = 1/\beta$, $T = \frac{n}{\sqrt{\log(en)}}$	$87\beta\frac{\sqrt{\log(en)}}{n}\mathbb{E}[\|W_1 - W_S^*\|_2^2$	Convex
$\eta_t = 2/(\beta+\gamma)$	$8\beta\left(\frac{\Delta}{n}e^{\frac{-2\gamma}{\beta+\gamma}T} + \frac{4\Delta^2}{n^2}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]$	γ-Strongly-Convex
$\eta_t = 2/(\beta+\gamma)$, $T = \frac{\beta+\gamma}{2\gamma}\log(n)$	$8\beta\frac{\Delta + 4\Delta^2}{n^2}\mathbb{E}[\|W_1 - W_S^*\|_2^2]$	γ-Strongly-Convex

Table 1: A list of the generalization error bounds for the full-batch GD under the interpolation regime with decreasing and constant learning rates. 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 $\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2]$, "$\epsilon_{\text{opt}}$" denotes the optimization error $\epsilon_{\text{opt}} \triangleq \mathbb{E}[R_S(A(S)) - R_S^*]$, and T is the total number of iterations. Lastly, $\pi_S \triangleq \pi(A(S))$ is the projection of the point $A(S)$ to the set of minimizers of $R_S(\cdot)$, while the constants "c" and "Δ" are defined as $c \triangleq 44\max\{\mathbb{E}[R_S(\pi_S) + R(\pi_S)], \mathbb{E}[R_S(W_1) - R_S^*]\}$ and $\Delta \triangleq \beta/\gamma(e^{4\gamma/(\beta+\gamma)} - 1)^{1/2}$.

show that full-batch GD with decreasing learning rate $\eta_t = 1/2\beta t$ 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 $T \leq n/\log(n)$. Further, for convex and strongly convex losses we provide strictly tighter generalization error bounds than those of SGD in prior works [7, 44, 45] for significantly larger learning rates, while our analysis does not require a Lipschitz or sub-Gaussian loss assumption. Specifically, for nonconvex and nonnegative smooth losses under the μ-PL condition with total number of iterations $T = 2\log(n)/\log(\beta/(\beta-\mu))$ and constant $\eta_t = 1/\beta$ learning rate, we also provide generalization bounds with rate $\mathcal{O}(1/n^2)$. This bound is tighter compared to known $\mathcal{O}(1/n)$ rates in prior works [33, 38] for SGD by one order of magnitude. In fact, while the SGD bounds in prior works [38, Theorem 5] require $T = n$ number of iterations to achieve excess risk rates of the order $\mathcal{O}(1/n)$ under a choice of a decreasing step-size, we show that full-batch GD requires only $T = \Omega(\log(n))$ to achieve $\mathcal{O}(1/n^2)$ rates with large

Full-Batch Gradient Descent		
Step Size	Excess Risk	Loss
$\eta_t = 1/\beta$	$\frac{c\beta}{\mu}\frac{\left(1-\frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{c\beta^2}{n^2\mu^2} + c\left(\frac{\beta}{\mu}+1\right)\left(1-\frac{\mu}{\beta}\right)^T$	μ-PL (Objective)
$\eta_t = 1/\beta$, $T = 2\log(n)/\log\left(\frac{\beta}{\beta-\mu}\right)$	$\frac{1}{n^2}\frac{c\beta}{\mu}\left(3+\frac{\beta}{\mu}\right)$	μ-PL (Objective)
$\eta_t = 1/\beta$	$64\beta\left(\frac{\sqrt{\log(eT)}}{n} + \frac{T\log(eT)}{n^2} + \frac{1}{T}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]$	Convex
$\eta_t = 1/\beta$, $T = \frac{n}{\sqrt{\log(en)}}$	$89\beta\frac{\sqrt{\log(en)}}{n}\mathbb{E}[\|W_1 - W_S^*\|_2^2$	Convex
$\eta_t = 2/(\beta+\gamma)$	$8\beta\left(\frac{\Delta}{n}e^{\frac{-2\gamma}{\beta+\gamma}T} + \frac{4\Delta^2}{n^2} + e^{\frac{-4T\gamma}{\beta+\gamma}}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]$	γ-Strongly-Convex
$\eta_t = 2/(\beta+\gamma)$, $T = \frac{\beta+\gamma}{2\gamma}\log(n)$	$8\beta\frac{1+\Delta+4\Delta^2}{n^2}\mathbb{E}[\|W_1 - W_S^*\|_2^2]$	γ-Strongly-Convex

Table 2: A list of excess risk bounds for the full-batch GD under the interpolation regime with decreasing and constant learning rates. We denote the number of samples by n, W_1 is the initial point of the algorithm, W_S^* is a point in the set of minimizers of the objective, and $\pi_S \triangleq \pi(A(S))$ is the projection of the point $A(S)$ to the set of the minimizers of $R_S(\cdot)$. The constants "c" and "Δ" are defined as $c \triangleq 44\max\{\mathbb{E}[R_S(\pi_S) + R(\pi_S)], \mathbb{E}[R_S(W_1) - R_S^*]\}$ and $\Delta \triangleq \beta/\gamma(e^{4\gamma/(\beta+\gamma)} - 1)^{1/2}$.

constant step-size. Similarly, for a β-smooth and γ-strongly convex loss with a constant stepsize (i.e., depending only on β, γ) and $T = C\log(n)$ (with $C = (\beta/\gamma + 1)/2$), we show bounds of the order $\mathcal{O}(1/n^2)$, while prior works [7, 44, 45] on SGD provide rates of the order $\mathcal{O}(1/n)$. This means that full batch GD provably attains improved generalization error rates by one order of magnitude. Additionally, for convex losses and for a fixed step-size $\eta_t = 1/\beta$ and $T \leq n/\sqrt{\log(en)}$, we show tighter generalization error bounds of the order $\mathcal{O}(\sqrt{\log(T)}/n)$, while bounds in prior work are of the order $\mathcal{O}(T/n)$ [7, 45]. As a consequence, for smooth convex losses with $T \leq n$[1], we provide tighter full-batch GD generalization error bounds than existing bounds on SGD. Essentially, we show that for smooth nonconvex μ-PL, convex and strongly convex losses, full-batch GD generalizes efficiently while training fast and longer, which provides an explanation of its good empirical performance in practice [34]. We refer the reader to Table 1 for an overview of our generalization error results, also showing some interesting trade-offs between alternative learning rates and the corresponding achievable bounds.

Our results indicate that full-batch GD in the interpolation regime also provably achieves substantially smaller excess error rates, as compared to the state-of-the-art excess error guarantees for SGD. In fact, our results can be easily converted to excess risk bounds, by combining them with well-known bounds on the optimization error of the (full-batch) GD algorithm [2, Theorem

[1] or $T \leq \sqrt{n}$, see prior works [15, 44] that consider such choices

2.1.14, Corollary 2.1.2][2]. Specifically, for nonconvex and nonnegative smooth losses under the μ-PL condition with total number of iterations $T = 2\log(n)/\log(\beta/(\beta-\mu))$ and constant learning rate $\eta_t = 1/\beta$, we also provide excess risk guarantees with rate $\mathcal{O}(1/n^2)$. For convex losses, we show that with constant step size and $T = n/\sqrt{\log(en)}$, the excess risk is of the order $\mathcal{O}(\sqrt{\log(n)}/n)$. Finally, if the loss is strongly convex, $T = (\beta/\gamma + 1)\log(n)/2$ iterations are required such that the excess risk is $\mathcal{O}(1/n^2)$, with a fixed learning rate. We refer the reader to Table 2 for an overview of the corresponding excess risk error bounds derived in tandem to our generalization error results. Trade-offs between learning rates and corresponding achievable bounds are also highlighted.

2 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 $R^* \triangleq \inf_w R(w)$. Since the distribution $\mathcal{D}$ is not known, we consider the empirical risk

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

and the corresponding empirical risk minimization (ERM) problem to find $W_S^* \in \arg\min_w R_S(w)$ (assuming minimizers on data exist for simplicity). 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} [7, Lemma 5.1], [46]

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

For the rest of the paper we assume that the loss is smooth and non-negative. This is the only required assumption on the loss function.

Assumption 1 (β-Lipshitz Gradient) *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 assume that the model has sufficient capacity to interpolate the data-set. Recall, that this is a property of the model, and holds independently of the algorithm and the number of iterations.

Assumption 2 (Model Capacity) *For almost every $S \in \mathbb{Z}^n$, it is true that $R_S(W_S^*) = 0$. Equivalently, it holds that $\mathbb{E}[R_S(W_S^*)] = 0$.*

Remark 1 *The model capacity assumption corresponds to the so-called interpolation setting and characterizes the minimal required capacity such that the model can memorize the data set. It is satisfied in a large class of models with nonconvex loss functions, when the number of parameters are at least the number of the data set size ($d \geq n$). It is well known that this property often appears while training neural networks [5, Theorem 4.1], [6]. As we show, the generalization error gap closes*

[2] In fact, the latter are also essential for bounding the optimization error term that also appears in the generalization error, see Theorem 13, and Theorem 17.

with sharp rates of the order $\mathcal{O}(1/n^2)$ in the PL and strongly convex cases. We conjecture that sharp generalization error and excess rates of the order $\mathcal{O}(1/n^2)$ under fast training (number of iteration $T = C\log(n)$, step size $\eta_t = 1/\beta$) can only be achieved in such an interpolation regime (see also [47] discussing the necessity of Assumption 2 in the setting of linear regression).

In the next section we provide a general theorem for the generalization error that holds for any symmetric deterministic algorithm and any smooth loss under memorization of the data-set.

3 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 i^{th} random element. Recall that an algorithm is symmetric if the output remains unchanged under permutations of vector input. Then [42, 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 $z_1, \ldots, z_n, z_1', \ldots, z_n'$. We define the model parameters $W_t, W_t^{(i)}$ evaluated at time t with corresponding inputs S and $S^{(i)}$. Further, it is true that for any $i, j \in \{1, \ldots, n\}$

$$\mathbb{E}[f(A(S); z_i)] = \mathbb{E}[f(A(S); z_j)] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[f(A(S); z_k)] = \mathbb{E}[R_S(A(S))]. \tag{4}$$

We show (4) 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.[3] For brevity, we also provide the next definition.

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

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 under Memorization) *Let $f(\cdot\,; z)$ be non-negative β-smooth loss for any $z \in \mathcal{Z}$. If $\mathbb{E}[R_S(W_S^*)] = 0$, then for any symmetric deterministic algorithm $A(\cdot)$ the generalization error is bounded as*

$$|\epsilon_{\text{gen}}| \leq 2\sqrt{2\beta\epsilon_{\text{opt}}\mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right]} + 2\beta\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2]. \tag{5}$$

Inequality (5) holds for any symmetric algorithm and smooth loss. In fact Theorem 3 shows that algorithm stability and a small expected optimization error at termination sufficiently provide an upper bound on the generalization error for smooth, but 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 or μ-PL objectives.

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

Proof of Theorem 3. The β-smooth property of $f(\cdot\,; z)$ for all $z \in \mathcal{Z}$ gives

$$f(A(S^{(i)}); z) - f(A(S); z) \leq \langle A(S^{(i)}) - A(S), \nabla f(A(S); z)\rangle + \frac{\beta \|A(S^{(i)}) - A(S)\|_2^2}{2}. \tag{6}$$

The expression $\epsilon_{\text{gen}} = \mathbb{E}[f(A(S^{(i)}); z_i) - f(A(S); z_i)]$ and the inequality (6) give

$$\epsilon_{\text{gen}} \leq \mathbb{E}\left[\langle A(S^{(i)}) - A(S), \nabla f(A(S); z_i)\rangle + \frac{\beta \|A(S^{(i)}) - A(S)\|_2^2}{2}\right] \tag{7}$$

We find an upper bound for the expectation of the inner product in (7) by applying Cauchy-Schwartz inequality as

$$\begin{aligned}
&\mathbb{E}\left[\langle A(S^{(i)}) - A(S), \nabla f(A(S); z_i)\rangle\right] \\
&\leq \mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2 \|\nabla f(A(S); z_i)\|_2\right] &(8)\\
&\leq \sqrt{\mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right] \mathbb{E}\left[\|\nabla f(A(S); z_i)]\|_2^2\right]}, &(9)
\end{aligned}$$

here we use the inequalities $\langle a, b\rangle \leq \|a\|_2 \|b\|_2$ and $\mathbb{E}^2[XY] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$ to derive the bounds in (8) and (9) respectively. By combining the inequalities (7) and (9) we find that for any $i \in \{1, \ldots, n\}$ it is true that

$$\epsilon_{\text{gen}} \leq \sqrt{\mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right] \mathbb{E}\left[\|\nabla f(A(S); z_i)\|_2^2\right]} + \frac{\beta}{2}\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2]. \tag{10}$$

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

$$-\epsilon_{\text{gen}} \leq \sqrt{\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right] \mathbb{E}\left[\|\nabla f(A(S^{(i)}); z_i)]\|_2^2\right]} + \frac{\beta}{2}\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]. \tag{11}$$

Then we find an upper bound on $\mathbb{E}[\|\nabla f(A(S^{(i)}); z_i)]\|_2^2]$ as follows

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(A(S^{(i)}); z_i)]\|_2^2\right] \\
&= \mathbb{E}\left[\|\nabla f(A(S^{(i)}); z_i) - \nabla f(A(S); z_i) + \nabla f(A(S); z_i)\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f(A(S^{(i)}); z_i) - \nabla f(A(S); z_i)]\|_2^2 + \|\nabla f(A(S); z_i)\|_2^2\right] \\
&\leq 2\beta^2 \mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2] + 2\mathbb{E}[\|\nabla f(A(S); z_i)]\|_2^2]. &(12)
\end{aligned}$$

The inequality (12) holds because of the β-smoothness of the loss. Additionally,

$$\begin{aligned}
&\sqrt{2\beta^2\mathbb{E}^2\left[\|A(S) - A(S^{(i)})\|_2^2\right] + 2\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right] \mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} \\
&\leq \sqrt{2\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right] \mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} + \sqrt{2}\beta\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]. &(13)
\end{aligned}$$

We combine (11), (12) and (13) to find

$$-\epsilon_{\text{gen}} \leq \sqrt{2\mathbb{E}\left[\|A(S) - A(S^{(i)})\|_2^2\right] \mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} + 2\beta\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2]. \tag{14}$$

Finally, through the inequalities (10) and (14) we find

$$|\epsilon_{\text{gen}}| \leq \sqrt{2\mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right] \mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2]} + 2\beta\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \tag{15}$$

We use the self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ [48, Lemma 3.1], to show

$$\|\nabla f(A(S); z_i)\|_2^2 \leq 4\beta f(A(S); z_i). \tag{16}$$

The last display, Assumption 2 and equality (4) give

$$\begin{aligned}\mathbb{E}[\|\nabla f(A(S); z_i)\|_2^2] \leq 4\beta \mathbb{E}[f(A(S); z_i)] &= 4\beta \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(A(S); z_i)] \\ &= 4\beta \mathbb{E}[R_S(A(S))] \\ &= 4\beta \left(\mathbb{E}[R_S(A(S))] - \mathbb{E}[R_S(W_S^*)]\right) \\ &= 4\beta \epsilon_{\text{opt}}. \end{aligned} \tag{17}$$

We combine the inequalities (15) and (17) to find

$$|\epsilon_{\text{gen}}| \leq 2\sqrt{2\beta\epsilon_{\text{opt}} \mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right]} + 2\beta \mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \tag{18}$$

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

Recall that the proof of Theorem 3 requires the symmetry of the algorithm and the Assumptions 1 & 2. This technique provides a way to decompose the expected inner product of the output stability and the gradient of the loss at termination into a product of the output stability multiplied by the expected optimization error. Through the bound of Theorem 3, we later provide generalization error guarantees by independently studying the expected output stability and the expected optimization error at termination (Definition 2).

Next we provide a corollary as a byproduct of Theorem 3, which holds for any symmetric algorithm, and μ-PL objective. Recall that the objective function is (in expectation) μ-PL if $\mathbb{E}[\|\nabla R_S(w)\|_2^2] \geq 2\mu \mathbb{E}[R_S(w) - R_S^*]$ for all $w \in \mathbb{R}^d$. For brevity, we also introduce some notation: 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 $\tilde{c} \triangleq \mathbb{E}[R_S(\pi_S) + R(\pi_S)]$.

Corollary 4 *For any symmetric algorithm, non-negative β-smooth loss function $f(\cdot; z)$ for all $z \in \mathcal{Z}$ and μ-PL objective such that $\mathbb{E}[R_S(W_S^*)] = 0$, it is true that*

$$|\epsilon_{\text{gen}}| \leq \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} + \frac{44\beta}{\mu}\epsilon_{\text{opt}}. \tag{19}$$

Corollary 4 provides a tighter version of the corresponding bound in prior works [38, Theorem 1] and [33]. In fact, the sample complexity dominant term, namely $1/n$ is multiplied by the square root of the expected optimization error. Thus for algorithms that converge at minimizers inequality (19) gives significantly tighter bounds. For instance, by applying Corollary 4, we derive generalization error bounds of the order $\mathcal{O}(1/n^2)$ for $T \geq 2\log(n)/\log\left(\beta/(\beta-\mu)\right)$ number of iterations as we later show (we refer the reader to Theorem 10). We provide the proof of Corollary 4 in Appendix C.

Remark 5 *As a consequence of Corollary 4 and the decomposition (2), it follows that for a non-negative smooth loss and a μ-PL objective in the interpolation regime the excess risk is bounded as*

$$\epsilon_{\text{excess}} \leq \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} + \frac{45\beta}{\mu}\epsilon_{\text{opt}}. \tag{20}$$

We note that a closely related result to that of Remark 5 has been shown in [38] for the SGD algorithm. In fact, [38, Therem 7] requires the interpolation assumption and an additional assumption, namely the inequality $\beta \leq n\mu/4$, to hold simultaneously. However, if $\beta \leq n\mu/4$ and $\mathbb{E}[R_S(\pi_S)] = 0$ (interpolation assumption), then [38, 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 $\pi_S \in \arg\min R_S(\cdot)$. Together with the interpolation assumption this case gives $\tilde{c} = 0$ in (20), and thus $\epsilon_{\text{excess}} \leq 45\beta\epsilon_{\text{opt}}/\mu$. We may combine the last inequality with known convergence rates for (stochastic) gradient descent [6] to find explicit excess risk guarantees or to recover the rate for SGD in [38, Therem 7]. We would like to emphasize that we do not require an assumption such as $\beta \leq n\mu/4$, and that our bounds are free of the related aforementioned issues.

4 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 (terminal) error ϵ_{opt}. These quantities will prominently appear in our analysis and results.

Definition 6 (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*

$$\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2]. \tag{21}$$

The ϵ_{path} term expresses the path-dependent quantity that appears in the generalization bounds in our results[4]. 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 [7, 45] 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 we derive explicit rates for both ϵ_{path} and ϵ_{opt} to characterize the generalization error.

4.1 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 7 (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*

$$\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2] \leq \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1 + \beta\eta_j)^2 . \tag{22}$$

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

The expected output stability in Theorem 7 is bounded by the product of the expected path error (Definition 6), 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 7, we derive the generalization error of full-batch GD for smooth nonconvex losses.

Theorem 8 *Assume that the (possibly) nonconvex loss $f(\cdot, z)$ is β-smooth for all $z \in \mathcal{Z}$ and $\mathbb{E}[R_S(W_S^*)] = 0$. 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 $\epsilon \triangleq \beta C < 1$. Then the generalization error of full-batch GD is bounded by*

$$\begin{aligned}|\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{2}}{n}\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}(eT)^{\epsilon}\min\left\{\epsilon+\frac{1}{2}, \epsilon\log(eT)\right\}^{\frac{1}{2}} + 8\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon}\min\left\{\epsilon+\frac{1}{2}, \epsilon\log(eT)\right\} \\ &\leq \frac{4\sqrt{3}(eT)^{\epsilon}}{n}\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}} + 12\frac{(eT)^{2\epsilon}}{n^2}\epsilon_{\text{path}}. \end{aligned} \tag{23}$$

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 8.

Corollary 9 *The generalization error of full-batch GD in Theorem 8 can be further bounded as follows*

$$|\epsilon_{\text{gen}}| \leq \left(\frac{8\sqrt{3}}{n}\sqrt{\log(eT)}(eT)^{\epsilon} + \frac{48}{n^2}\log(eT)(eT)^{2\epsilon}\right)\mathbb{E}[R_S(W_1)]. \tag{24}$$

The inequality (23) in Theorem 8 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 (24). We prove Theorem 7, Theorem 8 and Corollary 9 in Appendix B. By taking into account known convergence guarantees on the optimization error, we can also derive excess risk guarantees. As a consequence of the Corollary 4, we derive generalization error and excess risk guarantees for nonconvex losses $f(\cdot; z)$ by assuming that the PL condition holds for the objective function $R_S(\cdot)$, as $\mathbb{E}[\|\nabla R_S(w)\|_2^2] \geq 2\mu\mathbb{E}[R_S(w) - R_S^*]$ for all $w \in \mathbb{R}^d$.

Theorem 10 *Let the loss function $f(\cdot; z)$ be non-negative, nonconvex and β-smooth for all $z \in \mathcal{Z}$. Further, define the constant $c \triangleq 44\max\{\mathbb{E}[R_S(\pi_S) + R(\pi_S)], \mathbb{E}[R_S(W_1) - R_S^*]\}$ and let the objective be μ-PL such that $\mathbb{E}[R_S^*(W_S^*)] = 0$. 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*

$$|\epsilon_{\text{gen}}| \leq \frac{c\beta}{\mu}\frac{\left(1-\frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{c\beta^2}{n^2\mu^2} + \frac{c\beta}{\mu}\left(1-\frac{\mu}{\beta}\right)^{T}. \tag{25}$$

Additionally, the choice $T = 2\log(n)/\log\left(\frac{\beta}{\beta-\mu}\right)$ guarantees that

$$|\epsilon_{\text{gen}}| \leq \frac{1}{n^2}\frac{c\beta}{\mu}\left(2+\frac{\beta}{\mu}\right). \tag{26}$$

The linear convergence rate [49] for a step-size choice $\eta_t = 1/\beta$ provides the following optimization error bound $\epsilon_{\text{opt}} \leq (1-\mu/\beta)^T\mathbb{E}[R_S(W_1) - R_S^*]$, and combined with Theorem 10 gives exceptional excess risk guarantees that are similar to the strongly convex setting. For the proof of Theorem 10 we refer the reader to Appendix C.

Theorem 11 *Let the loss function $f(\cdot; z)$ be non-negative, nonconvex and β-smooth for all $z \in \mathcal{Z}$. Further, define the constant $c \triangleq 44 \max\{\mathbb{E}[R_S(\pi_S) + R(\pi_S)], \mathbb{E}[R_S(W_1) - R_S^*]\}$ and let the objective be μ-PL. Then the excess risk of the full-batch GD with step-size choice $\eta_t = 1/\beta$ and T total number of iterations is bounded as follows*

$$\epsilon_{\text{excess}} \leq \frac{c\beta}{\mu} \frac{\left(1 - \frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{c\beta^2}{n^2\mu^2} + c\left(\frac{\beta}{\mu} + 1\right)\left(1 - \frac{\mu}{\beta}\right)^T. \tag{27}$$

Additionally, the choice $T = 2\log(n)/\log\left(\frac{\beta}{\beta-\mu}\right)$ gives

$$\epsilon_{\text{excess}} \leq \frac{1}{n^2} \frac{c\beta}{\mu}\left(3 + \frac{\beta}{\mu}\right). \tag{28}$$

As a result of Theorem 10 and Theorem 11, the generalization error and excess risk of full-batch GD decay exponentially with T. Recent works [5, 6] have shown that the PL condition holds in deep neural networks. Theorem 11 verifies the efficient generalization error performance that have been observed in deep neural networks under the large dimensionality regime [34]. The exponential decay with respect to the total number of iteration, verifies that longer training decreases the generalization error, and shows that memorization (interpolation of the data, Assumption 2) closes the generalization error gap, as it has been observed in practice [34]. Additionally, the dependence of the generalization error with respect to the optimization error in Corollary 4 (and inequality (25)) indicates that full-batch GD generalizes better than SGD and other stochastic variants.

4.2 Convex Loss

Herein, we provide generalization error guarantees for convex losses. Starting from the stability bound of the output of the algorithm, we show that the dependence on the learning is weaker than that of the nonconvex case. That dependence and the fast convergence to the minimum guarantee tighter generalization error bounds with a larger fixed learning (for instance $\eta_t = 1/\beta$), as we later show. Together with the generalization error, they imply an excess risk bound using the decomposition (2). We refer the reader to the Table 2 for a summary of the excess risk guarantees. We continue by providing the stability bound for convex losses.

Theorem 12 (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 learning rate $\eta_t < 2/\beta$, for all $t \leq T + 1$. Then for outputs of the algorithm $W_{T+1} \equiv A(S)$, $W_{T+1}^{(i)} \equiv A(S^{(i)})$ it is true that*

$$\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2] \leq 4\frac{\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t. \tag{29}$$

In the convex case, the expected output stability (inequality (29)) is bounded by the product of the expected path error, the number of samples term $2/n^2$ and the accumulated learning rate. To guarantee descent at each iteration, recall that we choose $\eta_t < 2/\beta$. By combining Theorem 3 and Theorem 12, we show the next generalization error bound.

Theorem 13 *Let the loss function $f(\cdot, z)$ be convex and β-smooth for all $z \in \mathcal{Z}$ and $\mathbb{E}[R_S(W_S^*)] = 0$. Consider the full-batch GD where T denotes the total number of iterates. We chose the learning rate such that $\eta_t < 2/\beta$, for all $t \leq T + 1$. Then the generalization error of full-batch GD is bounded by*

$$|\epsilon_{\text{gen}}| \leq \frac{4\sqrt{2\beta\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n} \sqrt{\sum_{t=1}^{T} \eta_t} + 8\beta\frac{\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t. \tag{30}$$

We provide the proof of Theorem 12 and Theorem 13 in Appendix D. Similar to the nonconvex case (Theorem 8), the bound in Theorem 13 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 (30), 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. The standard choice of $\eta_t \leq 1/\beta$ (that also guarantees descent direction) gives the next known bound on the optimization error [2].

Lemma 14 *If $f(\cdot; z)$ is a convex and β-smooth function and $\eta_t \leq 1/\beta$, then*

$$\epsilon_{\text{opt}} = \mathbb{E}[R_S(A(S)) - R_S(W_S^*)] \leq \frac{\mathbb{E}[\|W_1 - W_S^*\|_2^2]}{\sum_{t=1}^{T} \eta_t \left(1 - \frac{\beta \eta_t}{2}\right)}. \tag{31}$$

Through Theorem 13 and Lemma 14, we derive explicit generalization error bounds for certain choices of the learning rate. In fact, we consider the standard choice $\eta_t = 1/\beta$ in the next result.

Theorem 15 *Let the loss function $f(\cdot, z)$ be convex and β-smooth for all $z \in \mathcal{Z}$ and $\mathbb{E}[R_S(W_S^*)] = 0$. If $\eta_t = 1/\beta$ for all $t \in \{1, \ldots, T\}$, then*

$$|\epsilon_{\text{gen}}| \leq \left(\frac{\sqrt{2 \log(eT)}}{n} + 4 \frac{T \log(eT)}{n^2} \right) 16 \beta \mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{32}$$

By combining the inequalities (32) with the corresponding optimization error (Lemma 14), inequality (2) gives the corresponding excess risk bounds found in Table 2. Through the Theorem 13 and Lemma 14, we observe that the fixed step-size choice $1/\beta$ excels the decreasing learning rate option, since the first provides smaller excess risk bounds than the second (for any $T \leq n$). Thus choices of decreasing step sizes are out of interest.

4.3 Strongly Convex Loss

For strongly convex loss functions, the stability and generalization error bounds are tighter than those of the convex loss setting. Further, the faster convergence also provides tighter bounds for the excess risk (see Table 2). In fact, full-batch GD generalizes better by training longer. This phenomenon has been observed by Hoffer et al. [34]. We proceed by providing the stability bound of the output, and then the generalization error guarantees.

Theorem 16 (Stability Error — Strongly Convex Loss) *Assume that the γ-strongly 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 $\eta_t \leq 2/(\beta + \gamma)$ denotes the learning rate, for all $t \leq T$. Then for outputs of the algorithm $A(S)$, $A(S^{(i)})$ it is true that*

$$\mathbb{E}[\|A(S) - A(S^{(i)})\|_2^2] \leq 4 \frac{\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2}\right). \tag{33}$$

By comparing the stability guarantee of Theorem 12 with Theorem 16, 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 E) that the ϵ_{path} term is smaller in the strongly convex case. Similar to the nonconvex and convex loss cases, Theorem 3 and the stability error bound in Theorem 16 provide the generalization error bound for strongly convex losses.

Theorem 17 *Let the loss function $f(\cdot, z)$ be γ-strongly convex and β-smooth for all $z \in \mathcal{Z}$ and $\mathbb{E}[R_S(W_S^*)] = 0$. 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*

$$|\epsilon_{\text{gen}}| \leq \frac{4\sqrt{2\beta\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)} + 8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right) \tag{34}$$

We prove Theorem 16 and Theorem 17 in Appendix E. Recall that the sum product term in the inequality (34) is smaller than the summation of the learning rates in Theorem 13. This attribute together with the tighter optimization error bound, provide a smaller excess risk than those of the convex losses. As in 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. For fixed and decreasing learning rate, the optimization error bounds appear in the next lemma.

Lemma 18 ([2, Theorem 2.1.14]) *If $f(\cdot; z)$ is a γ-strongly convex and β-smooth function and $\eta_t = 2/(\beta+\gamma)$, then*

$$\epsilon_{\text{opt}} \leq \frac{\beta}{2}\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{35}$$

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

$$\epsilon_{\text{opt}} \leq \frac{\beta}{2}T^{-\frac{2c\beta\gamma}{\beta+\gamma}}\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{36}$$

By combining Theorem 17, Lemma 18, and Lemma 20, (Appendix A.1), we derive generalization error guarantees for fixed and decreasing step sizes as follows in the next result.

Theorem 19 *Let the loss function $f(\cdot, z)$ be γ-strongly convex and β-smooth for all $z \in \mathcal{Z}$ and $\mathbb{E}[R_S(W_S^*)] = 0$. Define $\Delta \triangleq \beta/\gamma(e^{4\gamma/(\beta+\gamma)} - 1)^{1/2}$, and set the learning rate to $\eta_t = 2/(\beta+\gamma)$. Then it is true that*

$$|\epsilon_{\text{gen}}| \leq \left(\frac{8\exp\left(\frac{-2\gamma}{\beta+\gamma}T\right)}{n}\Delta + \frac{32}{n^2}\Delta^2\right)\beta\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{37}$$

The inequalities (37) shows that the generalization errors decay when T grows. Specifically, if we choose $T = (\beta/\gamma + 1)\log(n)/2$ in (37), then we find generalization error bounds of the order $\mathcal{O}(1/n^2)$ (see Table 1). To derive the excess risk bounds for strongly convex losses, we consider the excess risk decomposition (2) and we use generalization and optimization error bounds as appeared above (see also Lemma 18). Specifically, for the constant learning rate, we derive the excess risk through (37) and (35). The generalization error for the decreasing learning may be derived similarly, (through Theorem 17 and the corresponding optimization error in (36)), however it gives significantly looser generalization and excess risk bounds than the constant step size case and remains out of interest. We refer the reader to Tables 1 and 2 for a summary of the results.

5 Conclusion

In this paper we addressed the generalization error and excess risk guarantees of deterministic training on smooth losses in the interpolation setting. In particular, we provided full-batch GD sharp generalization and excess risk bounds. At the heart of our analysis is a novel technique, showing that the average algorithmic output stability and the gradient of loss at termination ensure generalization. Then by exploiting this sufficient condition, we showed a decomposition of the generalization error bound in terms of the number of samples, the learning rate, the number of iterations, a path-dependent quantity and the optimization error at termination. That decomposition allowed us to further explore the generalization ability of full-batch GD for different types of loss functions. Specifically, for nonconvex, convex and strongly convex smooth (possibly non-Lipschitz) losses, we derived upper bounds on the generalization error and excess risk.

In particular, for nonconvex losses, we derived generalization error bounds with decreasing learning rate, as well as excess risk guarantees provided the objective satisfies the PL condition. Through the aforementioned technique and by taking into account fast optimization error convergences of the full-batch GD for convex and strongly convex losses, we provided sharp bounds on the generalization and excess risk for constant and decreasing step sizes. Our theoretical results shed light on the recent empirical observations that full-batch gradient descent generalizes and stochastic training procedures are not only unnecessary but may even lead to higher generalization error.

References

[1] 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.

[2] 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.

[3] 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. arXiv:https://doi.org/10.1137/040605266, doi:10.1137/040605266.

[4] 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, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL: https://proceedings.mlr.press/v49/lee16.html.

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

[6] 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. Special Issue on Harmonic Analysis and Machine Learning. URL: https://www.sciencedirect.com/science/article/pii/S106352032100110X, doi:https://doi.org/10.1016/j.acha.2021.12.009.

[7] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings*

of Machine Learning Research, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL: https://proceedings.mlr.press/v48/hardt16.html.

[8] 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, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/05a624166c8eb8273b8464e8d9cb5bd9-Paper.pdf.

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

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

[11] Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $\mathcal{O}(1/n)$. *arXiv preprint arXiv:2103.12024*, 2021. URL: https://arxiv.org/abs/2103.12024.

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

[13] 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, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/qian19b.html.

[14] 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, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/8c01a75941549a705cf7275e41b21f0d-Paper.pdf.

[15] 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, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Paper.pdf.

[16] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5809–5819. PMLR, 13–18 Jul 2020. URL: https://proceedings.mlr.press/v119/lei20c.html.

[17] 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, 2021. URL: http://jmlr.org/papers/v22/19-716.html.

[18] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21216–21228. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/b1301141feffabac455e1f90a7de2054-Paper.pdf.

[19] 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, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21236–21246. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/f3173935ed8ac4bf073c1bcd63171f8a-Paper.pdf.

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

[21] 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}.

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

[23] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27290–27304. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf.

[24] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26523–26535. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/df1f1d20ee86704251795841e6a9405a-Paper.pdf.

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

[26] 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, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 605–638. PMLR, 06–09 Jul 2018. URL: https://proceedings.mlr.press/v75/mou18a.html.

[27] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019. URL: https://arxiv.org/abs/1902.00621.

[28] 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, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf.

[29] Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of SGD: Tightness Analysis and Improved Bounds. *arXiv preprint arXiv:2102.05274*, 2021. URL: https://arxiv.org/abs/2102.05274.

[30] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19836–19846. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/a4ee59dd868ba016ed2de90d330acb6a-Paper.pdf.

[31] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26080–26090. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/db2b4182156b2f1f817860ac9f409ad7-Paper.pdf.

[32] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24222–24234. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/cb77649f5d53798edfa0ff40dae46322-Paper.pdf.

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

[34] 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, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/a5e0ff62be0b08456fc7f1e88812af3d-Paper.pdf.

[35] Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. *arXiv preprint arXiv:2109.14119*, 2021. URL: https://arxiv.org/abs/2109.14119.

[36] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25033–25043. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/d27b95cac4c27feb850aaa4070cc4675-Paper.pdf.

[37] 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, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3526–3545. PMLR, 15–19 Aug 2021. URL: https://proceedings.mlr.press/v134/neu21a.html.

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

[39] 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.

[40] Konstantinos E Nikolakakis, Farzin Haddadpour, Dionysios S Kalogerias, and Amin Karbasi. Black-box generalization. *arXiv preprint arXiv:2202.06880*, 2022. URL: https://arxiv.org/abs/2202.06880.

[41] Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning. *arXiv preprint arXiv:2203.10036*, 2022. URL: https://arxiv.org/abs/2203.10036.

[42] 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.

[43] 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.

[44] Gergely Neu and Gábor Lugosi. Generalization bounds via convex analysis. *arXiv preprint arXiv:2202.04985*, 2022. URL: https://arxiv.org/abs/2202.04985.

[45] 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.

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

[47] Chen Cheng, John Duchi, and Rohith Kuditipudi. Memorize to Generalize: On the necessity of interpolation in high dimensional linear regression. *arXiv preprint arXiv:2202.09889*, 2022. URL: https://arxiv.org/abs/2202.09889.

[48] 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, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

[49] 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*, pages 795–811. Springer, 2016. URL: https://doi.org/10.1007/978-3-319-46128-1_50.

A Proofs

We provide the proofs of the results in these sections. We start by providing further 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).

A.1 Sum Product Terms in the Stability Bounds

Herein we show a lemma for the sum product terms associated with learning rate in Theorem 7 and Theorem 16. Then we will apply that lemma to derive the corresponding stability error bounds.

Lemma 20 *The following are true:*

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

$$\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2}\right) = 2\frac{1 - \left(1 - \frac{C\gamma}{2}\right)^T}{\gamma}, \tag{38}$$

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

$$\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2}\right) \leq C \log\left(e^2 \lceil \beta/\gamma \rceil\right). \tag{39}$$

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

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

Proof.

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

$$\begin{aligned}\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2}\right) &= C \sum_{t=1}^{T} \left(1 - \frac{C\gamma}{2}\right)^{T-t} = C \left(1 - \frac{C\gamma}{2}\right)^{T} \sum_{t=1}^{T} \left(1 - \frac{C\gamma}{2}\right)^{-t} \\ &= 2C\frac{1 - \left(1 - \frac{C\gamma}{2}\right)^T}{C\gamma} = 2\frac{1 - \left(1 - \frac{C\gamma}{2}\right)^T}{\gamma},\end{aligned}$$

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

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

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

$$\leq C\left(1 + \log(\lceil \beta/\gamma \rceil) + \left[1 - \left\lceil \frac{\beta}{T\gamma} \right\rceil\right]_+\right) \leq C \log\left(e^2 \lceil \beta/\gamma \rceil\right).$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2 &= \sum_{t=1}^{T} \frac{C}{t} \prod_{j=t+1}^{T} \left(1 + \beta\frac{C}{j}\right)^2 \\
&\leq \sum_{t=1}^{T} \frac{C}{t} \prod_{j=t+1}^{T} \exp\left(2\beta\frac{C}{j}\right) \\
&= \sum_{t=1}^{T} \frac{C}{t} \exp\left(2\beta \sum_{j=t+1}^{T} \frac{C}{j}\right) \\
&\leq \sum_{t=1}^{T} \frac{C}{t} \exp\left(2C\beta\left(\log(T) + 1 - \log(t+1)\right)\right) \\
&= Ce^{2C\beta}T^{2C\beta} \sum_{t=1}^{T} \frac{1}{t}\frac{1}{(t+1)^{2C\beta}} \\
&\leq Ce^{2C\beta}T^{2C\beta} \sum_{t=1}^{T} \frac{1}{t}\frac{1}{(t+1)^{2C\beta}} \\
&\leq Ce^{2C\beta}T^{2C\beta} \sum_{t=1}^{T} \frac{1}{t^{1+2C\beta}} \qquad (41) \\
&= Ce^{2C\beta}T^{2C\beta}\left(1 + \sum_{t=2}^{T} \frac{1}{t^{1+2C\beta}}\right) \\
&\leq Ce^{2C\beta}T^{2C\beta}\left(1 + \int_1^T \frac{1}{x^{1+2C\beta}}dx\right) \\
&= Ce^{2C\beta}T^{2C\beta}\left(1 + \frac{1}{2C\beta}\left(1 - T^{-2C\beta}\right)\right) \\
&= Ce^{2C\beta}T^{2C\beta}\left(1 + \frac{1}{2C\beta}\right) - Ce^{2C\beta} \\
&\leq Ce^{2C\beta}T^{2C\beta}\left(1 + \frac{1}{2C\beta}\right), \qquad (42)
\end{aligned}$$

  additionally $\sum_{t=1}^{T} 1/t \leq \log(eT)$, thus the term in the inequality (41) may be upper bounded by $Ce^{2C\beta}T^{2C\beta}\log(eT)$ for any $T \in \mathbb{N}$, and we conclude that

$$\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2 \leq Ce^{2C\beta}T^{2C\beta} \min\left\{1 + \frac{1}{2C\beta}, \log(eT)\right\}. \qquad (43)$$

The last inequality completes the proof. □

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

B Nonconvex Loss: Proof of Theorem 7 & Theorem 8

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

$$W_{t+1} = W_t - \frac{\eta_t}{n} \sum_{j=1}^{n} \nabla f(W_t, z_j), \tag{44}$$

$$W_{t+1}^{(i)} = W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) - \frac{\eta_t}{n} \nabla f(W_t^{(i)}, z_i'). \tag{45}$$

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

$$\begin{aligned}
&\|W_{t+1} - W_{t+1}^{(i)}\|_2 \\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \Big\| \sum_{j=1, j\neq i}^{n} \Big(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \Big) \Big\|_2 + \frac{\eta_t}{n} \|\nabla f(W_t, z_i) - \nabla f(W_t^{(i)}, z_i')\|_2 \\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \|\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\|_2 + \frac{\eta_t}{n} \Big(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \Big) \\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t (n-1)}{n} \beta \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \Big(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \Big) \qquad (46) \\
&= \Big(1 + \frac{n-1}{n} \beta \eta_t \Big) \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \Big(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \Big), \qquad (47)
\end{aligned}$$

inequality (46) comes from the smoothness of the loss. Then by solving the recursion we find

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

The last display gives

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

and by taking the expectation we find

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

$$\leq \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2 \,. \tag{48}$$

We evaluate the summation of the products in (48). Lemma 20 under the choice of decreasing learning rate $\eta_t \leq C/t \leq 2/\beta$ shows that

$$\sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2 \leq C e^{2C\beta} T^{2C\beta} \min\left\{1 + \frac{1}{2C\beta}, \log(eT)\right\}. \tag{49}$$

Through the inequalities (48), (49) and Theorem 3, we derive the bound on the generalization error as

$$\begin{aligned}
&|\epsilon_{\text{gen}}| \\
&\leq 2\sqrt{2\beta\epsilon_{\text{opt}}\mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right]} + 2\beta\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \\
&\leq \frac{4}{n}\sqrt{2\beta\epsilon_{\text{opt}}\epsilon_{\text{path}} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2} + 8\beta\frac{\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} (1+\beta\eta_j)^2 \\
&\leq \frac{4}{n}\sqrt{2C\beta\epsilon_{\text{opt}}\epsilon_{\text{path}}} e^{C\beta} T^{C\beta} \min\left\{1 + \frac{1}{2C\beta}, \log(eT)\right\}^{\frac{1}{2}} + 8C\beta\frac{\epsilon_{\text{path}}}{n^2} e^{2C\beta} T^{2C\beta} \min\left\{1 + \frac{1}{2C\beta}, \log(eT)\right\}
\end{aligned}$$

Under the choice $\eta_t \leq C/t < 1/\beta$ for all t, we choose $C < 1/\beta$ and define $\epsilon \triangleq \beta C < 1$, to get

$$\begin{aligned}
|\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{2}}{n}\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}(eT)^{\epsilon} \min\left\{\epsilon + \frac{1}{2}, \epsilon\log(eT)\right\}^{\frac{1}{2}} + 8\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon} \min\left\{\epsilon + \frac{1}{2}, \epsilon\log(eT)\right\} \\
&\leq \frac{4\sqrt{3}}{n}\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}(eT)^{\epsilon} + 12\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon}.
\end{aligned} \tag{50}$$

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 (50). We continue by proving the proof of Corollary 9.

B.1 Proof of Corollary 9.

The self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ [48, 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 2 and property (4) we find

$$\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \leq 4\beta\mathbb{E}[f(W_t, z_i)] = 4\beta\mathbb{E}[R_S(W_t)] = 4\beta\mathbb{E}[R_S(W_t) - R_S(W_S^*)]. \tag{51}$$

The definition of ϵ_{path} (Definition 6), and the decreasing learning rate ($\eta_t = C/t < 1/\beta t$) give

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

$$\leq 4\beta\mathbb{E}[R_S(W_1)]\sum_{t=1}^{T}\eta_t \tag{52}$$

$$< 4\mathbb{E}[R_S(W_1)]\sum_{t=1}^{T}\frac{1}{t}$$

$$\leq 4\mathbb{E}[R_S(W_1)]\log(eT), \tag{53}$$

and the inequality (52) holds since the learning rate $\eta_t < 2/\beta$ guarantees descent at each iteration. Similarly, $\epsilon_{\text{opt}} \triangleq \mathbb{E}[R_S(A(S)) - R_S(W_S^*)] \leq \mathbb{E}[R_S(W_1)]$. The last inequality together with the inequalities (53) and (50) give

$$|\epsilon_{\text{gen}}| \leq \frac{4\sqrt{3}}{n}\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}(eT)^{\epsilon} + 12\frac{\epsilon_{\text{path}}}{n^2}(eT)^{2\epsilon}$$

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

The last inequality provides the bound of the corollary.

C PL Objective and Nonconvex Loss

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 21 and Theorem 3 we derive generalization error algorithm for symmetric algorithms and smooth, 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. [38, Proof of Lemma B.2].

Lemma 21 *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*

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\mathbb{E}\left[R_S(\pi_S)\right] + \mathbb{E}[R(\pi_S)]\right). \tag{54}$$

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

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2]$$

$$\leq 4\mathbb{E}[\|A(S^{(i)}) - \pi_{S^{(i)}}\|_2^2] + 4\mathbb{E}[\|A(S) - \pi_S\|_2^2] + 2\mathbb{E}[\|\pi_{S^{(i)}} - \pi_S\|_2^2]$$

$$\leq \frac{8}{\mu}\mathbb{E}[R_{S^{(i)}}(A(S^{(i)})) - R_{S^{(i)}}^*] + \frac{8}{\mu}\mathbb{E}[R_S(A(S)) - R_S^*] + 2\mathbb{E}[\|\pi_{S^{(i)}} - \pi_S\|_2^2] \tag{55}$$

$$= \frac{16}{\mu}\epsilon_{\text{opt}} + 2\mathbb{E}[\|\pi_{S^{(i)}} - \pi_S\|_2^2]$$

$$\leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{4}{\mu}\left(\mathbb{E}[R_S(\pi_{S^{(i)}})] - \mathbb{E}[R_S(\pi_S)]\right), \tag{56}$$

the inequalities (55) and (56) come from the quadratic growth [49]. Recall that, the PL condition on the objective gives

$$\frac{1}{2\mu}\mathbb{E}[\|\nabla R_S(\pi_{S^{(i)}})\|_2^2] \geq \mathbb{E}[R_S(\pi_{S^{(i)}}) - R_S(\pi_S)]. \tag{57}$$

We combine the inequalities (56) and (57) to find

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{2}{\mu^2}\mathbb{E}[\|\nabla R_S(\pi_{S^{(i)}})\|_2^2]. \tag{58}$$

Also, it is true that

$$\begin{aligned}\|\nabla R_S(\pi_{S^{(i)}})\|_2^2 &= \|\nabla R_{S^{(i)}}(\pi_{S^{(i)}}) - \frac{1}{n}\nabla f(\pi_{S^{(i)}}; z_i') + \frac{1}{n}\nabla f(\pi_{S^{(i)}}; z_i)\|_2^2 \\ &= \frac{2}{n^2}\|\nabla f(\pi_{S^{(i)}}; z_i')\|_2^2 + \frac{2}{n^2}\|\nabla f(\pi_{S^{(i)}}; z_i)\|_2^2 \qquad (59) \\ &\leq \frac{4\beta}{n^2} f(\pi_{S^{(i)}}; z_i') + \frac{4\beta}{n^2} f(\pi_{S^{(i)}}; z_i), \qquad (60)\end{aligned}$$

equality (59) holds because $\nabla R_{S^{(i)}}(\pi_{S^{(i)}}) = 0$, and (60) holds for nonnegative losses [48, (Lemma 3.1]. Through (60) we find,

$$\begin{aligned}\mathbb{E}[\|\nabla R_S(\pi_{S^{(i)}})\|_2^2] &\leq \frac{4\beta}{n^2}\mathbb{E}[f(\pi_{S^{(i)}}; z_i')] + \frac{4\beta}{n^2}\mathbb{E}[f(\pi_{S^{(i)}}; z_i)] \qquad (61) \\ &= \frac{4\beta}{n^2}\mathbb{E}[f(\pi_S; z_i)] + \frac{4\beta}{n^2}\mathbb{E}[f(\pi_S; z_i')], \qquad (62)\end{aligned}$$

and the last equality holds because z_i, z_i' are exchangeable. We combine the inequalities (58) and (62) to find

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] &\leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(\pi_S; z_i)] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(\pi_S; z_i')\right) \qquad (63) \\ &= \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\mathbb{E}\left[R_S(\pi_S)\right] + \mathbb{E}[R(\pi_S)]\right). \qquad (64)\end{aligned}$$

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 $i \in \{1, \ldots, n\}$

$$\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \leq \frac{16}{\mu}\epsilon_{\text{opt}} + \frac{8\beta}{n^2\mu^2}\left(\mathbb{E}\left[R_S(\pi_S)\right] + \mathbb{E}[R(\pi_S)]\right). \tag{65}$$

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

Corollary 22 *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*

$$|\epsilon_{\text{gen}}| \leq \frac{8\beta\sqrt{\tilde{c}}}{n\mu}\sqrt{\epsilon_{\text{opt}}} + \frac{16\beta^2}{n^2\mu^2}\tilde{c} + \frac{44\beta}{\mu}\epsilon_{\text{opt}}. \tag{66}$$

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

$$|\epsilon_{\text{gen}}| \leq \frac{c\beta}{\mu}\frac{\left(1 - \frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{c\beta^2}{n^2\mu^2} + \frac{c\beta}{\mu}\left(1 - \frac{\mu}{\beta}\right)^T. \tag{67}$$

Proof. We apply Theorem 3 and Lemma 21 to find

$$\begin{aligned}
|\epsilon_{\text{gen}}| &\leq 2\sqrt{2\beta\epsilon_{\text{opt}}\mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right]} + 2\beta\mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2] \\
&\leq \left(\frac{8}{\sqrt{\mu}}\sqrt{\epsilon_{\text{opt}}} + \frac{4\sqrt{2\beta\mathbb{E}[R_S(\pi_S) + R(\pi_S)]}}{n\mu}\right)\sqrt{2\beta\epsilon_{\text{opt}}} + \frac{32\beta}{\mu}\epsilon_{\text{opt}} + \frac{16\beta^2}{n^2\mu^2}\mathbb{E}[R_S(\pi_S) + R(\pi_S)] \\
&= \frac{8\sqrt{2\beta}}{\sqrt{\mu}}\epsilon_{\text{opt}} + \frac{8\beta\sqrt{\mathbb{E}[R_S(\pi_S) + R(\pi_S)]}}{n\mu}\sqrt{\epsilon_{\text{opt}}} + \frac{32\beta}{\mu}\epsilon_{\text{opt}} + \frac{16\beta^2}{n^2\mu^2}\mathbb{E}[R_S(\pi_S) + R(\pi_S)] \\
&\leq \frac{8\beta\sqrt{\mathbb{E}[R_S(\pi_S) + R(\pi_S)]}}{n\mu}\sqrt{\epsilon_{\text{opt}}} + \frac{16\beta^2}{n^2\mu^2}\mathbb{E}[R_S(\pi_S) + R(\pi_S)] + \frac{44\beta}{\mu}\epsilon_{\text{opt}}. \qquad (68)
\end{aligned}$$

Further for the full-batch GD algorithm and μ-PL objective the expected optimization error is bounded as

$$\epsilon_{\text{opt}} \leq \left(1 - \frac{\mu}{\beta}\right)^T \mathbb{E}[R_S(W_1) - R_S^*]. \qquad (69)$$

Define the constants $\Lambda \triangleq \mathbb{E}[R_S(\pi_S) + R(\pi_S)]$ and $\Theta \triangleq \mathbb{E}[R_S(W_1) - R_S^*]$. The last inequality and (68) give

$$|\epsilon_{\text{gen}}| \leq 44\Theta\frac{\beta}{\mu}\left(1 - \frac{\mu}{\beta}\right)^T + \frac{8\beta\sqrt{\Theta\Lambda}}{\mu}\frac{\left(1 - \frac{\mu}{\beta}\right)^{T/2}}{n} + \frac{16\beta^2}{n^2\mu^2}\Lambda. \qquad (70)$$

Choose $T = 2\log(n)/\log\left(\frac{\beta}{\beta-\mu}\right)$, to get

$$|\epsilon_{\text{gen}}| \leq 44\Theta\frac{\beta}{\mu}\frac{1}{n^2} + \frac{8\beta\sqrt{\Theta\Lambda}}{\mu}\frac{1}{n^2} + \frac{16\beta^2}{n^2\mu^2}\Lambda = \frac{1}{n^2}\frac{\beta}{\mu}\left(44\Theta + 8\sqrt{\Theta\Lambda} + \frac{16\Lambda\beta}{\mu}\right). \qquad (71)$$

The last inequality and the definition $c \triangleq 44\max\{\Theta, \Lambda\}$ complete the proof. □

D Convex Loss: Proof of Theorem 12 and Theorem 13.

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 23 *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*

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

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

$$\|\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\|_2 \leq \beta\|W_t - W_t^{(i)}\|_2 \implies \qquad (73)$$

$$\|\sum_{j\in\mathcal{J}}\nabla f(W_t, z_j) - \sum_{j\in\mathcal{J}}\nabla f(W_t^{(i)}, z_j)\|_2 \leq \beta|\mathcal{J}|\|W_t - W_t^{(i)}\|_2. \qquad (74)$$

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

$$\sum_{j\in\mathcal{J}} \langle \nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)} \rangle \geq \frac{1}{\beta|\mathcal{J}|} \| \sum_{j\in\mathcal{J}} \nabla f(W_t, z_j) - \sum_{j\in\mathcal{J}} \nabla f(W_t^{(i)}, z_j) \|_2^2. \quad (75)$$

Then prove the inequality (72) as follows

$$\begin{aligned}
&\left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \Big(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \Big) \right\|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \langle \nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)} \rangle \quad (76) \\
&\quad + \frac{\eta_t^2}{n^2} \| \sum_{j=1, j\neq i}^{n} \Big(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \Big) \|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{n} \langle \sum_{j=1, j\neq i}^{n} \nabla f(W_t, z_j) - \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)} \rangle \\
&\quad + \frac{\eta_t^2}{n^2} \| \sum_{j=1, j\neq i}^{n} \nabla f(W_t, z_j) - \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) \|_2^2 \\
&\leq \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{\beta(n-1)n} \| \sum_{j=1, j\neq i}^{n} \nabla f(W_t, z_j) - \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) \|_2^2 \\
&\quad + \frac{\eta_t^2}{n^2} \| \sum_{j=1, j\neq i}^{n} \nabla f(W_t, z_j) - \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) \|_2^2 \quad (77) \\
&= \|W_t - W_t^{(i)}\|_2^2 + \frac{\eta_t}{n} \left(\frac{\eta_t}{n} - \frac{2}{\beta(n-1)} \right) \left\| \sum_{j=1, j\neq i}^{n} \Big(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \Big) \right\|_2^2 \\
&\leq \|W_t - W_t^{(i)}\|_2^2, \quad (78)
\end{aligned}$$

equation (76) holds from the expansion of the squared norm,(77) comes from the inequality (75). The inequality (78) holds under the choice $\eta_t < 2/\beta$ and completes the proof. □

Lemma 24 (Accumulated Path Error - Convex Loss) *Let the loss function $f(\cdot; z)$ be convex and β-smooth and $\eta_t = 1/\beta$. If Assumption 2 holds the expected path-error and the expected optimization error of the full-batch GD after T iterations are bounded as*

$$\epsilon_{\text{path}} \leq 8\beta \log(eT) \mathbb{E}[\|W_1 - W_S^*\|_2^2], \quad (79)$$

$$\epsilon_{\text{opt}} \leq \frac{2\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2]}{T}. \quad (80)$$

Proof. The self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ [48, 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 2 and property (4) we find

$$\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \leq 4\beta \mathbb{E}[f(W_t, z_i)] = 4\beta \mathbb{E}[R_S(W_t)] = 4\beta \mathbb{E}[R_S(W_t) - R_S(W_S^*)]. \quad (81)$$

Further, Lemma 14 and the choice of constant learning rate $\eta = 1/\beta$ give

$$\mathbb{E}[R_S(W_t) - R_S(W_S^*)] \leq \frac{\mathbb{E}[\|W_1 - W_S^*\|_2^2]}{\sum_{t'=1}^{t} \eta_{t'} \left(1 - \frac{\beta \eta_{t'}}{2}\right)} = \frac{2\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2]}{t}. \tag{82}$$

The definition of ϵ_{path} (Definition 6), the inequalities (81) and (82) and the constant learning rate ($\eta_t = 1/\beta$) give

$$\begin{aligned}
\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] &\leq 4\beta \sum_{t=1}^{T} \eta_t \mathbb{E}[R_S(W_t) - R_S(W_S^*)] \\
&\leq 8\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2] \sum_{t=1}^{T} \frac{1}{t} \\
&\leq 8\beta \mathbb{E}[\|W_1 - W_S^*\|_2^2] \log(eT).
\end{aligned} \tag{83}$$

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

D.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

$$W_{t+1} = W_t - \frac{\eta_t}{n} \sum_{j=1}^{n} \nabla f(W_t, z_j), \tag{84}$$

$$W_{t+1}^{(i)} = W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) - \frac{\eta_t}{n} \nabla f(W_t^{(i)}, z_i'). \tag{85}$$

Then for any $t \geq 1$

$$\begin{aligned}
&\|W_{t+1} - W_{t+1}^{(i)}\|_2 \\
&\leq \left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \right) \right\|_2 + \frac{\eta_t}{n} \|\nabla f(W_t, z_i) - \nabla f(W_t^{(i)}, z_i')\|_2 \\
&\leq \sqrt{\left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \right) \right\|_2^2} \\
&\quad + \frac{\eta_t}{n} \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right) \\
&\leq \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right).
\end{aligned} \tag{86}$$

The inequality (86) comes from Lemma 23. 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

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

$$\leq \frac{2}{n^2}\sum_{t=1}^{T}\eta_t\left(\|\nabla f(W_t,z_i)\|_2^2+\|\nabla f(W_t^{(i)},z_i')\|_2^2\right)\sum_{t=1}^{T}\eta_t. \tag{87}$$

Inequality (87) gives that for any $i\in\{1,\dots,n\}$

$$\begin{aligned}\mathbb{E}[\|W_{T+1}-W_{T+1}^{(i)}\|_2^2] &\leq \frac{2}{n^2}\sum_{t=1}^{T}\eta_t\left(\mathbb{E}[\|\nabla f(W_t,z_i)\|_2^2]+\mathbb{E}[\|\nabla f(W_t^{(i)},z_i')\|_2^2]\right)\sum_{t=1}^{T}\eta_t\\ &= \frac{4}{n^2}\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(W_t,z_i)\|_2^2]\sum_{t=1}^{T}\eta_t = \frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t. \end{aligned}\tag{88}$$

Recall that $W_{T+1}\equiv A(S)$ and $W_{T+1}^{(i)}\equiv A(S^{(i)})$. Theorem 3 and the inequality (88) give

$$\begin{aligned}|\epsilon_{\text{gen}}| &\leq 2\sqrt{2\beta\epsilon_{\text{opt}}\mathbb{E}\left[\|A(S^{(i)})-A(S)\|_2^2\right]}+2\beta\mathbb{E}[\|A(S^{(i)})-A(S)\|_2^2]\\ &\leq 2\sqrt{2\beta\epsilon_{\text{opt}}\frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t}+2\beta\frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\\ &= \frac{4\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{2\beta\sum_{t=1}^{T}\eta_t}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t. \end{aligned}\tag{89}$$

Under the choice of constant learning rate $\eta_t=1/\beta$, and if the Assumption 2 holds, then Lemma 24 together with (88) give $\epsilon_{\text{path}}\leq 8\beta\log(eT)\mathbb{E}[\|W_1-W_S^*\|_2^2]$, $\epsilon_{\text{opt}}\leq 2\beta\mathbb{E}[\|W_1-W_S^*\|_2^2]/T$. Thus

$$\mathbb{E}[\|A(S)-A(S^{(i)})\|_2^2]\leq\frac{32}{n^2}\beta\log(eT)\mathbb{E}[\|W_1-W_S^*\|_2^2]\sum_{t=1}^{T}\eta_t=\frac{32}{n^2}T\log(eT)\mathbb{E}[\|W_1-W_S^*\|_2^2], \tag{90}$$

and the inequality (89) gives

$$\begin{aligned}|\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{2\beta\sum_{t=1}^{T}\eta_t}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\\ &\leq \frac{4\sqrt{2\beta\mathbb{E}[\|W_1-W_S^*\|_2^2]8\beta\log(eT)\mathbb{E}[\|W_1-W_S^*\|_2^2]}}{n\sqrt{T}}\sqrt{2T}+8T\frac{8\beta\log(eT)\mathbb{E}[\|W_1-W_S^*\|_2^2]}{n^2}\\ &= \frac{16\beta\mathbb{E}[\|W_1-W_S^*\|_2^2]\sqrt{2\log(eT)}}{n}+64\beta\frac{T\log(eT)\mathbb{E}[\|W_1-W_S^*\|_2^2]}{n^2}\\ &= \left(\frac{\sqrt{2\log(eT)}}{n}+4\frac{T\log(eT)}{n^2}\right)16\beta\mathbb{E}[\|W_1-W_S^*\|_2^2]\end{aligned}$$

The last inequality completes the proof. □

E Strongly Convex Loss: Proof of Theorem 16 and Theorem 17

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

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

Proof. Under the assumption of β-Lipschitz gradients and γ-strong convexity of the loss, it is true that $\sum_{j\in\mathcal{J}} \nabla f(w, z_j)$ is $\beta|\mathcal{J}|$-Lipschitz and $\sum_{j\in\mathcal{J}} f(w, z_j)$ is $\gamma|\mathcal{J}|$-strongly convex. The co-coersivity of the sum of gradients gives

$$\begin{aligned}
&\langle \sum_{j\in|\mathcal{J}|} \nabla f(W_t, z_j) - \sum_{j\in|\mathcal{J}|} \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)}\rangle \\
&\geq \frac{\beta\gamma|\mathcal{J}|}{\beta+\gamma}\|W_t - W_t^{(i)}\|_2^2 + \frac{1}{(\beta+\gamma)|\mathcal{J}|}\|\sum_{j\in|\mathcal{J}|} \nabla f(W_t, z_j) - \sum_{j\in|\mathcal{J}|} \nabla f(W_t^{(i)}, z_j)\|_2^2. \qquad (91)
\end{aligned}$$

We start by expanding the squared norm as follows

$$\begin{aligned}
&\left\|W_t - W_t^{(i)} - \frac{\eta_t}{n}\sum_{j=1,j\neq i}^{n}\Big(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\Big)\right\|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{n}\sum_{j=1,j\neq i}^{n}\langle \nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)}\rangle \\
&\quad + \frac{\eta_t^2}{n^2}\|\sum_{j=1,j\neq i}^{n}\Big(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j)\Big)\|_2^2 \\
&= \|W_t - W_t^{(i)}\|_2^2 - 2\frac{\eta_t}{n}\langle \sum_{j=1,j\neq i}^{n}\nabla f(W_t, z_j) - \sum_{j=1,j\neq i}^{n}\nabla f(W_t^{(i)}, z_j), W_t - W_t^{(i)}\rangle \\
&\quad + \frac{\eta_t^2}{n^2}\|\sum_{j=1,j\neq i}^{n}\nabla f(W_t, z_j) - \sum_{j=1,j\neq i}^{n}\nabla f(W_t^{(i)}, z_j)\|_2^2 \qquad (92) \\
&\leq \|W_t - W_t^{(i)}\|_2^2 + \frac{\eta_t^2}{n^2}\|\sum_{j=1,j\neq i}^{n}\nabla f(W_t, z_j) - \sum_{j=1,j\neq i}^{n}\nabla f(W_t^{(i)}, z_j)\|_2^2 \\
&\quad - 2\frac{\eta_t}{n}\Bigg[\frac{\beta\gamma(n-1)}{\beta+\gamma}\|W_t - W_t^{(i)}\|_2^2 \\
&\qquad + \frac{1}{(\beta+\gamma)(n-1)}\|\sum_{j\in|\mathcal{J}|}\nabla f(W_t, z_j) - \sum_{j\in|\mathcal{J}|}\nabla f(W_t^{(i)}, z_j)\|_2^2\Bigg] \qquad (93) \\
&= \left(1 - 2\frac{\eta_t}{n}\frac{\beta\gamma(n-1)}{\beta+\gamma}\right)\|W_t - W_t^{(i)}\|_2^2 \\
&\quad + \frac{\eta_t}{n}\left(\frac{\eta_t}{n} - \frac{2}{(\beta+\gamma)(n-1)}\right)\|\sum_{j=1,j\neq i}^{n}\nabla f(W_t, z_j) - \sum_{j=1,j\neq i}^{n}\nabla f(W_t^{(i)}, z_j)\|_2^2 \\
&\leq \left(1 - 2\frac{\eta_t}{n}\frac{\beta\gamma(n-1)}{\beta+\gamma}\right)\|W_t - W_t^{(i)}\|_2^2. \qquad (94)
\end{aligned}$$

The inequality (92) comes from the convexity of the squared norm, we apply (91) to derive the inequality (93). Then inequality (94) holds under the choice $\eta_t \leq 2/(\beta+\gamma)$. Further, $2\beta \geq \beta+\gamma$, and for $n \geq 2$

$$2\frac{\eta_t}{n}\frac{\beta\gamma(n-1)}{\beta+\gamma} \geq \eta_t\gamma\frac{n-1}{n} \geq \frac{\eta_t\gamma}{2}. \qquad (95)$$

We combine (94) and (95) to derive the bound of the lemma. □

Lemma 26 (Accumulated Path Error - Strongly Convex Loss) *Let the loss function $f(\cdot; z)$ be γ-strongly convex and β-smooth and $\eta_t = 2/(\beta+\gamma)$. If Assumption 2 holds the expected path-error and the expected optimization error of the full-batch GD after T iterations are bounded as*

$$\epsilon_{\text{path}} \leq \frac{4\beta^2(e^{\frac{4\gamma}{\beta+\gamma}}-1)^{-1}}{\beta+\gamma}\mathbb{E}[\|W_1 - W_S^*\|_2^2], \tag{96}$$

$$\epsilon_{\text{opt}} \leq \frac{\beta}{2}\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{97}$$

Proof. The self-bounding property of the non-negative β-smooth loss function $f(\cdot; z)$ [48, 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 2 and property (4) we find

$$\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \leq 4\beta\mathbb{E}[f(W_t, z_i)] = 4\beta\mathbb{E}[R_S(W_t)] = 4\beta\mathbb{E}[R_S(W_t) - R_S(W_S^*)]. \tag{98}$$

Further, Lemma 18 and the choice of constant learning rate $\eta = 2/(\beta+\gamma)$ give

$$\mathbb{E}[R_S(W_t) - R_S(W_S^*)] \leq \frac{\beta}{2}\exp\left(\frac{-4t}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2]. \tag{99}$$

The definition of ϵ_{path} (Definition 6), the inequalities (98) and (99) and the constant learning rate ($\eta_t = 2/(\beta+\gamma)$) give

$$\begin{aligned}
\epsilon_{\text{path}} \triangleq \sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] &\leq 4\beta\sum_{t=1}^{T}\eta_t\mathbb{E}[R_S(W_t) - R_S(W_S^*)] \\
&\leq 4\beta\sum_{t=1}^{T}\frac{2}{\beta+\gamma}\frac{\beta}{2}\exp\left(\frac{-4t}{\frac{\beta}{\gamma}+1}\right)\mathbb{E}[\|W_1 - W_S^*\|_2^2] \\
&\leq \frac{4\beta^2}{\beta+\gamma}\mathbb{E}[\|W_1 - W_S^*\|_2^2]\sum_{t=1}^{T}\exp\left(\frac{-4t}{\frac{\beta}{\gamma}+1}\right) \\
&= \frac{4\beta^2}{\beta+\gamma}\mathbb{E}[\|W_1 - W_S^*\|_2^2]\exp\left(\frac{-4}{\frac{\beta}{\gamma}+1}\right)\frac{1-\exp\left(\frac{-4T}{\frac{\beta}{\gamma}+1}\right)}{1-\exp\left(\frac{-4}{\frac{\beta}{\gamma}+1}\right)} \\
&\leq \frac{4\beta^2(e^{\frac{4\gamma}{\beta+\gamma}}-1)^{-1}}{\beta+\gamma}\mathbb{E}[\|W_1 - W_S^*\|_2^2].
\end{aligned} \tag{100}$$

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

E.1 Proof of Theorem 16 and Theorem 17

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

$$W_{t+1} = W_t - \frac{\eta_t}{n}\sum_{j=1}^{n}\nabla f(W_t, z_j), \tag{101}$$

$$W_{t+1}^{(i)} = W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \nabla f(W_t^{(i)}, z_j) - \frac{\eta_t}{n} \nabla f(W_t^{(i)}, z_i'). \tag{102}$$

Then similarly to the inequality (86) we get

$$\begin{aligned}
&\|W_{t+1} - W_{t+1}^{(i)}\|_2 \\
&\leq \left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \right) \right\|_2 + \frac{\eta_t}{n} \|\nabla f(W_t, z_i) - \nabla f(W_t^{(i)}, z_i')\|_2 \\
&\leq \sqrt{\left\| W_t - W_t^{(i)} - \frac{\eta_t}{n} \sum_{j=1, j\neq i}^{n} \left(\nabla f(W_t, z_j) - \nabla f(W_t^{(i)}, z_j) \right) \right\|_2^2} \\
&\quad + \frac{\eta_t}{n} \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right) \\
&\leq \left(1 - \frac{\eta_t \gamma}{2} \right)^{\frac{1}{2}} \|W_t - W_t^{(i)}\|_2 + \frac{\eta_t}{n} \left(\|\nabla f(W_t, z_i)\|_2 + \|\nabla f(W_t^{(i)}, z_i')\|_2 \right)
\end{aligned} \tag{103}$$

and we apply Lemma 25 to derive (103). Then by solving the recursion we find

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

The last inequality provides the stability bound

$$\begin{aligned}
&\|W_{T+1} - W_{T+1}^{(i)}\|_2^2 \\
&\leq \frac{2}{n^2} \sum_{t=1}^{T} \eta_t \left(\|\nabla f(W_t, z_i)\|_2^2 + \|\nabla f(W_t^{(i)}, z_i')\|_2^2 \right) \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2} \right).
\end{aligned} \tag{104}$$

Inequality (104) gives that for any $i \in \{1, \ldots, n\}$

$$\begin{aligned}
\mathbb{E}[\|W_{T+1} - W_{T+1}^{(i)}\|_2^2] &\leq \frac{2}{n^2} \sum_{t=1}^{T} \eta_t \left(\mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] + \mathbb{E}[\|\nabla f(W_t^{(i)}, z_i')\|_2^2] \right) \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2} \right) \\
&= \frac{4}{n^2} \sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla f(W_t, z_i)\|_2^2] \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2} \right) \\
&= \frac{4\epsilon_{\text{path}}}{n^2} \sum_{t=1}^{T} \eta_t \prod_{j=t+1}^{T} \left(1 - \frac{\eta_j \gamma}{2} \right).
\end{aligned} \tag{105}$$

Recall that $W_{T+1} \equiv A(S)$ and $W_{T+1}^{(i)} \equiv A(S^{(i)})$. Theorem 3 and the inequality (105) give

$$|\epsilon_{\text{gen}}| \leq 2\sqrt{2\beta\epsilon_{\text{opt}} \mathbb{E}\left[\|A(S^{(i)}) - A(S)\|_2^2\right]} + 2\beta \mathbb{E}[\|A(S^{(i)}) - A(S)\|_2^2]$$

$$\leq 2\sqrt{2\beta\epsilon_{\text{opt}}\frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)}+2\beta\frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)$$
$$=\frac{4\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{2\beta\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right). \tag{106}$$

Under the choice of $\eta_t=\frac{2}{\beta+\gamma}<\frac{2}{\beta}$, the inequality (105), Lemmata 20 and 26 and give

$$\begin{aligned}\mathbb{E}[\|A(S)-A(S^{(i)})\|_2^2] &\leq \frac{4\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)\\ &=\frac{4\epsilon_{\text{path}}}{n^2}2\frac{1-\left(1-\frac{C\gamma}{2}\right)^T}{\gamma}\\ &\leq\frac{8\epsilon_{\text{path}}}{\gamma n^2}\\ &\leq\frac{32}{\gamma n^2}\frac{\beta^2(e^{\frac{\gamma}{\beta+\gamma}}-1)^{-1}}{\beta+\gamma},\end{aligned} \tag{107}$$

and similarly for the generalization error though the inequality (106) we find

$$\begin{aligned}|\epsilon_{\text{gen}}| &\leq \frac{4\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{2\beta\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)}+8\beta\frac{\epsilon_{\text{path}}}{n^2}\sum_{t=1}^{T}\eta_t\prod_{j=t+1}^{T}\left(1-\frac{\eta_j\gamma}{2}\right)\\ &=\frac{8\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{\beta\frac{1-\left(1-\frac{\gamma}{\beta+\gamma}\right)^T}{\gamma}}+16\beta\frac{\epsilon_{\text{path}}}{n^2}\frac{1-\left(1-\frac{\gamma}{\beta+\gamma}\right)^T}{\gamma}\\ &\leq\frac{8\sqrt{\epsilon_{\text{opt}}\epsilon_{\text{path}}}}{n}\sqrt{\frac{\beta}{\gamma}}+16\frac{\epsilon_{\text{path}}}{n^2}\frac{\beta}{\gamma}\\ &\leq\left(\frac{8\sqrt{2}\exp\left(\frac{-2T}{\frac{\beta}{\gamma}+1}\right)}{n}\sqrt{\frac{\beta^2}{\gamma(\beta+\gamma)(e^{\frac{4\gamma}{\beta+\gamma}}-1)}}+\frac{64}{n^2}\frac{\beta^2}{\gamma(\beta+\gamma)(e^{\frac{4\gamma}{\beta+\gamma}}-1)}\right)\beta\mathbb{E}[\|W_1-W_S^*\|_2^2]\\ &\leq\left(\frac{8\sqrt{2}\exp\left(\frac{-2T}{\frac{\beta}{\gamma}+1}\right)}{n}\sqrt{\frac{\beta^2}{2\gamma^2(e^{\frac{4\gamma}{\beta+\gamma}}-1)}}+\frac{64}{n^2}\frac{\beta^2}{2\gamma^2(e^{\frac{4\gamma}{\beta+\gamma}}-1)}\right)\beta\mathbb{E}[\|W_1-W_S^*\|_2^2]\\ &=\left(\frac{8\exp\left(\frac{-2\gamma}{\beta+\gamma}T\right)}{n}\sqrt{\frac{\beta^2}{\gamma^2(e^{\frac{4\gamma}{\beta+\gamma}}-1)}}+\frac{32}{n^2}\frac{\beta^2}{\gamma^2(e^{\frac{4\gamma}{\beta+\gamma}}-1)}\right)\beta\mathbb{E}[\|W_1-W_S^*\|_2^2].\end{aligned}$$

The last inequality provides the bound of the theorem and completes the proof. □