Stability and Sharper Risk Bounds with Convergence Rate $O(1/n^2)$

Anonymous Author(s)
Affiliation
Address
email

Abstract

The sharpest known high probability excess risk bounds are up to $O(1/n)$ for empirical risk minimization and projected gradient descent via algorithmic stability [Klochkov and Zhivotovskiy, 2021]. In this paper, we show that high probability excess risk bounds of order up to $O(1/n^2)$ are possible. We discuss how high probability excess risk bounds reach $O(1/n^2)$ under strongly convexity, smoothness and Lipschitz continuity assumptions for empirical risk minimization, projected gradient descent and stochastic gradient descent. Besides, to the best of our knowledge, our high probability results on the generalization gap measured by gradients for nonconvex problems are also the sharpest.

1 Introduction

Algorithmic stability is a fundamental concept in learning theory [Bousquet and Elisseeff, 2002], which can be traced back to the foundational works of Vapnik and Chervonenkis [1974] and has a deep connection with learnability [Rakhlin et al., 2005, Shalev-Shwartz et al., 2010, Shalev-Shwartz and Ben-David, 2014]. It is not difficult for only providing in-expectation error bounds via stability arguments. However, high probability bounds are beneficial to understand the robustness of optimization algorithms [Bousquet et al., 2020, Klochkov and Zhivotovskiy, 2021] and are much more challenging [Feldman and Vondrak, 2019, Bousquet et al., 2020, Lv et al., 2021]. In this paper, our goal is to improve the high probability risk bounds via algorithmic stability.

Let us start with some standard notations. We have a set of independent and identically distributed observations $S = \{z_1, \ldots, z_n\}$ sampled from a probability measure ρ defined on a sample space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Based on the training set S, our goal is to build a model $h : \mathcal{X} \mapsto \mathcal{Y}$ for prediction, where the model is determined by parameter $\mathbf{w}$ from parameter space $\mathcal{W} \subset \mathbb{R}^d$. The performance of a model $\mathbf{w}$ on an example z can be quantified by a loss function $f(\mathbf{w}; z)$, where $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$. Then the population risk and the empirical risk of $\mathbf{w} \in \mathcal{W}$, respectively as

$$F(\mathbf{w}) := \mathbb{E}_z \left[f(\mathbf{w}; z) \right], \quad F_S(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}; z_i),$$

where $\mathbb{E}_z$ denotes the expectation w.r.t. z.

Let $\mathbf{w}^* \in \arg\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w})$ be the model with the minimal population risk in $\mathcal{W}$ and $\mathbf{w}^*(S) \in \arg\min_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w})$ be the model with the minimal empirical risk w.r.t. dataset S. Let $A(S)$ be the output of a (possibly randomized) algorithm A on the dataset S. Let $\|\cdot\|_2$ denote the Euclidean norm and $\nabla g(\mathbf{w})$ denote a subgradient of g at $\mathbf{w}$.

Traditional generalization analysis aims to bound the generalization error $F(A(S)) - F_S(A(S))$ w.r.t the algorithm A and the dataset S. Based on the technique developed by Feldman and Vondrak [2018,

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

2019], Bousquet et al. [2020] provide the sharpest high probability bounds of $O\left(L/\sqrt{n}\right)$, where the loss function $f(\cdot,\cdot)$ is bounded by M. No matter how stable the algorithm is, the high probability generalization bound will not be smaller than $O(L/\sqrt{n})$. This is sampling error term scaling as $O\left(1/\sqrt{n}\right)$ that controls the generalization error [Klochkov and Zhivotovskiy, 2021].

A frequently used alternative to generalization bounds, that can avoid the sampling error, are the excess risk bounds. The excess risk of algorithm A w.r.t. the dataset S is $F(A(S)) - F(\mathbf{w}^*)$, which is more essential because it considers both generalization error and optimization error. Recently, Klochkov and Zhivotovskiy [2021] provided the best high probability excess risk bounds of order up to $O(\log n/n)$ for empirical risk minimization (ERM) and projected gradient descent (PGD) algorithms via algorithmic stability.

On the other hand, Zhang et al. [2017], Li and Liu [2021], Xu and Zeevi [2024] derived high probability excess risk bounds with $O\left(1/n^2\right)$ for ERM and stochastic gradient descent (SGD) via uniform convergence when the sample number satisfies $n = \Omega(d)$, which implied that the rate $O\left(1/n^2\right)$ is possible. However, the results obtained by the uniform convergence technique are related to the dimension d, which is unacceptable in high-dimensional learning problems. Since stability analysis can yield dimension-free bounds, we naturally have the following question:

Can algorithmic stability provide high probability excess risk bounds with the rate beyond $O(1/n)$?

The main results of this paper answers this question positively. We provides the first high probability bounds that are dimension-free with the rate $O(1/n^2)$ for ERM, PGD and SGD. Our framework can also be used to solve other stable algorithms.

To this end, we develop the generalization gap measured by gradients. Our bounds under *nonconvex setting* are tighter than existing works based on both algorithmic stability [Fan and Lei, 2024] and uniform convergence [Xu and Zeevi, 2024]. This is why we can achieve dimension-free excess risk bounds of order $O(1/n^2)$. In fact, in nonconvex problems, optimization algorithms can only find a local minimizer and we can only obtain optimization error bounds for $\|\nabla F_S(A(S))\|_2$ [Ghadimi and Lan, 2013]. Therefore, it is important to study the generalization behavior of $A(S)$ measured by gradients. Under Polyak-Lojasiewicz condition, we also obtain sharper results for both generalization bounds of gradients and excess risk bounds. Our route to excess risk bounds can also be applied to various stable algorithms and complex learning scenarios. In this paper, we take ERM, PGD, and SGD as examples to explore the stability of stochastic convex optimization algorithms with strongly convex losses. We provide tighter high probability dimension-free excess risk bounds of up to $O(1/n^2)$ comapring with existing works based on both algorithmic stability [Klochkov and Zhivotovskiy, 2021, Fan and Lei, 2024] and uniform convergence [Zhang et al., 2017, Li and Liu, 2021, Xu and Zeevi, 2024].

Besides, to obtain tighter bounds, we obtain a tighter p-moment bound for sums of vector-valued functions by introducing the optimal Marcinkiewicz-Zygmund's inequality for random variables taking values in a Hilbert space in the proof, which has more potential applications in vector-valued functional data.

This paper is organized as follows. The related work is reviewed in Section 2. In Section 3, we present our main results for stability and generalization. We give applications to ERM, PGD and SGD in Section 4. The conclusion is given in Section 5. All the proofs and additional lemmata are deferred to the Appendix.

2 Related Work

Algorithmic stability. Algorithmic stability is a classical approach in generalization analysis, which can be traced back to the foundational works of [Vapnik and Chervonenkis, 1974]. It gave the generalization bound by analyzing the sensitivity of a particular learning algorithm when changing one data point in the dataset. Modern framework of stability analysis was established by Bousquet and Elisseeff [2002], where they presented an important concept called uniform stability. Since then, a lot of works based on uniform stability have emerged. On one hand, generalization bounds with algorithmic stability have been significantly improved by Feldman and Vondrak [2018, 2019], Bousquet et al. [2020], Klochkov and Zhivotovskiy [2021]. On the other hand, different algorithmic stability measures such as uniform argument stability [Liu et al., 2017, Bassily et al., 2020], on average stability [Shalev-Shwartz et al., 2010, Kuzborskij and Lampert, 2018], hypothesis stability

[Bousquet and Elisseeff, 2002, Charles and Papailiopoulos, 2018], hypothesis set stability [Foster et al., 2019], pointwise uniform stability [Fan and Lei, 2024], PAC-Bayesian stability [Li et al., 2020], locally elastic stability [Deng et al., 2021], collective stability [London et al., 2016] and uniform stability in gradients [Lei, 2023, Fan and Lei, 2024] have been developed. Most of them provided the connection on stability and generalization in expectation. Bousquet and Elisseeff [2002], Elisseeff et al. [2005], Feldman and Vondrak [2018, 2019], Bousquet et al. [2020], Klochkov and Zhivotovskiy [2021], Fan and Lei [2024] considered high probability bounds. However, only Fan and Lei [2024] developed vector-valued bounds (eg: generalization bounds of gradients), which can be the order at most $O\left(M/\sqrt{n}\right)$ and remain improvement.

Uniform convergence. Uniform convergence is another popular approach in statistical learning theory to study generalization bounds [Fisher, 1922, Vapnik, 1999, Van der Vaart, 2000]. The main idea is to bound the generalization gap by its supremum over the whole (or a subset) of the hypothesis space via some space complexity measures, such as VC dimension, covering number and Rademacher complexity. For finite-dimensional problem, Kleywegt et al. [2002] provided that the generalization error is $O\left(\sqrt{d/n}\right)$ depended on the sample size n and the dimension of parameters d in high probability. In nonconvex settings, Mei et al. [2018] showed that the empirical of generalization error is $O(\sqrt{d/n})$. Xu and Zeevi [2024] developed a novel "uniform localized convergence" framework using generic chaining for the minimization problem and provided the localized generalization bounds in gradients $O\left(\max\left\{\|\mathbf{w}-\mathbf{w}^*\|_2, \frac{1}{n}\right\}\left(\sqrt{\frac{d}{n}}+\frac{d}{n}\right)\right)$, which is the optimal result when we only consider the order of n. However, uniform convergence results are related to the dimension d, which is unacceptable in high-dimensional learning problems.

3 Stability and Generalization

To derive sharper generalization bounds of gradients, we need to develop a novel concentration inequality which provide p-moment bound for sums of vector-valued functions. For a real-valued random variable Y, the L_p-norm of Y is defined by $\|Y\|_p := (\mathbb{E}[|Y|^p])^{\frac{1}{p}}$. Similarly, let $\|\cdot\|$ denote the norm in a Hilbert space $\mathcal{H}$. Then for a random variable X taking values in a Hilbert space, the L_p-norm of X is defined by $\|\|\mathbf{X}\|\|_p := (\mathbb{E}[\|\mathbf{X}\|^p])^{\frac{1}{p}}$.

3.1 A Moment Bound for Sums of Vector-valued Functions

Here we present our sharper moment bound for sums of vector-valued functions of n independent variables.

Theorem 1. *Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be a vector of independent random variables each taking values in $\mathcal{Z}$, and let $\mathbf{g}_1, \ldots, \mathbf{g}_n$ be some functions: $\mathbf{g}_i : \mathcal{Z}^n \mapsto \mathcal{H}$ such that the following holds for any $i \in [n]$:*

- $\|\mathbb{E}[\mathbf{g}_i(\mathbf{Z})|Z_i]\| \leq M$ *a.s.,*
- $\mathbb{E}\left[\mathbf{g}_i(\mathbf{Z})|Z_{[n]\setminus\{i\}}\right] = 0$ *a.s.,*
- $\mathbf{g}_i$ *satisfies the bounded difference property with β, namely, for any $i = 1, \ldots, n$, the following inequality holds*

$$\sup_{z_1,\ldots,z_n,z_j'} \|\mathbf{g}_i(z_1,\ldots,z_{j-1},z_j,z_{j+1},\ldots,z_n) - \mathbf{g}_i(z_1,\ldots,z_{j-1},z_j',z_{j+1},\ldots,z_n)\| \leq \beta. \quad (1)$$

Then, for any $p \geq 2$, we have

$$\left\|\left\|\sum_{i=1}^n \mathbf{g}_i\right\|\right\|_p \leq 2(\sqrt{2p}+1)\sqrt{n}M + 4\times 2^{\frac{1}{2p}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)n\beta\left\lceil\log_2 n\right\rceil.$$

Remark 1. The proof is motivated by Bousquet et al. [2020]. Under the same assumptions, Fan and Lei [2024] also established the following inequality[1]

$$\left\|\left\|\sum_{i=1}^{n}\mathbf{g}_i\right\|\right\|_p \leq 2(\sqrt{2}+1)\sqrt{np}M + 4(\sqrt{2}+1)np\beta\left\lceil\log_2 n\right\rceil. \tag{2}$$

It is easy to verify that our result is tighter than result provided by Fan and Lei [2024] for both the first and second term. Comparing Theorem 1 with (2), the larger p is, the tighter our result is relative to (2). In the worst case, when $p = 2$, the constant of our first term is 0.879 times tighter than (2), and the constant of our second term is 0.634 times tighter than (2). This is because we derive the optimal Marcinkiewicz-Zygmund's inequality for random variables taking values in a Hilbert space in the proof.

The improvement of this concentration inequality is meaningful. On one hand, we derive the optimal Marcinkiewicz-Zygmund's inequality for random variables taking values in a Hilbert space. On the other hand, in Section 3.2, we will carefully construct vector-valued functions which satisfies all the assumptions in Theorem 1 and ensures $M = 0$ at the same time. Under this condition, we can eliminate the first term. When we use Theorem 1 instead of (2) in the whole proofs, at least 0.634 times tighter bound can be obtained strictly.

3.2 Sharper Generalization Bounds in Gradients

Let $S = \{z_1, \ldots, z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$ and $S' = \{z_1', \ldots, z_n'\}$ be its independent copy. For any $i \in [n]$, define $S^{(i)} = \{z_i, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$ be a dataset replacing the i-th sample in S with another i.i.d. sample z_i'. We introduce some basic definitions here and we want to emphasize that our main Theorem 2 and Theorem 3 do not need smoothness assumption and PL condition.

Definition 1. *Let $g : \mathcal{W} \mapsto \mathbb{R}$. Let $\gamma, \mu < 0$.*

- *We say g is γ-smooth if*

$$\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \leq \gamma\|\mathbf{w} - \mathbf{w}'\|_2, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

- *Let $g* = \min_{\mathbf{w}\in\mathcal{W}} g(\mathbf{w})$. We say g satisfies the Polyak-Lojasiewicz (PL) condition with parameter $\mu > 0$ on $\mathcal{W}$ if*

$$g(\mathbf{w}) - g* \leq \frac{1}{2\mu}\|\nabla g(\mathbf{w})\|_2^2, \quad \forall \mathbf{w} \in \mathcal{W}.$$

Then we define uniform stability in gradients.

Definition 2 (Uniform Stability in Gradients)**.** *Let A be a randomized algorithm. We say A is β-uniformly-stable* in gradients *if for all neighboring datasets $S, S^{(i)}$, we have*

$$\sup_z \left\|\nabla f(A(S); z) - \nabla f(A(S^{(i)}); z)\right\|_2 \leq \beta. \tag{3}$$

Remark 2. Gradient-based stability was firstly introduced by Lei [2023], Fan and Lei [2024] to describe the generalization performance for nonconvex problems. In nonconvex problems, we can only find a local minimizer by optimization algorithms which may be far away from the global minimizer. Thus the convergence does not make much sense in function values. Instead, the convergence of $\|\nabla F_S(A(S))\|_2$ was often studied in the optimization community [Ghadimi and Lan, 2013]. Since the population risk of gradients $\|\nabla F(A(S))\|_2$ can be decomposed as the convergence of $\|\nabla F_S(A(S))\|_2$ and the generalization gap $\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2$, the generalization analysis of $\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2$ is important, which can be achieved by uniform stability in gradients.

Theorem 2 (Generalization via Stability in Gradients)**.** *Assume for any S and any z, $\|\nabla f(A(S); z)\|_2 \leq M$. If A is β-uniformly-stable in gradients, then for any $\delta \in (0, 1)$, the*

[1]They assume $n = 2^k, k \in \mathbb{N}$. Here we give the version of their result with general n.

following inequality holds with probability at least $1-\delta$

$$\begin{aligned}&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2\\ \leq& 2\beta + \frac{4M\left(1+e\sqrt{2\log(e/\delta)}\right)}{\sqrt{n}} + 8\times 2^{\frac{1}{4}}(\sqrt{2}+1)\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log(e/\delta).\end{aligned}$$

Remark 3. Theorem 2 is a direct application via Theorem 1 where we denote $\mathbf{g}_i(S) = \mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}),Z)\right] - \nabla f(A(S^{(i)}),z_i)\right]$ and find that $\mathbf{g}_i(S)$ satisfies all the assumptions in Theorem 1. As a comparison, Fan and Lei [2024] also developed high probability bounds under same assumptions, but our bounds are sharper since our moment inequality for sums of vector-valued functions are tighter as we have discussed in Remark 1. Next, we derive sharper generalization bound of gradients under same assumptions.

Theorem 3 (Sharper Generalization via Stability in Gradients)**.** *Assume for any S and any z, $\|\nabla f(A(S);z)\|_2 \leq M$. If A is β-uniformly-stable in gradients, then for any $\delta\in(0,1)$, the following inequality holds with probability at least* $1-\delta$

$$\begin{aligned}&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2\\ \leq& \sqrt{\frac{4\mathbb{E}_Z\left[\|\nabla f(A(S);Z)\|_2^2\right]\log\frac{6}{\delta}}{n}} + \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n}\\ &+ 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log(3e/\delta) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.\end{aligned}$$

Remark 4. Note that the factor in Theorem 2 before $1/\sqrt{n}$ is $O\left(M\sqrt{\log(e/\delta)}\right)$, which depends on the bound of $\|\nabla f(\cdot,\cdot)\|_2$. However, the factor in Theorem 3 before $1/\sqrt{n}$ is $O\left(\sqrt{\mathbb{E}_Z\left[\|\nabla f(A(S);Z)\|_2^2\right]\log 1/\delta} + \beta\log(1/\delta)\right)$, not involving the possibly large term M. As is known, optimization algorithms often provide parameters approaching the optimal solution, which make the term $\mathbb{E}_Z[\|\nabla f(A(S);Z)\|_2^2]$ much more smaller than M. We will give further reasonable results under more assumptions such as smoothness in Lemma 1 and Lemma 2.

On the other hand, best high probability bounds based on uniform convergence [Xu and Zeevi, 2024] is

$$\begin{aligned}&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2\\ \lesssim& \sqrt{\frac{\mathbb{E}_Z\left[\nabla\|f(\mathbf{w}^*;Z)\|_2^2\right]\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} + \max\left\{\|\mathbf{w}-\mathbf{w}^*\|_2, \frac{1}{n}\right\}\left(\sqrt{\frac{d}{n}}+\frac{d}{n}\right),\end{aligned} \tag{4}$$

which is the optimal result when we only consider the order of n. However, uniform convergence results are related to the dimension d, which is unacceptable in high-dimensional learning problems.

Note that (4) requires an additional smoothness-type assumption. As a comparison, when f is γ-smoothness, our result in Theorem 3 can be easily derived as

$$\begin{aligned}&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2\\ \lesssim& \beta\log n\log(1/\delta) + \frac{\log(1/\delta)}{n} + \sqrt{\frac{\mathbb{E}_Z\left[\nabla\|f(\mathbf{w}^*;Z)\|_2^2\right]\log(1/\delta)}{n}} + \|A(S)-\mathbf{w}^*\|\sqrt{\frac{\log(1/\delta)}{n}}.\end{aligned}$$

This result implies that when the uniformly stable in gradients parameter β is smaller than $1/\sqrt{n}$, our bound is tighter than (4) and is dimension independent. Note that Theorem 3 holds in nonconvex problems, to the best of our knowledge, this is the sharpest upper bound in both uniform convergence and algorithmic stability analysis.

Here we give the proof sketch of Theorem 3, which is motivated by the analysis in Klochkov and Zhivotovskiy [2021]. The key idea is to build vector functions $\mathbf{q}_i(S) = \mathbf{h}_i(S) - \mathbb{E}_{S\{z_i\}}[\mathbf{h}_i(S)]$ where we define $\mathbf{h}_i(S) = \mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}),Z)\right] - \nabla f(A(S^{(i)}),z_i)\right]$. These functions satisfy all the assumptions in Theorem 1 and ensure the factor M in Theorem 1 to 0. Then the term $O(1/\sqrt{n})$ can be eliminated.

Lemma 1. *Let assumptions in Theorem 3 hold. Suppose the function f is γ-smooth and the population risk F satisfies the PL condition with parameter μ. Then for any $\delta \in (0,1)$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq &\|\nabla F_S(A(S))\|_2 + 4\sqrt{\frac{\mathbb{E}_Z\left[\|\nabla f(\mathbf{w}^*; Z)\|_2^2\right] \log \frac{6}{\delta}}{n}} + 2\sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2 \log(3/\delta)\right) \log \frac{6}{\delta}}{n}} \\
&+ \frac{2M \log \frac{6}{\delta}}{n} + 32 \times 2^{\frac{3}{4}} \sqrt{e}\beta \left\lceil \log_2 n \right\rceil \log\left(3e/\delta\right) + 64\sqrt{e}\beta \left\lceil \log_2 n \right\rceil \sqrt{\log 3e/\delta}.
\end{aligned}$$

Remark 5. The following inequality can be easily derived using triangle inequality and Cauchy-Bunyakovsky-Schwarz inequality:

$$F(A(S)) - F(\mathbf{w}^*) \lesssim \|\nabla F_S(A(S))\|_2 + \frac{F(\mathbf{w}^*) \log\left(1/\delta\right)}{n} + \frac{\log^2(1/\delta)}{n^2} + \beta^2 \log^2 n \log^2(1/\delta). \quad (5)$$

Above inequality implies that excess risk can be bound by the optimization gradient error $\|\nabla F_S(A(S))\|_2$ and uniform stability in gradients β. Note that the assumption $F(\mathbf{w}^*) = O(1/n)$ is common and can be found in Srebro et al. [2010], Lei and Ying [2020], Liu et al. [2018], Zhang et al. [2017], Zhang and Zhou [2019]. This is natural since $F(\mathbf{w}^*)$ is the minimal population risk. On the other hand, we can derive that under μ-strongly convex and γ-smooth assumptions for the objective function f, uniform stability in gradients can be bounded of order $O(1/n)$ for ERM and PGD. Thus high probability excess risk can be bounded of order up to $O\left(1/n^2\right)$ under these common assumptions via algorithmic stability. Comparing with current best related work [Klochkov and Zhivotovskiy, 2021], they are insensitive to the stability parameter being smaller than $O(1/n)$ and their best rates can only up to $O(1/n)$. Although we involve extra smoothness and PL condition assumptions, these assumptions are also common in optimization community and our work can fully utilize these assumptions.

Besides, we discuss uniform stability in gradients for common algorithms such as ERM, PGD, and SGD in Section 4. Our results can be easily extended to other stable algorithms. Due to smoothness's property to link the uniform stability in gradients with uniform argument stability, many works [Bassily et al., 2020, Feldman and Vondrak, 2019, Hardt et al., 2016] exploring uniform argument stability can also use our framework.

Finally, the population risk of gradients $\|\nabla F(A(S))\|_2$ can be gracefully bounded by the empirical risk of gradients $\|\nabla F_S(A(S))\|_2$ under strong growth condition (SGC), that connects the rates at which the stochastic gradients shrink relative to the full gradient Vaswani et al. [2019].

Definition 3 (Strong Growth Condition)**.** *We say SGC holds if*

$$\mathbb{E}_Z\left[\|\nabla f(\mathbf{w}; Z)\|_2^2\right] \leq \rho \|\nabla F(\mathbf{w})\|_2^2.$$

Remark 6. There has been some related work that takes SGC into assumption Solodov [1998], Vaswani et al. [2019], Lei [2023]. Vaswani et al. [2019] has proved that the squared-hinge loss with linearly separable data and finite support features satisfies the SGC. Note that we only suppose this condition holds in Lemma 2.

Lemma 2 (SGC case)**.** *Let assumptions in Theorem 3 hold and suppose SGC holds. Then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\|\nabla F(A(S))\| \lesssim (1+\eta)\|\nabla F_S(A(S))\| + \frac{1+\eta}{\eta}\left(\frac{M}{n}\log\frac{6}{\delta} + \beta \log n \log\frac{1}{\delta}\right).$$

Remark 7. Lemma 2 build a connection between the population gradient error and the empirical gradient error under Lipschitz, nonconvex, nonsmooth and SGC case and elucidate that the population gradient error can be bounded of up to $O(1/n)$ under nonconvex problems.

4 Application

In this section, we analysis stochastic convex optimization with strongly convex losses. The most common setting is where at each round, the learner gets information on f through a stochastic gradient oracle [Rakhlin et al., 2012]. To derive uniform stability in gradients for algorithms, we firstly introduce the strongly convex assumption.

Definition 4. *We say g is μ-strongly convex if*

$$g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}') \rangle + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

4.1 Empirical Risk Minimizer

Empirical risk minimizer is one of the classical approaches for solving stochastic optimization (also referred to as sample average approximation (SAA)) in machine learning community. The following lemma shows the uniform stability in gradient for ERM under μ-strongly convexity and γ-smoothness assumptions.

Lemma 3 (Stability of ERM)**.** *Suppose the objective function f is μ-strongly-convex and γ-smooth. For any $\mathbf{w} \in \mathcal{W}$ and any z, suppose that $\|\nabla f(\mathbf{w}; z) \leq M\|$. Let $\hat{\mathbf{w}}^*(S^{(i)})$ be the ERM of $F_{S^{(i)}}(\mathbf{w})$ that denotes the empirical risk on the samples $S^{(i)} = \{z_1, ..., z_i', ..., z_n\}$ and $\hat{\mathbf{w}}^*(S)$ be the ERM of $F_S(\mathbf{w})$ on the samples $S = \{z_1, ..., z_i, ..., z_n\}$. For any $S^{(i)}$ and S, there holds the following uniform stability bound of ERM:*

$$\forall z \in \mathcal{Z}, \quad \left\|\nabla f(\hat{\mathbf{w}}^*(S^{(i)}); z) - \nabla f(\hat{\mathbf{w}}^*(S); z)\right\|_2 \leq \frac{4M\gamma}{n\mu}.$$

Then, we present the application of our main sharper Theorem 3. In the strongly convex and smooth case, we provide a up to $O\left(1/n^2\right)$ high probability excess risk guarantee valid for any algorithms depending on the optimal population error $F(\mathbf{w}^*)$.

Theorem 4. *Let assumptions in Theorem 3 and Lemma 3 hold. Suppose the function f is nonnegative. Then for any $\delta \in (0, 1)$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1 - \delta$, we have*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \lesssim \frac{F(\mathbf{w}^*) \log(1/\delta)}{n} + \frac{\log^2 n \log^2(1/\delta)}{n^2}.$$

Furthermore, assume $F(\mathbf{w}^*) = O(\frac{1}{n})$, we have

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \lesssim \frac{\log^2 n \log^2(1/\delta)}{n^2}.$$

Remark 8. Theorem 4 shows that when the objective function f is μ-strongly convex, γ-smooth and nonnegative, high probability risk bounds can even up to $O\left(1/n^2\right)$ for ERM. The most related work to ours is Zhang et al. [2017]. They also obtain the $O\left(1/n^2\right)$-type bounds for ERM by uniform convergence of gradients approach. However, they need the sample number $n = \Omega(\gamma d/\mu)$, which is related to the dimension d. Our risk bounds are dimension independent and only require the sample number $n = \Omega(\gamma^2/\mu^2)$. Comparing with Klochkov and Zhivotovskiy [2021], we add two assumptions, smoothness and $F(\mathbf{w}^*) = O(1/n)$, but our bounds also tighter, from $O(1/n)$ to $O\left(1/n^2\right)$.

4.2 Projected Gradient Descent

Note that when the objective function f is strongly convex and smooth, the optimization error can be ignored. However, the generalization analysis framework proposed by Klochkov and Zhivotovskiy [2021] does not use smoothness assumption, which only derive high probability excess risk bound of order $O(1/n)$ after $T = O(\log n)$ steps under strongly convex and smooth assumptions. In this subsection, we provide sharper risk bound under the same iteration steps, which is because our generalization analysis also fully utilized the smooth assumptions. Here we give the definition of PGD.

Definition 5 (Projected Gradient Descent)**.** *Let $\mathbf{w}_1 = o \in \mathbb{R}^d$ be an initial point and $\{\eta_t\}_t$ be a sequence of positive step sizes. PGD updates parameters by*

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\left(\mathbf{w}_t - \eta_t \nabla F_S\left(\mathbf{w}_t\right)\right),$$

where $\nabla F_S(\mathbf{w}_t)$ denotes a subgradient of F w.r.t. $\mathbf{w}_t$ and $\Pi_{\mathcal{W}}$ is the projection operator onto $\mathcal{W}$.

Lemma 4 (Stability of Gradient Descent)**.** *Suppose the objective function f is μ-strongly-convex and γ-smooth. For any $\mathbf{w} \in \mathcal{W}$ and any z, suppose that $\|\nabla f(\mathbf{w}; z)\|_2 \leq M$. Let $\mathbf{w}_t^i$ be the output of $F_{S^{(i)}}(\mathbf{w})$ on t-th iteration on the samples $S^{(i)} = \{z_1, ..., z_i', ..., z_n\}$ in running PGD, and $\mathbf{w}_t$ be the output of $F_S(\mathbf{w})$ on t-th iteration on the samples $S = \{z_1, ..., z_i, ..., z_n\}$ in running PGD. Let the constant step size $\eta_t = 1/\gamma$. For any $S^{(i)}$ and S, there holds the following uniform stability bound of PGD:*

$$\forall z \in \mathcal{Z}, \quad \left\|\nabla f(\hat{\mathbf{w}}^*(S^{(i)}); z) - \nabla f(\hat{\mathbf{w}}^*(S); z)\right\|_2 \leq \frac{4M\gamma}{n\mu}.$$

Remark 9. The derivations of Feldman and Vondrak [2019] in Section 4.1.2 (See also Hardt et al. [2016] in Section 3.4) imply that if the objective function f is γ-smooth in addition to μ-strongly convexity and M-Lipschitz property, then PGD with the constant step size $\eta = 1/\gamma$ is $\left(\frac{2M}{n\mu}\right)$-uniformly argument stable for any number of steps, which means that PGD is $\left(\frac{2M\gamma}{n\mu}\right)$-uniformly-stable in gradients regardless of iteration steps.

Theorem 5. *Let assumptions in Theorem 3 and Lemma 3 hold. Suppose the function f is nonnegative. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by PGD with $\eta_t = 1/\gamma$. Then for any $\delta \in (0, 1)$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1 - \delta$, we have*

$$F(\mathbf{w}) - F(\mathbf{w}^*) \lesssim \left(1 - \frac{\mu}{\gamma}\right)^{2T} + \frac{F(\mathbf{w}^*) \log(1/\delta)}{n} + \frac{\log^2 n \log^2(1/\delta)}{n^2}.$$

Furthermore, assume $F(\mathbf{w}^*) = O(\frac{1}{n})$ and let $T \asymp \log n$, we have

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \lesssim \frac{\log^2 n \log^2(1/\delta)}{n^2}.$$

Remark 10. Theorem 5 shows that under the same assumptions as Klochkov and Zhivotovskiy [2021], our bound is $O\left(\frac{F(\mathbf{w}^*)\log(1/\delta)}{n} + \frac{\log^2 n \log^2(1/\delta)}{n^2}\right)$. Comparing with their bound $O\left(\frac{\log n \log(1/\delta)}{n}\right)$, we are sharper because $F(\mathbf{w}^*)$ is the minimal population risk, which is a common assumption towards sharper risk bounds Srebro et al. [2010], Lei and Ying [2020], Liu et al. [2018], Zhang et al. [2017], Zhang and Zhou [2019].

4.3 Stochastic Gradient Descent

Stochastic gradient descent optimization algorithm has been widely used in machine learning due to its simplicity in implementation, low memory requirement and low computational complexity per iteration, as well as good practical behavior. Here we give the definition of standard SGD.

Definition 6 (Stochastic Gradient Descent)**.** *Let $\mathbf{w}_1 = o \in \mathbb{R}^d$ be an initial point and $\{\eta_t\}_t$ be a sequence of positive step sizes. SGD updates parameters by*

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\left(\mathbf{w}_t - \eta_t \nabla f\left(\mathbf{w}_t; z_{i_t}\right)\right),$$

where $\nabla f(\mathbf{w}_t; z_{i_t})$ denotes a subgradient of f w.r.t. $\mathbf{w}_t$ and i_t is independently drawn from the uniform distribution over $[n] := \{1, 2, \ldots, n\}$.

Lemma 5 (Stability of SGD)**.** *Suppose the objective function f is μ-strongly-convex and γ-smooth. For any $\mathbf{w} \in \mathcal{W}$ and any z, suppose that $\|\nabla f(\mathbf{w}; z)\|_2 \leq M$. Let $\mathbf{w}_t^i$ be the output of $F_{S^{(i)}}(\mathbf{w})$ on t-th iteration on the samples $S^{(i)} = \{z_1, ..., z_i', ..., z_n\}$ in running PGD and and $\mathbf{w}_t$ be the output of $F_S(\mathbf{w})$ on t-th iteration on the samples $S = \{z_1, ..., z_i, ..., z_n\}$ in running SGD. For any $S^{(i)}$ and S, there holds the following uniform stability bound of SGD:*

$$\left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2 \leq 2\gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu}, \quad \forall z \in \mathcal{Z},$$

where $\epsilon_{opt}(\mathbf{w}_t) = F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*(S))$ and $\hat{\mathbf{w}}^*(S)$ is the ERM of $F_S(\mathbf{w})$.

Next, we introduce a necessary assumption in stochastic optimization theory.

Assumption 1. *Assume the existence of $\sigma > 0$ satisfying*

$$\mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}_t; z_{i_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2] \leq \sigma^2, \quad \forall t \in \mathbb{N}, \tag{6}$$

where $\mathbb{E}_{i_t}$ denotes the expectation w.r.t. i_t.

Remark 11. Assumption 1 is a standard assumption from the stochastic optimization theory [Nemirovski et al., 2009, Ghadimi and Lan, 2013, Ghadimi et al., 2016, Kuzborskij and Lampert, 2018, Zhou et al., 2018, Bottou et al., 2018, Lei and Tang, 2021], which essentially bounds the variance of the stochastic gradients for dataset S.

Theorem 6. *Let assumptions in Theorem 3 and Lemma 5 hold. Suppose Assumption 1 holds and the function f is nonnegative. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by SGD with $\eta_t = \eta_1 t^{-\theta}, \theta \in (0,1)$ and $\eta_1 < \frac{1}{2\gamma}$. Then for any $\delta \in (0,1)$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\left(\sum_{t=1}^T \eta_t\right)^{-1} \sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 \\
&= \begin{cases} O\left(\frac{\log^2 n \log^3(1/\delta)}{T^{-\theta}}\right) + O\left(\frac{\log^2 n \log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right), & \text{if } \theta < 1/2 \\ O\left(\frac{\log^2 n \log^3(1/\delta)}{T^{-\frac{1}{2}}}\right) + O\left(\frac{\log^2 n \log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right), & \text{if } \theta = 1/2 \\ O\left(\frac{\log^2 n \log^3(1/\delta)}{T^{\theta-1}}\right) + O\left(\frac{\log^2 n \log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right), & \text{if } \theta > 1/2. \end{cases}
\end{aligned}$$

Remark 12. When $\theta < 1/2$, we take $T \asymp n^{2/\theta}$. When $\theta = 1/2$, we take $T \asymp n^4$ and when $\theta > 1/2$, we set $T \asymp n^{2/(1-\theta)}$. Then according to Theorem 6, the population risk of gradient is bounded by $O\left(\frac{\log^2 n \log^3(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right)$. To the best of our knowledge, this is the first high probability population gradient bound $\|\nabla F_(\mathbf{w}_t)\|_2$ for SGD via algorithmic stability.

Theorem 7. *Let Assumptions in Theorem 3 and Lemma 5 hold. Suppose Assumption 1 holds and the function f is nonnegative. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by SGD with $\eta_t = \frac{2}{\mu(t+t_0)}$ such that $t_0 \geq \max\left\{\frac{4\gamma}{\mu}, 1\right\}$. Then for any $\delta > 0$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$ and $T \asymp n^2$, with probability at least $1-\delta$, we have*

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = O\left(\frac{\log^4 n \log^5(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log(1/\delta))}{n}\right).$$

Furthermore, assume $F(\mathbf{w}^*) = O(\frac{1}{n})$, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = O\left(\frac{\log^4 n \log^5(1/\delta)}{n^2}\right).$$

Remark 13. Theorem 7 implies that high probability risk bounds for SGD optimization algorithm can be up to $O(1/n^2)$ and the rate is dimension-free in high-dimensional learning problems. We compare Theorem 7 with most related work. For algorithmic stability, high probability risk bounds in Fan and Lei [2024] is up to $O(1/n)$ when choosing optimal iterate number T for SGD optimization algorithm. To the best of knowledge, we are faster than all the existing bounds. The best high probability risk bounds of order $O(1/n^2)$ are given by Li and Liu [2021] via uniform convergence, which require the sample number $n = \Omega(\gamma d/\mu)$ depending on dimension d.

5 Conclusion

In this paper, we improve a p-moment concentration inequality for sums of vector-valued functions. By carefully constructing functions, we apply this moment concentration to derive sharper generalization bounds in gradients in nonconvex problems, which can further be used to obtain sharper high probability excess risk bounds for stable optimization algorithms. In application, we study three common algorithms: ERM, PGD, SGD. To the best of our knowledge, we provide the sharpest high probability dimension independent $O(1/n^2)$-type for these algorithms.

References

R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4381–4391, 2020.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

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

O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

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

P. J. Davis. Gamma function and related functions. *Handbook of mathematical functions*, 256, 1972.

V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Z. Deng, H. He, and W. Su. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning*, pages 2590–2600. PMLR, 2021.

A. Elisseeff, T. Evgeniou, M. Pontil, and L. P. Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

J. Fan and Y. Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.

V. Feldman and J. Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604): 309–368, 1922.

D. J. Foster, S. Greenberg, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Hypothesis set stability and generalization. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

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

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML*, pages 795–811. Springer, 2016.

A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12(2):479–502, 2002.

Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.

I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2815–2824. PMLR, 2018.

R. Latała and K. Oleszkiewicz. On the best constant in the khinchin-kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994.

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

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

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

J. Li, X. Luo, and M. Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

S. Li and Y. Liu. Improved learning rates for stochastic optimization: Two theoretical viewpoints. *arXiv preprint arXiv:2107.08686*, 2021.

M. Liu, X. Zhang, L. Zhang, R. Jin, and T. Yang. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *Advances in Neural Information Processing Systems*, 31, 2018.

T. Liu, G. Lugosi, G. Neu, and D. Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.

B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

X. Luo and D. Zhang. Khintchine inequality on normed spaces and the application to banach-mazur distance. *arXiv preprint arXiv:2005.03728*, 2020.

s. Lv, J. Wang, J. Liu, and Y. Liu. Improved learning rates of a functional lasso-type svm with sparse multi-kernel representation. *Advances in Neural Information Processing Systems*, 34:21467–21479, 2021.

S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

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

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04): 397–417, 2005.

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

O. Rivasplata, E. Parrado-Hernández, J. S. Shawe-Taylor, S. Sun, and C. Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. *Advances in Neural Information Processing Systems*, 31, 2018.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11:23–35, 1998.

N. Srebro, K. Sridharan, and A. Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. 1974.

V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Y. Xu and A. Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 2024.

L. Zhang and Z.-H. Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the o(1/t) convergence rate. In *Conference on Learning Theory*, pages 3160–3179. PMLR, 2019.

L. Zhang, T. Yang, and R. Jin. Empirical risk minimization for stochastic convex optimization: O(1/n)-and o(1/n**2)-type of risk bounds. In *Conference on Learning Theory*, pages 1954–1979. PMLR, 2017.

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

A Additional definitions and lemmata

Lemma 6 (Equivalence of tails and moments for random vectors [Bassily et al., 2020]). *Let X be a random variable with*

$$\|X\|_p \leq \sqrt{p}a + pb$$

for some $a, b \geq 0$ and for any $p \geq 2$. Then for any $\delta \in (0,1)$ we have, with probability at least $1-\delta$,

$$|X| \leq e\left(a\sqrt{\log\left(\frac{e}{\delta}\right)} + b\log\frac{e}{\delta}\right).$$

Lemma 7 (Vector Bernstein's inequality [Pinelis, 1994, Smale and Zhou, 2007]). *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables taking values in a real separable Hilbert space. Assume that $\mathbb{E}[X_i] = \mu$, $\mathbb{E}[\|X_i - \mu\|^2] = \sigma^2$, and $\|X_i\| \leq M$, $\forall 1 \leq i \leq n$, then for all $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right\| \leq \sqrt{\frac{2\sigma^2\log(\frac{2}{\delta})}{n}} + \frac{M\log\frac{2}{\delta}}{n}.$$

Definition 7 (Weakly self-Bounded Function). *Assume that $a, b > 0$. A function $f : \mathcal{Z}^n \mapsto [0, +\infty)$ is said to be (a,b)-weakly self-bounded if there exist functions $f_i : \mathcal{Z}^{n-1} \mapsto [0,+\infty)$ that satisfies for all $Z^n \in \mathcal{Z}^n$,*

$$\sum_{i=1}^n (f_i(Z^n) - f(Z^n))^2 \leq af(Z^n) + b.$$

Lemma 8 ([Klochkov and Zhivotovskiy, 2021]). *Suppose that $z_1, \ldots, z_n$ are independent random variables and the function $f : \mathcal{Z}^n \mapsto [0,+\infty)$ is (a,b)-weakly self-bounded and the corresponding function f_i satisfy $f_i(Z^n) \geq f(Z^n)$ for $\forall i \in [n]$ and any $Z^n \in \mathcal{Z}^n$. Then, for any $t > 0$,*

$$Pr(\mathbb{E}f(z_1,\ldots,z_n) \geq f(z_1,\ldots,z_n) + t) \leq \exp\left(-\frac{t^2}{2a\mathbb{E}f(z_1,\ldots,z_n) + 2b}\right).$$

Definition 8. *A Rademacher random variable is a Bernoulli variable that takes values ± 1 with probability $\frac{1}{2}$ each.*

B Proofs of Section 3.1

The proof of Theorem 1 is motivated by Bousquet et al. [2020], which need the Marcinkiewicz-Zygmund's inequality for random variables taking values in a Hilbert space and the McDiarmid's inequality for vector-valued functions.

Firstly, we derive the optimal constants in the Marcinkiewicz-Zygmund's inequality for random variables taking values in a Hilbert space.

Lemma 9 (Marcinkiewicz-Zygmund's Inequality for Random Variables Taking Values in a Hilbert Space). *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random variables taking values in a Hilbert space with $\mathbb{E}[\mathbf{X}_i] = 0$ for all $i \in [n]$. Then for $p \geq 2$ we have*

$$\left\|\left\|\sum_{i=1}^n \mathbf{X}_i\right\|\right\|_p \leq 2\cdot 2^{\frac{1}{2p}}\sqrt{\frac{np}{e}}\left(\frac{1}{n}\sum_{i=1}^n \left\|\,\|\mathbf{X}_i\|\,\right\|_p^p\right)^{\frac{1}{p}}.$$

Remark 14. Comparing with Marcinkiewicz-Zygmund's inequality given by Fan and Lei [2024], we provide best constants. Next, we give the proof of Lemma 9.

The Marcinkiewicz-Zygmund's inequality can be proved by using its connection to Khintchine-Kahane's inequality. Thus, we introduce the best constants in Khintchine-Kahane's inequality for random variables taking values from a Hilbert space here.

Lemma 10 (Best constants in Khintchine-Kahane's inequality in Hilbert space [Latała and Oleszkiewicz, 1994, Luo and Zhang, 2020])**.** *For all $p \in [2, \infty)$ and for all choices of Hilbert space $\mathcal{H}$, finite sets of vectors $\mathbf{X}_i, \ldots, \mathbf{X}_n \in \mathcal{X} \in \mathcal{H}$, and independent Rademacher variables $r_1, \ldots, r_n$,*

$$\left[\mathbb{E}\left\|\sum_{i=1}^{n} r_i \mathbf{X}_i\right\|^p\right]^{\frac{1}{p}} \leq C_p \cdot \left[\sum_{i=1}^{n} \|\mathbf{X}_i\|^2\right]^{\frac{1}{2}},$$

where $C_p = 2^{\frac{1}{2}} \left\{\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right\}^{\frac{1}{p}}$.

Proof of Lemma 9. The symmetrization argument goes as follows: Let $(r_1, \ldots, r_n)$ be i.i.d. with $\mathbb{P}(r_i = 1) = \mathbb{P}(r_i = -1) = 1/2$ and besides such that $r_1, \ldots, r_n$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ are independent. Then by independence and symmetry, according to Lemma 1.2.6 of De la Pena and Giné [2012], conditioning on $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ yields

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} \mathbf{X}_i\right\|^p\right] = 2^p \mathbb{E}\left[\left\|\sum_{i=1}^{n} r_i \mathbf{X}_i\right\|^p\right] \leq 2^p \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^{n} r_i \mathbf{X}_i\right\|^p \middle| \mathbf{X}_1, \ldots, \mathbf{X}_n\right]\right]. \tag{7}$$

As for the conditional expectation in (7), notice that by independence

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} r_i \mathbf{X}_i\right\|^p \middle| \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n\right] = \mathbb{E}\left[\left\|\sum_{i=1}^{n} r_i \mathbf{X}_i\right\|^p\right] \tag{8}$$

According to Lemma 10, for v_n-almost every $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^n$, where $v_n := \mathbb{P} \circ (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{-1}$ denotes the distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, we have

$$\left[\mathbb{E}\left\|\sum_{i=1}^{n} r_i \mathbf{x}_i\right\|^p\right] \leq C \cdot \left[\sum_{i=1}^{n} \|\mathbf{x}_i\|^2\right]^{\frac{p}{2}}, \tag{9}$$

where $C = 2^{\frac{p}{2}} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$ and C is optimal. This means that for any constant C' such that

$$\left[\mathbb{E}\left\|\sum_{i=1}^{n} r_i \mathbf{x}_i\right\|^p\right] \leq C' \cdot \left[\sum_{i=1}^{n} \|\mathbf{x}_i\|^2\right]^{\frac{p}{2}}, \tag{10}$$

for all $n \in \mathbb{N}$ and for each collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, it follows that $C' \geq C$.

From (8) and (9), we can infer that

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} r_i \mathbf{X}_i\right\|^p \middle| \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n\right] \leq C \cdot \left[\sum_{i=1}^{n} \|\mathbf{X}_i\|^2\right]^{\frac{p}{2}}.$$

Taking expectations in the above inequalities and (7) yield that

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} \mathbf{X}_i\right\|^p\right] \leq C \cdot \mathbb{E}\left[\sum_{i=1}^{n} \|\mathbf{X}_i\|^2\right]^{\frac{p}{2}}. \tag{11}$$

To see optimality let the above statement hold for some constants C' in place of C. Then if we choose $\mathbf{X}_i := \mathbf{x}_i r_i, 1 \leq i \leq n$ with arbitrary reals vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, it follows that

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} r_i \mathbf{x}_i\right\|^p\right] \leq C' \cdot \mathbb{E}\left[\sum_{i=1}^{n} \|\mathbf{x}_i\|^2\right]^{\frac{p}{2}},$$

whence we can conclude from (10) that $C' \geq C$. Thus we obtain that $C' = C$.

Notice that by Holder's inequality

$$\left[\sum_{i=1}^{n}\|\mathbf{X}_i\|^2\right]^{\frac{p}{2}} \leq n^{p/2-1}\sum_{i=1}^{n}\|\mathbf{X}_i\|^p. \tag{12}$$

Plugging (12) into (11), we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n}\mathbf{X}_i\right\|^p\right] \leq C\cdot 2^p n^{p/2-1}\cdot\mathbb{E}\left[\sum_{i=1}^{n}\|\mathbf{X}_i\|^p\right],$$

where $C = 2^{\frac{p}{2}}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$ is a constant.

Next, we use the following form of Stirling's formula for the Gamma-function, which follows from (6.1.5), (6.1.15) and (6.1.38) in Davis [1972] to bound the constant C. For every $x > 0$, there exists a $\mu(x) \in (0, 1/(12x))$ such that

$$\Gamma(x) = \sqrt{2\pi}x^{x-1/2}e^{-x}e^{\mu(x)}.$$

Thus

$$C = 2^{\frac{p}{2}}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} = g(p)\sqrt{2}e^{-p/2}p^{p/2},$$

with $g(p) = \left(1+\frac{1}{p}\right)^{p/2}e^{v(p)-1/2}$, where $0 < v(p) < 1/(6(p+1))$. By Taylor's formula we have that

$$\log(1+x) = \sum_{m=1}^{\infty}\frac{1}{m}(-1)^{m-1}x^m, \quad \forall x \in (-1,1],$$

and that for every $k \in \mathbb{N}_0$

$$\sum_{m=1}^{2k}\frac{1}{m}(-1)^{m-1}x^m \leq \log(1+x) \leq \sum_{m=1}^{2k+1}\frac{1}{m}(-1)^{m-1}x^m, \forall x \geq 0.$$

Therefor we obtain with $k = 1$ that

$$\log g(p) = \frac{p}{2}\log(1+\frac{1}{p}) + v(p) - \frac{1}{2} \leq -\frac{1}{4p} + \frac{1}{6p^2} + \frac{1}{6(p+1)} \leq -\frac{1}{18p},$$

where the last equality follows from elementary calculus. Similarly,

$$\log g(p) = \frac{p}{2}\log(1+\frac{1}{p}) + v(p) - \frac{1}{2} \geq -\frac{1}{4p} + v(p) \geq -\frac{1}{4p},$$

Thus, we have

$$e^{-\frac{1}{4p}}\sqrt{2}e^{-p/2}p^{p/2} < C < e^{-\frac{1}{18p}}\sqrt{2}e^{-p/2}p^{p/2},$$

which implies that C is strictly smaller than $\sqrt{2}e^{-p/2}p^{p/2}$ for all $p \geq 2$.

Since $C = \frac{1}{g(p)}\sqrt{2}e^{-p/2}p^{p/2}$ and $g(p) \geq e^{-\frac{1}{4p}}$, we can obtain that the relative error between C and $\sqrt{2}e^{-p/2}p^{p/2}$ is equal to

$$\frac{1}{g(p)} - 1 \leq e^{-\frac{1}{4p}} - 1 \leq \frac{1}{4p}e^{\frac{1}{4p}}$$

using Mean Value Theorem. This implies that the corresponding relative errors between C and $\sqrt{2}e^{-p/2}p^{p/2}$ converge to zero as p tends to infinity.

The proof is complete.

□

Then we introduce the McDiarmid's inequality for vector-valued functions. We firstly consider real-valued functions, which follows from the standard tail-bound of McDiarmid's inequality and Proposition 2.5.2 in Vershynin [2018].

Lemma 11 (McDiarmid's Inequality for real-valued functions)**.** *Let $Z_i, \dots, Z_n$ be independent random variables, and $f : \mathcal{Z}^n \mapsto \mathbb{R}$ such that the following inequality holds for any $z_i, \dots, z_{i-1}, z_{i+1}, \dots, z_n$*

$$\sup_{z_i, z_i'} |f(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) - f(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)| \le \beta,$$

Then for any $p > 1$ we have

$$\|f(Z_1, \dots, Z_n) - \mathbb{E}f(Z_1, \dots, Z_n)\|_p \le \sqrt{2pn}\beta.$$

To derive the McDiarmid's inequality for vector-valued functions, we need the expected distance between $\mathbf{f}(Z_1, \dots, Z_n)$ and its expectation.

Lemma 12 ([Rivasplata et al., 2018])**.** *Let $Z_i, \dots, Z_n$ be independent random variables, and $\mathbf{f} : \mathcal{Z}^n \mapsto \mathcal{H}$ is a function into a Hilbert space $\mathcal{H}$ such that the following inequality holds for any $z_i, \dots, z_{i-1}, z_{i+1}, \dots, z_n$*

$$\sup_{z_i, z_i'} \|\mathbf{f}(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) - \mathbf{f}(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)\| \le \beta,$$

Then we have

$$\mathbb{E}\left[\|\mathbf{f}(Z_1, \dots, Z_n) - \mathbb{E}\mathbf{f}(Z_1, \dots, Z_n)\|\right] \le \sqrt{n}\beta.$$

Now, we can easily derive the p-norm McDiarmid's inequality for vector-valued functions which refines from Fan and Lei [2024] with better constants.

Lemma 13 (McDiarmid's inequality for vector-valued functions)**.** *Let $Z_i, \dots, Z_n$ be independent random variables, and $\mathbf{f} : \mathcal{Z}^n \mapsto \mathcal{H}$ is a function into a Hilbert space $\mathcal{H}$ such that the following inequality holds for any $z_i, \dots, z_{i-1}, z_{i+1}, \dots, z_n$*

$$\sup_{z_i, z_i'} \|\mathbf{f}(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) - \mathbf{f}(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)\| \le \beta, \tag{13}$$

Then for any $p > 1$ we have

$$\left\|\|\mathbf{f}(Z_1, \dots, Z_n) - \mathbb{E}\mathbf{f}(Z_1, \dots, Z_n)\|\right\|_p \le (\sqrt{2p} + 1)\sqrt{n}\beta.$$

Proof of Lemma 13. Define a real-valued function $h : \mathcal{Z}^n \mapsto \mathbb{R}$ as

$$h(z_1, \dots, z_n) = \|\mathbf{f}(z_1, \dots, z_n) - \mathbb{E}[\mathbf{f}(Z_1, \dots, Z_n)]\|.$$

We notice that this function satisfies the increment condition. For any i and $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$, we have

$$\begin{aligned}
&\sup_{z_i, z_i'} |h(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) - h(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)| \\
=&\sup_{z_i, z_i'} \left|\|\mathbf{f}(z_1, \dots, z_n) - \mathbb{E}[\mathbf{f}(Z_1, \dots, Z_n)]\| - \|\mathbf{f}(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n) - \mathbb{E}[\mathbf{f}(Z_1, \dots, Z_n)]\|\right| \\
\le&\sup_{z_i, z_i'} \left|\|\mathbf{f}(z_1, \dots, z_n) - \mathbf{f}(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)\| \le \beta.\right.
\end{aligned}$$

Therefore, we can apply Lemma 11 to the real-valued function h and derive the following inequality

$$\|h(Z_1, \dots, Z_n) - \mathbb{E}[h(Z_1, \dots, Z_n)]\|_p \le \sqrt{2pn}\beta.$$

According to Lemma 12, we know the following inequality $\mathbb{E}[h(Z_1, \dots, Z_n)] \le \sqrt{n}\beta$. Combing the above two inequalities together and we can derive the following inequality

$$\begin{aligned}
&\left\|\|\mathbf{f}(Z_1, \dots, Z_n) - \mathbb{E}\mathbf{f}(Z_1, \dots, Z_n)\|\right\|_p \\
\le&\|h(Z_1, \dots, Z_n) - \mathbb{E}[h(Z_1, \dots, Z_n)]\|_p + \|\mathbb{E}[h(Z_1, \dots, Z_n)]\|_p \\
\le&(\sqrt{2p} + 1)\sqrt{n}\beta.
\end{aligned}$$

The proof is complete.

□

Proof of Theorem 1. For $\mathbf{g}(Z_1, \dots, Z_n)$ and $A \subset [n]$, we write $\|\|\mathbf{g}\|\|_p(Z_A) = (\mathbb{E}\left[\|f\|^p \, Z_A\right])^{\frac{1}{p}}$. Without loss of generality, we suppose that $n = 2^k$. Otherwise, we can add extra functions equal to zero, increasing the number of therms by at most two times.

Consider a sequence of partitions $\mathcal{P}_0, \dots, \mathcal{P}_k$ with $\mathcal{P}_0 = \{\{i\} : i \in [n]\}, \mathcal{P}_k$ with $\mathcal{P}_n = \{[n]\}$, and to get $\mathcal{P}_l$ from $\mathcal{P}_{l+1}$ we split each subset in $\mathcal{P}_{l+1}$ into two equal parts. We have

$$\mathcal{P}_0 = \{\{1\}, \dots, \{2^k\}\}, \quad \mathcal{P}_1 = \{\{1,2\}, \{3,4\}, \dots, \{2^k - 1, 2^k\}\}, \quad \mathcal{P}_k = \{\{1, \dots, 2^k\}\}.$$

We have $|\mathcal{P}_l| = 2^{k-l}$ and $|P| = 2^l$ for each $P \in \mathcal{P}_l$. For each $i \in [n]$ and $l = 0, \dots, k$, denote by $P^l(i) \in \mathcal{P}_l$ the only set from $\mathcal{P}_l$ that contains i. In particular, $P^0(i) = \{i\}$ and $P^K(i) = [n]$.

For each $i \in [n]$ and every $l = 0, \dots, k$ consider the random variables

$$\mathbf{g}_i^l = \mathbf{g}_i^l(Z_i, Z_{[n]\setminus P^l(i)}) = \mathbb{E}[\mathbf{g}_i | Z_i, Z_{[n]\setminus P^l(i)}],$$

i.e. conditioned on z_i and all the variables that are not in the same set as Z_i in the partition $\mathcal{P}_l$. In particular, $\mathbf{g}_i^0 = \mathbf{g}_i$ and $\mathbf{g}_i^k = \mathbb{E}[\mathbf{g}_i | Z_i]$. We can write a telescopic sum for each $i \in [n]$,

$$\mathbf{g}_i - \mathbb{E}[\mathbf{g}_i | Z_i] = \sum_{l=1}^{k-1} \mathbf{g}_i^l - \mathbf{g}_i^{l+1}.$$

Then, by the triangle inequality

$$\left\|\left\|\sum_{i=1}^n \mathbf{g}_i\right\|\right\|_p \le \left\|\left\|\sum_{i=1}^n \mathbb{E}[\mathbf{g}_i | Z_i]\right\|\right\|_p + \sum_{l=0}^{k-1} \left\|\left\|\sum_{i=1}^n \mathbf{g}_i^l - \mathbf{g}_i^{l+1}\right\|\right\|_p. \tag{14}$$

To bound the first term, since $\|\mathbb{E}[\mathbf{g}_i | Z_i]\| \le M$, we can check that the vector-valued function $\mathbf{f}(Z_1, \dots, Z_n) = \sum_{i=1}^n \mathbb{E}[\mathbf{g}_i | Z_i]$ satisfies (13) with $\beta = 2M$, and $\mathbb{E}[\mathbb{E}[\mathbf{g}_i | Z_i]] = 0$, applying Lemma 13 with $\beta = 2M$, we have

$$\left\|\left\|\sum_{i=1}^n \mathbb{E}[\mathbf{g}_i | Z_i]\right\|\right\|_p \le 2(\sqrt{2p} + 1)\sqrt{n} M. \tag{15}$$

Then we start to bound the second term of the right hand side of (14). Observe that

$$\mathbf{g}_i^{l+1}(Z_i, Z_{[n]\setminus P^{l+1}(i)}) = \mathbb{E}\left[\mathbf{g}_i^l(Z_i, Z_{[n]\setminus P^l(i)}) \middle| Z_i, Z_{[n]\setminus P^{l+1}(i)}\right],$$

where the expectation is taken with respect to the variables $Z_j, j \in P^{l+1}(i)\setminus P^l(i)$. Changing any Z_j would change $\mathbf{g}_i^l$ by β. Therefore, we apply Lemma 13 with $\mathbf{f} = \mathbf{g}_i^l$ where there are 2^l random variables and obtain a uniform bound

$$\left\|\left\|\mathbf{g}_i^l - \mathbf{g}_i^{l+1}\right\|\right\|_p (Z_i, Z_{[n]\setminus P^{l+1}(i)}) \le (\sqrt{2p} + 1)\sqrt{2^l}\beta, \quad \forall p \ge 2,$$

Taking integration over $(Z_i, Z_{[n]\setminus P^{l+1}(i)})$, we have $\left\|\left\|\mathbf{g}_i^l - \mathbf{g}_i^{l+1}\right\|\right\|_p \le (\sqrt{2p} + 1)\sqrt{2^l}\beta$ as well.

Next, we turn to the sum $\sum_{i \in P^l} \mathbf{g}_i^l - \mathbf{g}_i^{l+1}$ for any $P^l \in \mathcal{P}_l$. Since $\mathbf{g}_i^l - \mathbf{g}_i^{l+1}$ for $i \in P^l$ depends only on $Z_i, Z_{[n]\setminus P^l}$, the terms are independent and zero mean conditioned on $Z_{[n]\setminus P^l}$. Applying Lemma 9, we have for any $p \ge 2$,

$$\left\|\left\|\sum_{i \in P^l} \mathbf{g}_i^l - \mathbf{g}_i^{l+1}\right\|\right\|_p^p (Z_{[n]\setminus P^l}) \le \left(2 \cdot 2^{\frac{1}{2p}} \sqrt{\frac{2^l p}{e}}\right)^p \frac{1}{2^l} \sum_{i \in P^l} \left\|\left\|\mathbf{g}_i^l - \mathbf{g}_i^{l+1}\right\|\right\|_p^p (Z_{[n]\setminus P^l}).$$

Integrating with respect to $(Z_{[n]\setminus P^l})$ and using $\left\|\left\|\mathbf{g}_i^l - \mathbf{g}_i^{l+1}\right\|\right\|_p \le (\sqrt{2p} + 1)\sqrt{2^l}\beta$, we have

$$\begin{aligned}\left\|\left\|\sum_{i \in P^l} g_i^l - g_i^{l+1}\right\|\right\|_p &\le \left(2 \cdot 2^{\frac{1}{2p}} \sqrt{\frac{2^l p}{e}}\right) \frac{1}{2^l} \times 2^l (\sqrt{2p} + 1)\sqrt{2^l}\beta \\ &= 2^{1 + \frac{1}{2p}} \left(\sqrt{\frac{p}{e}}\right) (\sqrt{2p} + 1) 2^l \beta.\end{aligned}$$

Then using triangle inequality over all sets $P^l \in \mathcal{P}_l$, we have

$$\begin{aligned}\left\|\left\|\sum_{i\in[n]}\mathbf{g}_i^l-\mathbf{g}_i^{l+1}\right\|\right\|_p \leq& \sum_{P^l\in\mathcal{P}_l}\left\|\left\|\sum_{i\in P^l}\mathbf{g}_i^l-\mathbf{g}_i^{l+1}\right\|\right\|_p\\ \leq& 2^{k-l}\times 2^{1+\frac{1}{2p}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)2^l\beta\\ \leq& 2^{1+\frac{1}{2p}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)2^k\beta.\end{aligned}$$

Recall that $2^k \leq n$ due to the possible extension of the sample. Then we have

$$\sum_{i=0}^{k-1}\left\|\left\|\sum_{i=1}^{n}\mathbf{g}_i^l-\mathbf{g}_i^{i+1}\right\|\right\|_p \leq 2^{2+\frac{1}{2p}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)n\beta\left\lceil\log_2 n\right\rceil.$$

We can plug the above bound together with (15) into (14), to derive the following inequality

$$\left\|\left\|\sum_{i=1}^{n}\mathbf{g}_i\right\|\right\|_p \leq 2(\sqrt{2p}+1)\sqrt{n}M+2^{2+\frac{1}{2p}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)n\beta\left\lceil\log_2 n\right\rceil.$$

The proof is completed.

□

C Proofs of Section 3

Proof of Theorem 2. Let $S=\{z_1,\ldots,z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$ and $S'=\{z_1',\ldots,z_n'\}$ be its independent copy. For any $i\in[n]$, define $S^{(i)}=\{z_i,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n\}$ be a dataset replacing the i-th sample in S with another i.i.d. sample z_i'. Then we can firstly write the following decomposition

$$\begin{aligned}&n\nabla F(A(S))-n\nabla F_S(A(S))\\ =&\sum_{i=1}^{n}\mathbb{E}_Z\left[\nabla f(A(S);Z)\right]-\mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}),Z)\right]\Big]\\ &+\sum_{i=1}^{n}\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}),Z)\right]-\nabla f(A(S^{(i)}),z_i)\right]\\ &+\sum_{i=1}^{n}\mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}),z_i)\right]-\sum_{i=1}^{n}\nabla f(A(S),z_i).\end{aligned}$$

We denote that $\mathbf{g}_i(S)=\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}),Z)\right]-\nabla f(A(S^{(i)}),z_i)\right]$, thus we have

$$\begin{aligned}&\|n\nabla F(A(S))-n\nabla F_S(A(S))\|_2\\ =&\Bigg\|\sum_{i=1}^{n}\mathbb{E}_Z\left[\nabla f(A(S);Z)\right]-\mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}),Z)\right]\Big]\\ &+\sum_{i=1}^{n}\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}),Z)\right]-\nabla f(A(S^{(i)}),z_i)\right]\\ &+\sum_{i=1}^{n}\mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}),z_i)\right]-\sum_{i=1}^{n}\nabla f(A(S),z_i)\Bigg\|_2\\ \leq& 2n\beta+\left\|\sum_{i=1}^{n}\mathbf{g}_i(S)\right\|_2,\end{aligned}\tag{16}$$

where the inequality holds from the definition of uniform stability in gradients.

According to our assumptions, we get $\|\mathbf{g}_i(S)\|_2 \leq 2M$ and

$$\begin{aligned}\mathbb{E}_{z_i}[\mathbf{g}_i(S)] &= \mathbb{E}_{z_i}\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)});Z)\right] - \nabla f(A(S^{(i)});z_i)\right] \\ &= \mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)});Z)\right] - \mathbb{E}_{z_i}\left[\nabla f(A(S^{(i)});z_i)\right]\right] = 0,\end{aligned}$$

where this equality holds from the fact that z_i and Z follow from the same distribution. For any $i \in [n]$, any $j \neq i$ and any z_j'', we have

$$\begin{aligned}&\left\|\mathbf{g}_i(z_1,\ldots,z_{j-1},z_j,z_{j+1},\ldots,z_n) - \mathbf{g}_i(z_1,\ldots,z_{j-1},z_j'',z_{j+1},\ldots,z_n)\right\|_2 \\ \leq&\left\|\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)});Z)\right] - \nabla f(A(S^{(i)});z_i)\right] - \mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S_j^{(i)});Z)\right] - \nabla f(A(S_j^{(i)});z_i)\right]\right\|_2 \\ \leq&\left\|\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)});Z) - \nabla f(A(S_j^{(i)});Z)\right]\right]\right\|_2 + \left\|\mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)});Z)\right] - \nabla f(A(S_j^{(i)});z_i)\right]\right\|_2 \\ \leq&2\beta,\end{aligned}$$

where $S^{(i)} = \{z_i,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n\}$. Thus, we have verified that three conditions in Theorem 1 are satisfied for $\mathbf{g}_i(S)$. We have the following result for any $p > 2$

$$\left\|\left\|\sum_{i=1}^n \mathbf{g}_i(S)\right\|\right\|_p \leq 4(\sqrt{2p}+1)\sqrt{n}M + 8\times 2^{\frac{1}{4}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)n\beta\left\lceil\log_2 n\right\rceil.$$

We can combine the above inequality and (16) to derive the following inequality

$$\begin{aligned}&n\left\|\left\|\nabla F(A(S)) - n\nabla F_S(A(S))\right\|\right\|_p \\ \leq&2n\beta + 4(\sqrt{2p}+1)\sqrt{n}M + 8\times 2^{\frac{1}{4}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)n\beta\left\lceil\log_2 n\right\rceil.\end{aligned}$$

According to Lemma 6 for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}&n\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\ \leq&2n\beta + 4\sqrt{n}M + 8\times 2^{\frac{3}{4}}\sqrt{e}n\beta\left\lceil\log_2 n\right\rceil\log(e/\delta) + (4e\sqrt{2n}M + 8\times 2^{\frac{1}{4}}\sqrt{e}n\beta\left\lceil\log_2 n\right\rceil)\sqrt{\log e/\delta}.\end{aligned}$$

This implies that

$$\begin{aligned}&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\ \leq&2\beta + \frac{4M\left(1+e\sqrt{2\log(e/\delta)}\right)}{\sqrt{n}} + 8\times 2^{\frac{1}{4}}(\sqrt{2}+1)\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log(e/\delta).\end{aligned}$$

The proof is completed. □

Proof of Theorem 3. We can firstly write the following decomposition

$$\begin{aligned}&n\nabla F(A(S)) - n\nabla F_S(A(S)) \\ =&\sum_{i=1}^n \mathbb{E}_Z\left[\nabla f(A(S);Z)] - \mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}),Z)\right]\right] \\ &+\sum_{i=1}^n \mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}),Z)\right] - \nabla f(A(S^{(i)}),z_i)\right] \\ &+\sum_{i=1}^n \mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}),z_i)\right] - \sum_{i=1}^n \nabla f(A(S),z_i).\end{aligned}$$

We denote that $\mathbf{h}_i(S) = \mathbb{E}_{z_i'}\left[\mathbb{E}_Z\left[\nabla f(A(S^{(i)}), Z)\right] - \nabla f(A(S^{(i)}), z_i)\right]$, we have

$$\begin{aligned}
&n\nabla F(A(S)) - n\nabla F_S(A(S)) - \sum_{i=1}^n \mathbf{h}_i(S) \\
=&\sum_{i=1}^n \mathbb{E}_Z\left[\nabla f(A(S); Z)\right] - \mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}), Z)\right]\Big] \\
&+ \sum_{i=1}^n \mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}), z_i)\right] - \sum_{i=1}^n \nabla f(A(S), z_i),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left\| n\nabla F(A(S)) - n\nabla F_S(A(S)) - \sum_{i=1}^n \mathbf{h}_i(S) \right\|_2 \\
=&\left\| \sum_{i=1}^n \mathbb{E}_Z\left[\nabla f(A(S); Z)\right] - \mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}), Z)\right]\Big] \right. \\
&\left. + \sum_{i=1}^n \mathbb{E}_{z_i'}\left[\nabla f(A(S^{(i)}), z_i)\right] - \sum_{i=1}^n \nabla f(A(S), z_i) \right\|_2 \\
&\leq 2n\beta,
\end{aligned} \tag{17}$$

where the inequality holds from the definition of uniform stability in gradients.

Then, for any $i = 1, \ldots, n$, we define $\mathbf{q}_i(S) = \mathbf{h}_i(S) - \mathbb{E}_{S\{z_i\}}[\mathbf{h}_i(S)]$. It is easy to verify that $\mathbb{E}_{S\setminus\{z_i\}}[\mathbf{q}_i(S)] = \mathbf{0}$ and $\mathbb{E}_{z_i}[\mathbf{h}_i(S)] = \mathbb{E}_{z_i}[\mathbf{q}_i(S)] - \mathbb{E}_{z_i}\mathbb{E}_{S\setminus\{z_i\}}[\mathbf{q}_i(S)] = \mathbf{0} - \mathbf{0} = \mathbf{0}$. Also, for any $j \in [n]$ with $j \neq i$ and $z_j'' \in \mathcal{Z}$, we have the following inequality

$$\begin{aligned}
&\|\mathbf{q}_i(S) - \mathbf{q}_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\|_2 \\
\leq&\|\mathbf{h}_i(S) - \mathbf{h}_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\|_2 \\
&+ \|\mathbb{E}_{S\setminus\{z_i\}}[\mathbf{h}_i(S)] - \mathbb{E}_{S\setminus\{z_i\}}[\mathbf{h}_i(1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)]\|_2.
\end{aligned}$$

For the first term $\|\mathbf{h}_i(S) - \mathbf{h}_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\|_2$, it can be bounded by 2β according to the definition of uniform stability. Similar result holds for the second term $\|\mathbb{E}_{S\setminus\{z_i\}}[\mathbf{h}_i(S)] - \mathbb{E}_{S\setminus\{z_i\}}[\mathbf{h}_i(1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)]\|_2$ according to the uniform stability. By a combination of the above analysis, we get $\|\mathbf{q}_i(S) - \mathbf{q}_i(1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\|_2 \leq \|\mathbf{h}_i(S) - \mathbf{h}_i(1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\|_2 \leq 4\beta$.

Thus, we have verified that three conditions in Theorem 1 are satisfied for $\mathbf{q}_i(S)$. We have the following result for any $p \geq 2$

$$\left\| \left\| \sum_{i=1}^n \mathbf{q}_i(S) \right\| \right\|_p \leq 2^{4+\frac{1}{4}} \left(\sqrt{\frac{p}{e}}\right) (\sqrt{2p} + 1) n\beta \left\lceil \log_2 n \right\rceil. \tag{18}$$

Furthermore, we can derive that

$$\begin{aligned}
&n\nabla F(A(S)) - n\nabla F_S(A(S)) - \sum_{i=1}^n \mathbf{h}_i(S) + \sum_{i=1}^n \mathbf{q}_i(S) \\
=&n\nabla F(A(S)) - n\nabla F_S(A(S)) - \sum_{i=1}^n \mathbb{E}_{S\setminus\{z_i\}}[\mathbf{h}_i(S)] \\
=&n\nabla F(A(S)) - n\nabla F_S(A(S)) - n\mathbb{E}_{S'}[\nabla F(A(S'))] + n\mathbb{E}_S[\nabla F(A(S))].
\end{aligned}$$

Due to the i.i.d. property between S and S', we know that $\mathbb{E}_{S'}[\nabla F(A(S'))] = \mathbb{E}_S[\nabla F(A(S))]$. Thus, combined above equality, (17) and (18), we have

$$
\begin{aligned}
&\left\|\left\|n\nabla F(A(S)) - n\nabla F_S(A(S)) - n\mathbb{E}_S[\nabla F(A(S))] + n\mathbb{E}_{S'}[\nabla F_S(A(S'))]\right\|\right\|_p \\
\leq &\left\|\left\|n\nabla F(A(S)) - n\nabla F_S(A(S)) - \sum_{i=1}^n \mathbf{h}_i(S)\right\|\right\|_p \\
+ &\left\|\left\|\sum_{i=1}^n \mathbf{h}_i(S) - n\mathbb{E}_S[\nabla F(A(S))] + n\mathbb{E}_{S'}F_S[A(S')]\right\|\right\|_p \\
= &\left\|\left\|n\nabla F(A(S)) - n\nabla F_S(A(S)) - \sum_{i=1}^n \mathbf{h}_i(S)\right\|\right\|_p + \left\|\left\|\sum_{i=1}^n \mathbf{q}_i(S)\right\|\right\|_p \\
\leq &2n\beta + 2^{4+\frac{1}{4}}\left(\sqrt{\frac{p}{e}}\right)(\sqrt{2p}+1)n\beta\left\lceil\log_2 n\right\rceil \\
\leq &16\times 2^{\frac{3}{4}}\left(\sqrt{\frac{1}{e}}\right)pn\beta\left\lceil\log_2 n\right\rceil + 32\left(\sqrt{\frac{1}{e}}\right)\sqrt{p}n\beta\left\lceil\log_2 n\right\rceil.
\end{aligned}
$$

According to Lemma 6 for any $\delta \in (0,1)$, with probability at least $1-\delta/3$, we have

$$
\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq &\|\mathbb{E}_{S'}[\nabla F_S(A(S'))] - \mathbb{E}_S[\nabla F(A(S))]\|_2 \\
&+ 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned} \tag{19}
$$

Next, we need to bound the term $\|\mathbb{E}_{S'}[\nabla F_S(A(S'))] - \mathbb{E}_S[\nabla F(A(S))]\|_2$. There holds that $\|\mathbb{E}_S\mathbb{E}_{S'}[\nabla F_S(A(S'))]\|_2 = \|\mathbb{E}_S[\nabla F(A(S))]\|_2$. Then, by the Bernstein inequality in Lemma 7, we obtain the following inequality with probability at least $1-\delta/3$,

$$
\left\|\mathbb{E}_{S'}[\nabla F_S(A(S'))] - \mathbb{E}_S[\nabla F(A(S))]\right\|_2 \leq \sqrt{\frac{2\mathbb{E}_{z_i}[\|\mathbb{E}_{S'}\nabla f(A(S');z_i)\|_2^2]\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n}. \tag{20}
$$

Then using Jensen's inequality, we have

$$
\begin{aligned}
\mathbb{E}_{z_i}[\|\mathbb{E}_{S'}\nabla f(A(S');z_i)\|_2^2] &\leq \mathbb{E}_{z_i}\mathbb{E}_{S'}\|\nabla f(A(S');z_i)\|_2^2 \\
&= \mathbb{E}_Z\mathbb{E}_{S'}\|\nabla f(A(S');Z)\|_2^2 \\
&= \mathbb{E}_Z\mathbb{E}_S\|\nabla f(A(S);Z)\|_2^2.
\end{aligned} \tag{21}
$$

Combing (19), (20) with (21), we finally obtain that with probability at least $1-2\delta/3$,

$$
\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq &\sqrt{\frac{2\mathbb{E}_Z\mathbb{E}_S\|\nabla f(A(S);Z)\|_2^2\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&+ 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned} \tag{22}
$$

Next, since $S = \{z_i, \ldots, z_n\}$, we define $p = p(z_1, \ldots, z_n) = \mathbb{E}_Z[\|\nabla f(A(S);Z)\|_2^2]$ and $p_i = p_i(z_1, \ldots, z_n) = \sup_{z_i\in\mathcal{Z}} p(z_i, \ldots, z_n)$. So there holds $p_i \geq p$ for any $i = 1, \ldots, n$ and any

$\{z_1, \ldots, z_n\} \in \mathcal{Z}^n$. Also, there holds that

$$
\begin{aligned}
&\sum_{i=1}^{n}(p_i - p)^2 \\
=&\sum_{i=1}^{n}\left(\sup_{z_i \in \mathcal{Z}} \mathbb{E}_Z[\|\nabla f(A(S'); Z)\|_2^2] - \mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2]\right)^2 \\
\leq&\sum_{i=1}^{n}\left(\mathbb{E}_Z\left[\sup_{z_i \in \mathcal{Z}} \|\nabla f(A(S'); Z)\|_2^2 - \|\nabla f(A(S); Z)\|_2^2\right]\right)^2 \\
=&\sum_{i=1}^{n}\left(\mathbb{E}_Z\left[\left(\sup_{z_i \in \mathcal{Z}} \|\nabla f(A(S'); Z)\|_2 - \|\nabla f(A(S); Z)\|_2\right)\left(\sup_{z_i \in \mathcal{Z}} \|\nabla f(A(S'); Z)\|_2 + \|\nabla f(A(S); Z)\|_2\right)\right]\right)^2 \\
\leq&\sum_{i=1}^{n}\beta^2\left(\mathbb{E}_Z\left[\|\nabla f(A(S); Z)\|_2 + \sup_{z_i \in \mathcal{Z}} \|\nabla f(A(S); Z)\|_2\right]\right)^2 \\
\leq&n\beta^2\left(2\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2 + \beta]\right)^2 \\
\leq&8n\beta^2 p + 2n\beta^4,
\end{aligned}
\tag{23}
$$

where the first inequality follows from the Jensen's inequality. The second and third inequalities follow from the definition of uniform stability in gradients. The last inequality holds from that $(a+b)^2 \leq 2a^2 + 2b^2$.

From (23), we know that p is $(8n\beta^2, 2n\beta^4)$ weakly self-bounded. Thus, by Lemma 8, we obtain that with probability at least $1 - \delta/3$,

$$
\begin{aligned}
&\mathbb{E}_Z\mathbb{E}_S[\|\nabla f(A(S); Z)\|_2^2] - \mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] \\
\leq&\sqrt{(16n\beta^2\mathbb{E}_S\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] + 4n\beta^4)\log(3/\delta)} \\
=&\sqrt{(\mathbb{E}_S\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] + \frac{1}{4}\beta^2)16n\beta^2\log(3/\delta)} \\
\leq&\frac{1}{2}(\mathbb{E}_S\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] + \frac{1}{4}\beta^2) + 8n\beta^2\log(3/\delta),
\end{aligned}
$$

where the last inequality follows from that $\sqrt{ab} \leq \frac{a+b}{2}$ for all $a, b > 0$. Thus, we have

$$
\mathbb{E}_Z\mathbb{E}_S[\|\nabla f(A(S); Z)\|_2^2] \leq 2\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] + \frac{1}{4}\beta^2 + 16n\beta^2\log(3/\delta). \tag{24}
$$

Substituting (24) into (22), we finally obtain that with probability at least $1 - \delta$

$$
\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{2\left(2\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] + \frac{1}{4}\beta^2 + 16n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&+ 16 \times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned}
\tag{25}
$$

According to inequality $\sqrt{a+b} = \sqrt{a} + \sqrt{b}$ for any $a, b > 0$, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2]\log\frac{6}{\delta}}{n}} + \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&+ 16 \times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned}
$$

The proof is complete.

□

Proof of Remark 4. According to the proof in Theorem 3, we have the following inequality that with probability at least $1-\delta$

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(A(S);Z)\|_2^2]\log\frac{6}{\delta}}{n}} + \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&+ 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned} \tag{26}$$

Since $f(\mathbf{w})$ is γ-smooth, we have

$$\begin{aligned}
&\mathbb{E}_Z[\|\nabla f(A(S);Z)\|_2^2] \\
\leq&\mathbb{E}_Z[\|\nabla f(A(S);Z) - \nabla f(\mathbf{w}^*;Z)\|_2^2 + \|\nabla f(\mathbf{w}^*;Z)\|_2^2] \\
\leq&\gamma^2\|A(S)-\mathbf{w}^*\|_2^2 + \mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]
\end{aligned} \tag{27}$$

Plugging (27) into (26), we have

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{4(\gamma^2\|A(S)-\mathbf{w}^*\|_2^2 + \mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2])\log\frac{6}{\delta}}{n}} + \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} \\
&+ \frac{M\log\frac{6}{\delta}}{n} + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta} \\
\leq&2\gamma\|A(S)-\mathbf{w}^*\|_2\sqrt{\frac{\log\frac{6}{\delta}}{n}} + \sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log\frac{6}{\delta}}{n}} \\
&+ \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&+ 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta},
\end{aligned} \tag{28}$$

where the second inequality holds because $\sqrt{a+b}+\sqrt{a}+\sqrt{b}$ for any $a,b>0$, which means that

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\lesssim&\beta\log n\log(1/\delta) + \frac{\log(1/\delta)}{n} + \sqrt{\frac{\mathbb{E}_Z\left[\nabla\|f(\mathbf{w}^*;Z)\|_2^2\right]\log(1/\delta)}{n}} + \|A(S)-\mathbf{w}^*\|\sqrt{\frac{\log(1/\delta)}{n}}.
\end{aligned}$$

The proof is complete.

□

Proof of Lemma 1. Inequality (28) implies that

$$\begin{aligned}
&\|\nabla F(A(S))\|_2 - \|\nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{4(\gamma^2\|A(S)-\mathbf{w}^*\|_2^2 + \mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2])\log\frac{6}{\delta}}{n}} + \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} \\
&+ \frac{M\log\frac{6}{\delta}}{n} + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta} \\
\leq&2\gamma\|A(S)-\mathbf{w}^*\|_2\sqrt{\frac{\log\frac{6}{\delta}}{n}} + \sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log\frac{6}{\delta}}{n}} + \sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} \\
&+ \frac{M\log\frac{6}{\delta}}{n} + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta},
\end{aligned}$$

When $F(\mathbf{w})$ satisfies the PL condition, there holds the following error bound property (refer to Theorem 2 in Karimi et al. [2016])

$$\|\nabla F(A(S))\|_2 \geq \mu\|A(S)-\mathbf{w}^*\|_2.$$

Thus, we have

$$\begin{aligned}
&\mu\|A(S)-\mathbf{w}^*\|_2 \leq \|\nabla F(A(S))\|_2 \\
\leq&\|\nabla F_S(A(S))\|_2 + 2\gamma\|A(S)-\mathbf{w}^*\|_2\sqrt{\frac{\log\frac{6}{\delta}}{n}} + \sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log\frac{6}{\delta}}{n}} \\
&+\sqrt{\frac{\left(\frac{1}{2}\beta^2+32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&\quad + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned}$$

When $n \geq \frac{16\gamma^2\log\frac{6}{\delta}}{\mu^2}$, we have $2\gamma\sqrt{\frac{\log\frac{6}{\delta}}{n}} \leq \frac{\mu}{2}$, then we can derive that

$$\begin{aligned}
&\mu\|A(S)-\mathbf{w}^*\|_2 \leq \|\nabla F(A(S))\|_2 \\
\leq&\|\nabla F_S(A(S))\|_2 + \frac{\mu}{2}\|A(S)-\mathbf{w}^*\|_2 + \sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log\frac{6}{\delta}}{n}} \\
&+\sqrt{\frac{\left(\frac{1}{2}\beta^2+32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&\quad + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\|A(S)-\mathbf{w}^*\|_2 \\
\leq&\frac{2}{\mu}\Big(\|\nabla F_S(A(S))\|_2 + \sqrt{\frac{4\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log\frac{6}{\delta}}{n}} \\
&+\sqrt{\frac{\left(\frac{1}{2}\beta^2+32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} \\
&\quad + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 32\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}\Big).
\end{aligned} \tag{29}$$

Then, substituting (29) into (28), when $n \geq \frac{16\gamma^2\log\frac{6}{\delta}}{\mu^2}$, with probability at least $1-\delta$

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\| \\
\leq&\|\nabla F_S(A(S))\| + 4\sqrt{\frac{\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|^2]\log\frac{6}{\delta}}{n}} + 2\sqrt{\frac{\left(\frac{1}{2}\beta^2+32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} \\
&+\frac{2M\log\frac{6}{\delta}}{n} + 32\times 2^{\frac{3}{4}}\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right) + 64\sqrt{e}\beta\left\lceil\log_2 n\right\rceil\sqrt{\log 3e/\delta}.
\end{aligned}$$

The proof is complete. □

Proof of Remark 5. Here we briefly prove the results given in Remark 5. Since F satisfies the PL condition with μ, we have

$$F(A(S)) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(A(S))\|^2}{2\mu}, \quad \forall \mathbf{w} \in \mathcal{W}. \tag{30}$$

So to bound $F(A(S)) - F(A(S))$, we need to bound the term $\|\nabla F(A(S))\|^2$. And there holds

$$\|\nabla F(A(S))\|_2^2 = 2\|\nabla F(A(S)) - \nabla F_S(A(S))\|^2 + 2\|\nabla F_S(A(S))\|_2^2. \tag{31}$$

From Lemma 1, if f is M-Lipschitz and γ-smooth and F satisfies PL condition with μ, for any $\delta > 0$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1-\delta$, there holds

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq& \|\nabla F_S(A(S))\|_2 + C\left(\sqrt{\frac{2\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*; Z)\|_2^2]\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} + e\beta\lceil\log_2 n\rceil\log(3e/\delta)\right) \\
\leq& \|\nabla F_S(A(S))\|_2 + C\left(\sqrt{\frac{8\gamma F(\mathbf{w}^*)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} + e\beta\lceil\log_2 n\rceil\log(3e/\delta)\right),
\end{aligned}$$

where C is a positive constant and the last inequality follows from Lemma 4.1 of Srebro et al. [2010] when f is nonegative and γ-smooth (see (44)).

Combing above inequality with (30), (31), we can derive that

$$F(A(S)) - F(\mathbf{w}^*) \lesssim \|\nabla F_S(A(S))\|_2 + \frac{F(\mathbf{w}^*)\log(1/\delta)}{n} + \frac{M\log^2(1/\delta)}{n^2} + \beta^2\log^2 n\log^2(1/\delta).$$

The proof is complete.

□

Proof of Lemma 2. According to the proof in Theorem 3, we have the following inequality with probability at least $1-\delta$

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{2\left(2\mathbb{E}_Z[\|\nabla f(A(S); Z)\|_2^2] + \frac{1}{4}\beta^2 + 16n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} \\
&+ \frac{M\log\frac{6}{\delta}}{n} + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\lceil\log_2 n\rceil\log(3e/\delta) + 32\sqrt{e}\beta\lceil\log_2 n\rceil\sqrt{\log 3e/\delta}.
\end{aligned} \tag{32}$$

Since SGC implies that $\mathbb{E}_Z[\|\nabla f(\mathbf{w}; Z)\|_2^2] \leq \rho\|\nabla F(\mathbf{w})\|_2^2$, according to inequalities $\sqrt{ab} \leq \eta a + \frac{1}{\eta}b$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b, \eta > 0$, we have the following inequality with probability at least $1-\delta$

$$\begin{aligned}
&\|\nabla F(A(S)) - \nabla F_S(A(S))\|_2 \\
\leq&\sqrt{\frac{2\left(2\rho\|\nabla F(A(S))\|_2^2 + \frac{1}{4}\beta^2 + 16n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} \\
&+ \frac{M\log\frac{6}{\delta}}{n} + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\lceil\log_2 n\rceil\log(3e/\delta) + 32\sqrt{e}\beta\lceil\log_2 n\rceil\sqrt{\log 3e/\delta} \\
\leq&\sqrt{\frac{\left(\frac{1}{2}\beta^2 + 32n\beta^2\log(3/\delta)\right)\log\frac{6}{\delta}}{n}} + \frac{\eta}{1+\eta}\|\nabla F(A(S))\| + \frac{1+\eta}{\eta}\frac{4\rho M\log\frac{6}{\delta}}{n} \\
&+ \frac{M\log\frac{6}{\delta}}{n} + 16\times 2^{\frac{3}{4}}\sqrt{e}\beta\lceil\log_2 n\rceil\log(3e/\delta) + 32\sqrt{e}\beta\lceil\log_2 n\rceil\sqrt{\log 3e/\delta}.
\end{aligned}$$

which implies that

$$\|\nabla F(A(S))\|_2 \leq (1+\eta)\|\nabla F_S(A(S))\|_2 + C\frac{1+\eta}{\eta}\left(\frac{M}{n}\log\frac{6}{\delta} + \beta\log n\log\frac{1}{\delta}\right).$$

The proof is complete. □

D Proofs of ERM

Proof of Lemma 3. Since $F_{S^{(i)}}(\mathbf{w}) = \frac{1}{n}\left(f(\mathbf{w}; z_i') + \sum_{j\neq i} f(\mathbf{w}, z_j)\right)$, we have

$$\begin{aligned}
&F_S(\hat{\mathbf{w}}^*(S^{(i)})) - F_S(\hat{\mathbf{w}}^*(S)) \\
=& \frac{f(\hat{\mathbf{w}}^*(S^{(i)}); z_i) - f(\hat{\mathbf{w}}^*(S); z_i)}{n} + \frac{\sum_{j\neq i}(f(\hat{\mathbf{w}}^*(S^{(i)}); z_j) - f(\hat{\mathbf{w}}^*(S); z_j))}{n} \\
=& \frac{f(\hat{\mathbf{w}}^*(S^{(i)}); z_i) - f(\hat{\mathbf{w}}^*(S); z_i)}{n} + \frac{f(\hat{\mathbf{w}}^*(S); z_i') - f(\hat{\mathbf{w}}^*(S^{(i)}); z_i')}{n} \\
&+ \left(F_{S^{(i)}}(\hat{\mathbf{w}}^*(S^{(i)})) - F_{S^{(i)}}(\hat{\mathbf{w}}^*(S))\right) \\
\leq& \frac{f(\hat{\mathbf{w}}^*(S^{(i)}); z_i) - f(\hat{\mathbf{w}}^*(S); z_i)}{n} + \frac{f(\hat{\mathbf{w}}^*(S); z_i') - f(\hat{\mathbf{w}}^*(S^{(i)}); z_i')}{n} \\
\leq& \frac{2M}{n}\|\hat{\mathbf{w}}^*(S^{(i)}) - \hat{\mathbf{w}}^*(S)\|_2,
\end{aligned}$$

where the first inequality follows from the fact that $\hat{\mathbf{w}}^*(S^{(i)})$ is the ERM of $F_{S^{(i)}}$ and the second inequality follows from the Lipschitz property. Furthermore, for $\hat{\mathbf{w}}^*(S^{(i)})$, the convexity of f and the strongly-convex property of F_S imply that its closest optima point of F_S is $\hat{\mathbf{w}}^*(S)$ (the global minimizer of F_S is unique). Then, there holds that

$$F_S(\hat{\mathbf{w}}^*(S^{(i)})) - F_S(\hat{\mathbf{w}}^*(S)) \geq \frac{\mu}{2}\|\hat{\mathbf{w}}^*(S^{(i)}) - \hat{\mathbf{w}}^*(S)\|_2^2.$$

Then we get

$$\frac{\mu}{2}\|\hat{\mathbf{w}}^*(S^{(i)}) - \hat{\mathbf{w}}^*(S)\|_2^2 \leq F_S(\hat{\mathbf{w}}^*(S^{(i)})) - F_S(\hat{\mathbf{w}}^*(S)) \leq \frac{2M}{n}\|\hat{\mathbf{w}}^*(S^{(i)}) - \hat{\mathbf{w}}^*(S)\|_2,$$

which implies that $\|\hat{\mathbf{w}}^*(S^{(i)}) - \hat{\mathbf{w}}^*(S)\|_2 \leq \frac{4M}{n\mu}$. Combined with the smoothness property of f we obtain that for any $S^{(i)}$ and S

$$\forall z \in \mathcal{Z}, \quad \left\|\nabla f(\hat{\mathbf{w}}^*(S^{(i)}); z) - \nabla f(\hat{\mathbf{w}}^*(S); z)\right\|_2 \leq \frac{4M\gamma}{n\mu}.$$

The proof is complete. □

Proof of Theorem 4. Since F is μ-strongly convex, we have

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w})\|_2^2}{2\mu}, \quad \forall \mathbf{w} \in \mathcal{W}. \tag{33}$$

So to bound $F(\hat{\mathbf{w}}^*) - F(\mathbf{w}^*)$, we need to bound the term $\|\nabla F(\hat{\mathbf{w}}^*)\|_2^2$. And there holds

$$\|\nabla F(\hat{\mathbf{w}}^*)\|_2^2 = 2\|\nabla F(\hat{\mathbf{w}}^*) - \nabla F_S(\hat{\mathbf{w}}^*)\|_2^2 + 2\|\nabla F_S(\hat{\mathbf{w}}^*)\|_2^2. \tag{34}$$

From Lemma 1, if f is M-Lipschitz and γ-smooth and F_S is μ-strongly convex, for any $\delta > 0$, when $n \geq \frac{16\gamma^2 \log\frac{6}{\delta}}{\mu^2}$, with probability at least $1-\delta$, there holds

$$\begin{aligned}
&\|\nabla F(\hat{\mathbf{w}}^*) - \nabla F_S(\hat{\mathbf{w}}^*)\|_2 \\
\leq& \|\nabla F_S(\hat{\mathbf{w}}^*)\|_2 + C\left(\sqrt{\frac{2\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*; Z)\|_2^2]\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} + e\hat{\beta}\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right)\right) \\
\leq& \|\nabla F_S(\hat{\mathbf{w}}^*)\|_2 + C\left(\sqrt{\frac{8\gamma F(\mathbf{w}^*)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} + e\hat{\beta}\left\lceil\log_2 n\right\rceil\log\left(3e/\delta\right)\right),
\end{aligned} \tag{35}$$

where the last inequality follows from Lemma 4.1 of Srebro et al. [2010] when f is nonegative (see (44)) and γ-smooth and $\hat{\beta} = \|\nabla f(\hat{\mathbf{w}}^*(S); z) - \nabla f(\hat{\mathbf{w}}^*(S'); z)\|_2$. C is a positive constant.

From Lemma 3, we have $\|\nabla f(\hat{\mathbf{w}}^*(S); z) - \nabla f(\hat{\mathbf{w}}^*(S'); z)\|_2 \leq \frac{4M\gamma}{n\mu}$. Since $\nabla F_S(\hat{\mathbf{w}}^*) = 0$, we have $\|\nabla F_S(\hat{\mathbf{w}}^*)\|_2 = 0$, then we can derive that

$$F(\mathbf{w}) - F(\mathbf{w}^*) \lesssim \frac{F(\mathbf{w}^*)\log(1/\delta)}{n} + \frac{\log^2 n \log^2(1/\delta)}{n^2}.$$

□

E Proofs of PGD

Proof of Theorem 5. According to smoothness assumption and $\eta = 1/\gamma$, we can derive that

$$\begin{aligned}
&F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \\
\leq &\langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle + \frac{\gamma}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
= &-\eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2 + \frac{\gamma}{2}\eta_t^2 \|\nabla F_S(\mathbf{w}_t)\|_2^2 \\
= &\left(\frac{\gamma}{2}\eta_t^2 - \eta_t\right)\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\
\leq &-\frac{1}{2}\eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

According to above inequality and the assumptions that F_S is μ-strongly convex, we can prove that

$$F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \leq -\frac{1}{2}\eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2 \leq -\mu\eta_t (F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*)),$$

which implies that

$$F_S(\mathbf{w}_{t+1}) - F_S(\hat{\mathbf{w}}^*) \leq (1 - \mu\eta_t)(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*)).$$

According to the property for γ-smooth for F_S and the property for μ-strongly convex for F_S, we have

$$\frac{1}{2\gamma}\|\nabla F_S(\mathbf{w})\|_2^2 \leq F_S(\mathbf{w}) - F_S(\hat{\mathbf{w}}^*) \leq \frac{1}{2\mu}\|\nabla F_S(\mathbf{w})\|_2^2,$$

which means that $\frac{\mu}{\gamma} \leq 1$.

Then If $\eta_t = 1/\gamma$, $0 \leq 1 - \mu\eta_t < 1$, taking over T iterations, we get

$$F_S(\mathbf{w}_{t+1}) - F_S(\hat{\mathbf{w}}^*) \leq (1 - \mu\eta_t)^T (F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*)). \tag{36}$$

Combined (36), the smoothness of F_S and the nonnegative property of f, it can be derive that

$$\|\nabla F_S(\mathbf{w}_{T+1}))\|_2^2 = O\left((1 - \frac{\mu}{\gamma})^T\right).$$

Furthermore, since F is μ-strongly convex, we have

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w})\|_2^2}{2\mu}, \quad \forall \mathbf{w} \in \mathcal{W}. \tag{37}$$

So to bound $F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)$, we need to bound the term $\|\nabla F(\mathbf{w}_{T+1})\|_2^2$. And there holds

$$\|\nabla F(\mathbf{w}_{T+1})\|_2^2 = 2\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|_2^2 + 2\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2. \tag{38}$$

From Lemma 1, if f is M-Lipschitz and γ-smooth and F_S is μ-strongly convex, for any $\delta > 0$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1-\delta$, there holds

$$\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|_2 \\
&\leq \|\nabla F_S(\mathbf{w}_{T+1})\|_2 + C\left(\sqrt{\frac{2\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} + e\beta\lceil\log_2 n\rceil\log(3e/\delta)\right) \\
&\leq \|\nabla F_S(\mathbf{w}_{T+1})\|_2 + C\left(\sqrt{\frac{8\gamma F(\mathbf{w}^*)\log\frac{6}{\delta}}{n}} + \frac{M\log\frac{6}{\delta}}{n} + e\beta\lceil\log_2 n\rceil\log(3e/\delta)\right),
\end{aligned} \tag{39}$$

where the last inequality follows from Lemma 4.1 of Srebro et al. [2010] when f is nonegative and γ-smooth (see (44)) and $\beta = \|\nabla f(\mathbf{w}_{T+1}(S);z) - \nabla f(\mathbf{w}_{T+1}(S');z)\|_2$. C is a positive constant.

From Lemma 4, we have $\beta = \|\nabla f(\mathbf{w}_{T+1}(S);z) - \nabla f(\mathbf{w}_{T+1}(S');z)\|_2 \leq \frac{2M\gamma}{n\mu}$. Since $\|\nabla F_S(\mathbf{w}_{T+1})\|_2 = O\left((1-\frac{\mu}{\gamma})^T\right)$, then we can derive that

$$F(\mathbf{w}) - F(\mathbf{w}^*) \lesssim \left(1-\frac{\mu}{\gamma}\right)^{2T} + \frac{F(\mathbf{w}^*)\log(1/\delta)}{n} + \frac{\log^2 n\log^2(1/\delta)}{n^2}.$$

Let $T \asymp \log n$, we have

$$F(\mathbf{w}) - F(\mathbf{w}^*) \lesssim \frac{F(\mathbf{w}^*)\log(1/\delta)}{n} + \frac{\log^2 n\log^2(1/\delta)}{n^2}.$$

The proof is complete. □

F Proofs of SGD

We first introduce some necessary lemmata on the empirical risk.

Lemma 14 ([Lei and Tang, 2021]). *Let $\{\mathbf{w}_t\}_t$ be the sequence produced by SGD with $\eta_t \leq \frac{1}{2\gamma}$ for all $t \in \mathbb{N}$. Suppose Assumption 1 hold. Assume for all z, the function $\mathbf{w} \mapsto f(\mathbf{w};z)$ is M-Lipschitz and γ-smooth. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, there holds that*

$$\sum_{k=1}^t \eta_k \|\nabla F_S(\mathbf{w}_k)\|_2^2 = O\left(\log\frac{1}{\delta} + \sum_{k=1}^t \eta_k^2\right).$$

Lemma 15 ([Lei and Tang, 2021]). *Let $\{\mathbf{w}_t\}_t$ be the sequence produced by SGD with $\eta_t = \frac{2}{\mu(t+t_0)}$ such that $t_0 \geq \max\{\frac{4\gamma}{\mu}, 1\}$ for all $t \in \mathbb{N}$. Suppose Assumption 1 hold. Assume for all z, the function $\mathbf{w} \mapsto f(\mathbf{w};z)$ is M-Lipschitz and γ-smooth and assume F_S satisfies PL condition with parameter μ. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, there holds that*

$$F_S(\mathbf{w}_{T+1}) - F_S(\hat{\mathbf{w}}^*) = O\left(\frac{\log(T)\log^3(1/\delta)}{T}\right).$$

Lemma 16 ([Lei and Tang, 2021]). *Let e be the base of the natural logarithm. There holds the following elementary inequalities.*

- *If $\theta \in (0,1)$, then $\sum_{k=1}^t k^{-\theta} \leq t^{1-\theta}/(1-\theta)$;*
- *If $\theta = 1$, then $\sum_{k=1}^t k^{-\theta} \leq \log(et)$;*
- *If $\theta > 1$, then $\sum_{k=1}^t k^{-\theta} \leq \frac{\theta}{\theta-1}$.*

Proof of Lemma 5. We have known that $F_{S^{(i)}}(\mathbf{w}) = \frac{1}{n}\left(f(\mathbf{w};z_i') + \sum_{j\neq i} f(\mathbf{w};z_j)\right)$. We denote $\hat{\mathbf{w}}^*(S^{(i)})$ be the ERM of $F_{S^{(i)}}(\mathbf{w})$ and $\hat{\mathbf{w}}_S^*$ be the ERM of $F_S(\mathbf{w})$. From Lemma 3, we know that

$$\forall z \in \mathcal{Z}, \quad \left\|\nabla f(\hat{\mathbf{w}}^*(S^{(i)});z) - f(\hat{\mathbf{w}}^*(S);z)\right\|_2 \leq \frac{4M\gamma}{n\mu}.$$

Also, for $\mathbf{w}_t$, the convexity of f and the strongly-convex property implies that its closest optima point of F_S is $\hat{\mathbf{w}}^*(S)$ (the global minimizer of F_S is unique). Then, there holds that

$$\frac{\mu}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}^*(S)\|_2^2 \leq F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*(S)) = \epsilon_{opt}(\mathbf{w}_t).$$

Thus we have $\|\mathbf{w}_t - \hat{\mathbf{w}}^*(S)\|_2 \leq \sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}}$. A similar relation holds between $\hat{\mathbf{w}}^*(S^{(i)})$ and $\mathbf{w}_t^i$. Combined with the Lipschitz property of f we obtain that for $\forall z \in \mathcal{Z}$, there holds that

$$\begin{aligned}
&\left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2 \\
\leq& \|\nabla f(\mathbf{w}_t; z) - \nabla f(\hat{\mathbf{w}}^*(S); z)\|_2 + \left\|\nabla f(\hat{\mathbf{w}}^*(S); z) - \nabla f(\hat{\mathbf{w}}^*(S^{(i)}); z)\right\|_2 \\
&+ \left\|\nabla f(\hat{\mathbf{w}}^*(S^{(i)}); z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2 \\
\leq& \gamma\|\mathbf{w}_t - \hat{\mathbf{w}}^*(S)\|_2 + \frac{4M\gamma}{n\mu} + \gamma\|\hat{\mathbf{w}}^*(S^{(i)}) - \mathbf{w}_t^i\|_2 \\
\leq& \gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu} + \gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t^i)}{\mu}}.
\end{aligned}$$

According to Lemma 15, for any dataset S, the optimization error $\epsilon_{opt}(\mathbf{w}_t)$ is uniformly bounded by the same upper bound. Therefore, we write $\left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2 \leq 2\gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu}$ here.

The proof is complete. □

Now We begin to prove Lemma 6.

Proof of Lemma 6. If f is L-Lipschitz and γ-smooth and F_S is μ-strongly convex. According to Lemma 1, we know that for all $\mathbf{w} \in \mathcal{W}$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta/2$, when $n > \frac{16\gamma^2 \log\frac{6}{\delta}}{\mu^2}$, we have

$$\begin{aligned}
&\left(\sum_{t=1}^T \eta_t\right)^{-1} \sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 \\
\leq& 16\left(\sum_{t=1}^T \eta_t\right)^{-1} \sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2 + \frac{4C^2L^2\log^2\frac{6}{\delta}}{n^2} + \frac{8C^2\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*; Z)\|_2^2]\log^2\frac{6}{\delta}}{n} \\
&+ \left(\sum_{t=1}^T \eta_t\right)^{-1} \sum_{t=1}^T \eta_t C^2 e^2 \beta_t^2 \left\lceil \log_2 n \right\rceil^2 \log^2(3e/\delta),
\end{aligned} \tag{40}$$

where $\beta_t = \left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2$ and C is a positive constant.

From Lemma 5, we have $\left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2 \leq 2\gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu}$, thus

$$\begin{aligned}
\beta_t^2 &= \left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2^2 \\
&\leq \left(2\gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu}\right)^2 \\
&\leq \frac{16\gamma^2(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*(S)))}{\mu} + \frac{32M^2\gamma^2}{n^2\mu^2} \\
&\leq \frac{8\gamma^2\|\nabla F_S(\mathbf{w}_t)\|_2^2}{\mu^2} + \frac{32M^2\gamma^2}{n^2\mu^2},
\end{aligned} \tag{41}$$

where the second inequality holds from Cauchy-Bunyakovsky-Schwarz inequality and the second inequality satisfies because F_S is μ-strongly convex.

Plugging (41) into (40), with probability at least $1-\delta/2$, when $n > \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, we have

$$\begin{aligned}
&\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 \\
\leq &\left(16+\frac{8\gamma^2C^2e^2\left\lceil\log_2 n\right\rceil^2\log^2\left(6e/\delta\right)}{\mu^2}\right)\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\
&+\frac{4C^2L^2\log^2\frac{12}{\delta}}{n^2}+\frac{8C^2\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2]\log^2\frac{12}{\delta}}{n}+\frac{32L^2\gamma^2C^2e^2\left\lceil\log_2 n\right\rceil^2\log^2\left(6e/\delta\right)}{n^2\mu^2},
\end{aligned} \tag{42}$$

When $\eta_t = \eta_1 t^{-\theta}, \theta \in (0,1)$, with $\eta_1 \leq \frac{1}{2\beta}$ and Assumption 1, according to Lemma 14 and Lemma 16, we obtain the following inequality with probability at least $1-\delta/2$,

$$\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 = \begin{cases} O\left(\frac{\log(1/\delta)}{T^{-\theta}}\right), & \text{if } \theta < 1/2 \\ O\left(\frac{\log(1/\delta)}{T^{-\frac{1}{2}}}\right), & \text{if } \theta = 1/2 \\ O\left(\frac{\log(1/\delta)}{T^{\theta-1}}\right), & \text{if } \theta > 1/2. \end{cases} \tag{43}$$

On the other hand, when f is nonegative and γ-smooth, from Lemma 4.1 of Srebro et al. [2010], we have

$$\|\nabla f(\mathbf{w}^*;z)\|_2^2 \leq 4\gamma f(\mathbf{w}^*;z),$$

which implies that

$$\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*;Z)\|_2^2] \leq 4\gamma\mathbb{E}_Z f(\mathbf{w}^*;Z) = 4\gamma F(\mathbf{w}^*). \tag{44}$$

Plugging (44), (43) into (42), with probability at least $1-\delta$, we derive that

$$\begin{aligned}
&\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 \\
=&\begin{cases} O\left(\frac{\log^2 n\log^3(1/\delta)}{T^{-\theta}}\right)+O\left(\frac{\log^2 n\log^2(1/\delta)}{n^2}+\frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right), & \text{if } \theta < 1/2 \\ O\left(\frac{\log^2 n\log^3(1/\delta)}{T^{-\frac{1}{2}}}\right)+O\left(\frac{\log^2 n\log^2(1/\delta)}{n^2}+\frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right), & \text{if } \theta = 1/2 \\ O\left(\frac{\log^2 n\log^3(1/\delta)}{T^{\theta-1}}\right)+O\left(\frac{\log^2 n\log^2(1/\delta)}{n^2}+\frac{F(\mathbf{w}^*)\log^2(1/\delta)}{n}\right), & \text{if } \theta > 1/2. \end{cases}
\end{aligned}$$

When $\theta < 1/2$, we set $T \asymp n^{\frac{2}{\theta}}$ and assume $F(\mathbf{w}^*) = O(\frac{1}{n})$, then we obtain the following result with probability at least $1-\delta$

$$\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = O\left(\frac{\log^2 n\log^3(1/\delta)}{n^2}\right).$$

When $\theta = 1/2$, we set $T \asymp n^4$ and assume $F(\mathbf{w}^*) = O(\frac{1}{n})$, then we obtain the following result with probability at least $1-\delta$

$$\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = O\left(\frac{\log^2 n\log^3(1/\delta)}{n^2}\right).$$

When $\theta > 1/2$, we set $T \asymp n^{\frac{2}{1-\theta}}$ and assume $F(\mathbf{w}^*) = O(\frac{1}{n})$, then we obtain the following result with probability at least $1-\delta$

$$\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = O\left(\frac{\log^2 n\log^3(1/\delta)}{n^2}\right).$$

The proof is complete.

□

Proof of Theorem 7. Since F is μ-strongly convex, we have

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w})\|_2^2}{2\mu}, \quad \forall \mathbf{w} \in \mathcal{W}. \tag{45}$$

So to bound $F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)$, we need to bound the term $\|\nabla F(\mathbf{w}_{T+1})\|_2^2$. And there holds

$$\|\nabla F(\mathbf{w}_{T+1})\|_2^2 = 2\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 + 2\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2. \tag{46}$$

From Lemma 1, if f is L-Lipschitz and γ-smooth and F_S is μ-strongly convex, for all $\mathbf{w} \in \mathcal{W}$ and any $\delta > 0$, when $n \geq \frac{16\gamma^2 \log \frac{6}{\delta}}{\mu^2}$, with probability at least $1 - \delta/2$, there holds

$$\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|_2 \\
\leq &\|\nabla F_S(\mathbf{w}_{T+1})\|_2 + C\left(\sqrt{\frac{2\mathbb{E}_Z[\|\nabla f(\mathbf{w}^*; Z)\|_2^2]\log\frac{12}{\delta}}{n}} + \frac{M\log\frac{12}{\delta}}{n} + e\beta\left\lceil\log_2 n\right\rceil\log\left(6e/\delta\right)\right) \\
\leq &\|\nabla F_S(\mathbf{w}_{T+1})\|_2 + C\left(\sqrt{\frac{8\gamma F(\mathbf{w}^*)\log\frac{12}{\delta}}{n}} + \frac{M\log\frac{12}{\delta}}{n} + e\beta\left\lceil\log_2 n\right\rceil\log\left(6e/\delta\right)\right),
\end{aligned} \tag{47}$$

where the last inequality follows from Lemma 4.1 of Srebro et al. [2010] when f is nonegative and γ-smooth (see (44)) and C is a positive constant. Then we can derive that

$$\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|_2^2 \\
\leq &4\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2 + \frac{32C^2\gamma F(\mathbf{w}^*)\log\frac{12}{\delta}}{n} + \frac{4M^2C^2\log^2\frac{12}{\delta}}{n^2} + 4e^2\beta_{T+1}^2\left\lceil\log_2 n\right\rceil^2\log^2\left(6e/\delta\right).
\end{aligned} \tag{48}$$

From Lemma 5, we have $\left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2 \leq 2\gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu}$, thus

$$\begin{aligned}
\beta_t^2 = &\left\|\nabla f(\mathbf{w}_t; z) - \nabla f(\mathbf{w}_t^i; z)\right\|_2^2 \\
\leq &\left(2\gamma\sqrt{\frac{2\epsilon_{opt}(\mathbf{w}_t)}{\mu}} + \frac{4M\gamma}{n\mu}\right)^2 \\
\leq &\frac{16\gamma^2(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*(S)))}{\mu} + \frac{32M^2\gamma^2}{n^2\mu^2} \\
\leq &\frac{8\gamma^2\|\nabla F_S(\mathbf{w}_t)\|_2^2}{\mu^2} + \frac{32M^2\gamma^2}{n^2\mu^2},
\end{aligned} \tag{49}$$

where the second inequality holds from Cauchy-Bunyakovsky-Schwarz inequality and the second inequality satisfies because F_S is μ-strongly convex.

Plugging (49) into (48), with probability at least $1 - \delta/2$, when , we have

$$\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|_2^2 \\
\leq &\left(4 + 32e^2\left\lceil\log_2 n\right\rceil^2\log^2\left(6e/\delta\right)\right)\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2 + \frac{32C^2\gamma F(\mathbf{w}^*)\log\frac{6}{\delta}}{n} \\
&+ \frac{4L^2C^2\log^2\frac{12}{\delta}}{n^2} + \frac{128M^2\gamma^2e^2\left\lceil\log_2 n\right\rceil^2\log^2\left(6e/\delta\right)}{n^2\mu^2}.
\end{aligned} \tag{50}$$

According to the smoothness property of F_S and Lemma 15, it can be derived that with propability at least $1 - \delta/2$

$$\|\nabla F_S(\mathbf{w}_{T+1})\|_2^2 = O\left(\frac{\log T\log^3(1/\delta)}{T}\right). \tag{51}$$

Substituting (51), (50) into (46), we derive that

$$\begin{aligned}&\|\nabla F(\mathbf{w}_{T+1})\|_2^2\\ =&O\left(\frac{\lceil\log_2 n\rceil^2\log T\log^5(1/\delta)}{T}\right)+O\left(\frac{\lceil\log_2 n\rceil^2\log^2(1/\delta)}{n^2}+\frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right).\end{aligned}\tag{52}$$

Further substituting (52) into (45) and choosing $T \asymp n^2$, we finally obtain that when n, with probability at least $1-\delta$

$$F(\mathbf{w}_{T+1})-F(\mathbf{w}^*)=O\left(\frac{\log^4 n\log^5(1/\delta)}{n^2}+\frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right).$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have checked that the abstract and introduction accurately reflect our contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have clearly stated the required assumptions for each theorem and lemma, and the conditions for the assumptions to hold are also stated in the main text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have clearly stated the required assumptions for each theorem and lemma, and all proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper focuses on learning theory.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper focuses on learning theory and does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper focuses on learning theory and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper focuses on learning theory and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper focuses on learning theory and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper focuses on learning theory and there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper focuses on learning theory and poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper focuses on learning theory and does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.