Towards Sharper Risk Bounds for Minimax Problems

Bowei Zhu[1,2], **Shaojie Li**[1,2], **Yong Liu**[1,2*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
{bowei.zhu, lishaojie95, liuyonggsai}@ruc.edu.cn

Abstract

Minimax problems have achieved success in machine learning such as adversarial training, robust optimization, reinforcement learning. For theoretical analysis, current optimal excess risk bounds, which are composed by generalization error and optimization error, present 1/n-rates in strongly-convex-strongly-concave (SC-SC) settings. Existing studies mainly focus on minimax problems with specific algorithms for optimization error, with only a few studies on generalization performance, which limit better excess risk bounds. In this paper, we study the generalization bounds measured by the gradients of primal functions using uniform localized convergence. We obtain a sharper high probability generalization error bound for nonconvex-strongly-concave (NC-SC) stochastic minimax problems. Furthermore, we provide dimension-independent results under Polyak-Lojasiewicz condition for the outer layer. Based on our generalization error bound, we analyze some popular algorithms such as empirical saddle point (ESP), gradient descent ascent (GDA) and stochastic gradient descent ascent (SGDA). We derive better excess primal risk bounds with further reasonable assumptions, which, to the best of our knowledge, are n times faster than exist results in minimax problems.

1 Introduction

Modern machine learning settings such as reinforcement learning [Du *et al.*, 2017; Dai *et al.*, 2018], adversarial learning [Goodfellow *et al.*, 2016], robust optimization [Chen *et al.*, 2017; Namkoong and Duchi, 2017] often need to solve minimax problems, which divide the training process into two groups: one for minimization and one for maximization. To solve the problems, various efficient optimization algorithms such as gradient descent ascent (GDA), stochastic gradient descent ascent (SGDA) have been proposed and widely used in application.

*Corresponding author.

In theoretical analysis, an essential issue is the excess risk, which compares the risk of certain parameters to the Bayes optimal parameters. The standard technique to bound excess risk is to divide it into generalization error and optimization error. Current optimal excess primal risk[1] bounds are $O(1/n)$ in strongly-convex-strongly-concave (SC-SC) minimax problems, which are derived by [Li and Liu, 2021a]. In this paper, we derive $O(1/n^2)$ excess primal risk bounds with some reasonable assumptions, which is, to the best of our knowledge, the optimal results in minimax problems.

Since excess risk can be bounded by generalization error and optimization error, most of existing studies such as [Palaniappan and Bach, 2016; Hsieh *et al.*, 2019; Lin *et al.*, 2020; Luo *et al.*, 2020] were focused on iteration complexity for certain algorithms, which only considered the optimization error. In contrast, the generalization performance analysis is less considered, which is an important measure to foresee their prediction behavior after training and limits better excess risk bounds.

In this paper, our goal is to improve the generalization error bounds and further derives better excess risk bounds. We use local methods to consider variance information and obtain a tighter generalization error bound comparing with Rademacher complexity method [Zhang *et al.*, 2022]. Note that we introduce a novel "uniform localized convergence" framework using generic chaining developed by [Xu and Zeevi, 2020] to minimax problems which is different from traditional local Rademacher complexity technique [Bartlett *et al.*, 2002].

Our contributions are summarized as follows:

- We introduce local uniform convergence using new generic chaining techniques. Comparing with traditional uniform convergence results in [Zhang *et al.*, 2022], we derive sharper generalization bounds measured by the gradients of primal functions for NC-SC minimax problems. It provides problem independent results that can be used in various minimax algorithms.
- Under the Polyak-Lojasiewicz condition for the outer layer, we provide dimension-independent results and remove the dimension of parameters d from our generalization bound when the sample size n is large enough,

[1]Primal function is one of the common measures in minimax problems. Please refer to Section 3 for details.

which is, to our knowledge, the first result in minimax problems.

- We extend our main theorems into various algorithms such as ESP, GDA, SGDA. We gain $O(1/n^2)$ bounds with further assumptions that the optimal population risk is small. To our best knowledge, it is the first time to gain $O(1/n^2)$ for PL-SC minimax problems with expectation version and the first excess primal risk bounds for $O(1/n^2)$ with high probability for SC-SC settings.

This paper is organized as follows. In Section 2, we review the related work. In Section 3, we introduce the notations and assumptions about the problems. Section 4 presents our improved generalization error bounds. Then we apply our main theorems into various algorithms and give sharper bounds for different settings in Section 5. Section 6 concludes our paper. All the proofs in our paper are given in Appendix.

2 Related Work

Minimax optimization. Minimax optimization analysis has been widely studied in different settings. For example, one of the most popular SGDA algorithm and its variants have been analyzed in several recent works including [Palaniappan and Bach, 2016; Hsieh *et al.*, 2019] for SC-SC cases, [Nedić and Ozdaglar, 2009; Nemirovski *et al.*, 2009] for convex-concave (C-C) cases, [Lin *et al.*, 2020; Luo *et al.*, 2020; Yan *et al.*, 2020; Rafique *et al.*, 2022] for NC-SC problems, [Thekumparampil *et al.*, 2019; Yan *et al.*, 2020] for nonconvex-concave (NC-C) cases and [Loizou *et al.*, 2020; Liu *et al.*, 2021; Yang *et al.*, 2020] for nonconvex-nonconcave (NC-NC) minimax optimization problems. All these works focus on the iteration complexity (or the gradient complexity) of the algorithms, which only proved the optimization error bounds for the sum of T iteration's gradient of primal empirical function in expectation. Recently [Li and Liu, 2021a; Lei *et al.*, 2021] gave optimization bounds with high probability for Primal-Dual risk. We notice that the optimization error of the gradients of primal functions with high probability haven't been studied yet.

Algorithmic stability. Algorithmic stability is a classical approach in generalization analysis, which was presented by [Rogers and Wagner, 1978]. 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 [Bousquet *et al.*, 2020; Feldman and Vondrak, 2018; Feldman and Vondrak, 2019; 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], collective stability [London *et al.*, 2016] have been developed. For minimax problems, many useful stability measures have also been extended, for example, weak stability [Lei *et al.*, 2021], argument stability [Lei *et al.*, 2021; Li and Liu, 2021a], and uniform stability [Lei *et al.*, 2021; Li and Liu, 2021a; Zhang *et al.*, 2021; Farnia and Ozdaglar, 2021; Ozdaglar *et al.*, 2022]. Most of them focused on the expectation generalization bounds and only [Lei *et al.*, 2021; Li and Liu, 2021a] established some high probability bounds.

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 $O(\sqrt{d/n})$ depended on the sample size n and the dimension of parameters d in high probability. For nonconvex settings, [Mei *et al.*, 2018] showed that the empirical of generalization error is $O(\sqrt{d/n})$. [Xu and Zeevi, 2020] developed a novel "uniform localized convergence" framework using generic chaining for the minimization problems and [Li and Liu, 2021b] extended it to analyze stochastic algorithms. In minimax problems, [Zhang *et al.*, 2022] established the first uniform convergence and showed that the empirical generalization error of the gradients for primal functions is $O(\sqrt{d/n})$ under NC-SC settings.

3 Preliminaries

Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \mathbb{R}^{d'}$ be two nonempty closed convex parameters spaces. Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z}$. We consider the following minimax optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) := \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{x},\mathbf{y};\mathbf{z})], \tag{1}$$

where $f : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is continuously differentiable and Lipschitz smooth jointly in $\mathbf{x}$ and $\mathbf{y}$ for any $\mathbf{z}$. This above minimax objective called as the population minimax problem represents an expectation of a cost function $f(\mathbf{x},\mathbf{y};\mathbf{z})$ for minimization variable $\mathbf{x}$, maximization variable $\mathbf{y}$ and data variable $\mathbf{z}$. In this paper, we focus on the NC-SC problem which means that f is nonconvex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$. Obviously, our goal is to gain the optimal solution $(\mathbf{x}^*,\mathbf{y}^*)$ to (1). Since the distribution $\mathbb{P}$ is unavailable, we can only gain a dataset $S = \{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ drawn n times independently from $\mathbb{P}$. Therefore, we solve the following empirical minimax problem instead

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F_S(\mathbf{x},\mathbf{y}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x},\mathbf{y};\mathbf{z}_i). \tag{2}$$

Next we introduce one of the common measures in minimax problems called primal functions.

Definition 1 (Primal (empirical/population) function)**.** *The primal empirical function and the primal population function are given by*

$$\Phi_S(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} F_S(\mathbf{x},\mathbf{y}) \quad \textit{and} \quad \Phi(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}).$$

Since F_S and F are nonconvex in $\mathbf{x}$, it is difficult to find the global optimal solution in general. In practice, we design an algorithm $\mathcal{A}$ that finds an ϵ-stationary point

$$\|\nabla\Phi(\mathcal{A}_{\mathbf{x}}(S))\| \leq \epsilon, \tag{3}$$

where $\mathcal{A}_{\mathbf{x}}(S)$ is the x-component of the output using any algorithm $\mathcal{A}(S) = (\mathcal{A}_{\mathbf{x}}(S), \mathcal{A}_{\mathbf{y}}(S))$ for solving (2). Then the optimization error for solve the population minimax problem (1) can be decomposed into two terms:

$$\begin{aligned}\|\nabla\Phi(\mathcal{A}_{\mathbf{x}}(S))\| \leq &\|\nabla\Phi_S(\mathcal{A}_{\mathbf{x}}(S))\| \\ &+ \|\nabla\Phi(\mathcal{A}_{\mathbf{x}}(S)) - \nabla\Phi_S(\mathcal{A}_{\mathbf{x}}(S))\|,\end{aligned}$$

where the first term on the right-hand-side corresponds to the optimization error of solving the empirical minimax problem (2) and the second term corresponds to the generalization error of the gradients for primal functions. The above inequality satisfies from the triangle inequality.

Let $\|\cdot\|$ be the Euclidean norm for simplicity and $B(\mathbf{x}_0, R) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}_0\| \leq R\}$ denotes a ball with center $\mathbf{x}_0 \in \mathbb{R}^d$ and radius R. For the closed convex set $\mathcal{X}$, we assume that there is a radius R_1 such that $\mathcal{X} \in B(\mathbf{x}^*, R_1)$. Let $\mathcal{A}(S) := (\mathcal{A}_{\mathbf{x}}(S), \mathcal{A}_{\mathbf{y}}(S))$ denote the output of an algorithm $\mathcal{A}$ for solving the empirical minimax problem (2) with dataset S and $\nabla f = (\nabla_{\mathbf{x}} f, \nabla_{\mathbf{y}} f)$ denote the gradient of a function f.

Definition 2 (Strongly convex function)**.** *Let $\mu_{\mathbf{y}} > 0$. A differentiable function $g : \mathcal{W} \to \mathbb{R}$ is called μ-strongly-convex in $\mathbf{w}$ if the following inequality holds for every $\mathbf{w}_1$, $\mathbf{w}_2$:*

$$g(\mathbf{w}_1) - g(\mathbf{w}_2) \geq \langle \nabla g(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle + \frac{\mu}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|^2,$$

we say g is μ-strongly-concave if $-g$ is μ-strongly-convex.

Definition 3 (Smooth function)**.** *Let $\beta > 0$. A function $f : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is β-smooth in $(\mathbf{x}, \mathbf{y})$ if the function is continuous differentiable and for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ and $\mathbf{z} \in \mathcal{Z}$, $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ satisfies*

$$\left\| \begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}_1; \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}_2; \mathbf{z}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1; \mathbf{z}) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2; \mathbf{z}) \end{pmatrix} \right\| \leq \beta \left\| \begin{pmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{pmatrix} \right\|.$$

Assumption 1 (Nonconvex-strongly-concave minimax problem)**.** *In order to obtain meaningful conclusions, we make the following assumptions:*

- *Let $\mu_{\mathbf{y}} > 0$. The function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is $\mu_{\mathbf{y}}$-strongly concave in $\mathbf{y} \in \mathcal{Y}$ for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$.*
- *The function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is β-smooth in $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ for any $\mathbf{z}$.*
- *$\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets, which means that there exist constants $D_{\mathcal{X}}, D_{\mathcal{Y}} > 0$ such that for any $\mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|^2 \leq D_{\mathcal{X}}$ and for any $\mathbf{y} \in \mathcal{Y}$, $\|\mathbf{y}\|^2 \leq D_{\mathcal{Y}}$.*

The first two assumptions in Assumption 1 are standard in NC-SC minimax problems [Zhang *et al.*, 2021; Farnia and Ozdaglar, 2021; Lei *et al.*, 2021; Li and Liu, 2021a] and the last one in Assumption 1 is widely used in uniform convergence analysis [Kleywegt *et al.*, 2002; Davis and Drusvyatskiy, 2022; Zhang *et al.*, 2022].

Assumption 2 (Lipschitz continuity)**.** *Let $L > 0$, assume that for any $\mathbf{x} \in \mathcal{X}$ and any $\mathbf{y} \in \mathcal{Y}$ respectively for any $\mathbf{z}$, the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ satisfies*

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\| \leq L \quad \text{and} \quad \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\| \leq L.$$

Lipschitz assumption is also the standard assumption and widely used in literature such as [Zhang *et al.*, 2021; Farnia and Ozdaglar, 2021; Lei *et al.*, 2021; Li and Liu, 2021a]. But we need to emphasize that our main Theorem 1 and Theorem 3 do not require the Lipschitz assumption. Instead, we introduce a weaker assumption called Bernstein condition in minimax problems.

Definition 4 (Bernstein condition)**.** *Given a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[X^2] - \mu^2$, we say that Bernstein's condition holds if there exists $B > 0$ such that for all $k \geq 2, k \in \mathbb{N}$,*

$$\left|\mathbb{E}\left[(X - \mu)^k\right]\right| \leq \frac{k!}{2}\sigma^2 B^{k-2}.$$

Remark 1. Bernstein condition has been widely used to obtain tail bounds that may be tighter than the Hoeffding bounds. It is easy to verify that any bounded variable satisfies Bernstein condition. Next, we introduce a straightforward generalization of Bernstein condition to minimax problems. We formally state these extension in the following assumptions.

Assumption 3. *In minimax problems, the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ satisfies Bernstein condition in $\mathbf{x}^*$ for $\mathbf{y}^*$: there exists $B_{\mathbf{x}^*} > 0$ such that for all $k \geq 2, k \in \mathbb{N}$,*

$$\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^k\right] \leq \frac{k!}{2}\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2\right] B_{\mathbf{x}^*}^{k-2}.$$

And the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ satisfies Bernstein condition in $\mathbf{y}^*(\mathbf{x})$ for any fixed $\mathbf{x}$: there exists $B_{\mathbf{y}^*} > 0$ such that for all $k \geq 2, k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^k\right] \leq \frac{k!}{2}\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2\right] B_{\mathbf{y}^*}^{k-2},$$

Remark 2. We can easily obtain that Assumption 2 can derive Assumption 3. For example, if function f is L-Lipschitz continuous, then $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\| \leq L$. Thus for any $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$ and for all $k \geq 2, k \in \mathbb{N}$, we have $\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^k\right] \leq \frac{k!}{2}\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^2\right] L^{k-2}$, which means that the function f satisfies Bernstein condition for any $\mathbf{x}, \mathbf{y}$. Similarly, $\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^k\right] \leq \frac{k!}{2}\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^2\right] L^{k-2}$ can be easily derived. Moreover, Bernstein condition is milder than the bounded assumption of random variables and is also satisfied by various unbounded variables. For example, a random variable is sub-exponetial if it satisfies Bernstein condition [Wainwright, 2019]. Please refer to [Wainwright, 2019] for more discussions. Furthermore, Bernstein condition assumption is pretty mild since $B_{\mathbf{y}^*}$ and $B_{\mathbf{x}^*}$ only depends on gradients at $(\mathbf{x}^*, \mathbf{y}^*)$.

4 Uniform Localized Convergence and Generalization Bounds

Uniform convergence of the gradients for primal functions measures the deviation between the gradients of the primal population function $\nabla\Phi(\mathbf{x})$ and the gradients of the primal empirical function $\nabla\Phi_S(\mathbf{x})$. In this section, we provide the sharper uniform convergence of the gradients for primal functions comparing with [Zhang *et al.*, 2022].

Theorem 1. *Under Assumption 1 and 3, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds for all $\mathbf{x} \in \mathcal{X}$ that*

$$\begin{aligned}&\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\| \\ \le & \frac{\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}\|\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2 \log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{8}{\delta}}{n}\right) \\ &+ \sqrt{\frac{2\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2 \log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} \\ &+ \frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\frac{(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\} \times \\ &\left(\sqrt{\frac{d+\log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}} + \frac{d+\log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}\right),\end{aligned}$$

where C is a absolute constant.

There is only one uniform convergence of gradients for primal functions in minimax problems given in [Zhang *et al.*, 2022]. Here is their main theorem in NC-SC settings.

Theorem 2 (Theorem in [Zhang *et al.*, 2022])**.** *Under Assumption 1 and 2, we have*

$$\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\|\right] = \tilde{O}\left(\frac{L(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\sqrt{\frac{d}{n}}\right),$$

where $\tilde{O}(\cdot)$ hides logarithmic factors.

Remark 3. We now compare our uniform convergence of gradient for primal functions with [Zhang *et al.*, 2022]. Firstly, our result is the only one with high-probability format. Besides, we successfully relax the assumptions. Theorem 2 requires the Lipschitz continuity assumption, while our result only needs Bernstein condition assumption. Please refer to Remark 1 Remark 2 for the detailed comparison between these assumptions. Then, the factor in Theorem 2 is $\frac{L(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}$, while our result in Theorem 1 is $\frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\frac{(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\}$, not involving the term L, which may be very large and even infinite without Lipschitz continuity assumption. Finally, while [Zhang *et al.*, 2022] studied the worst-case upper bounds on the parameters, results based on generic chaining yield upper bound related to the parameters. As shown Theorem 1, we have the term $\max\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\}$ before the term $O(\sqrt{d/n})$, indicating that our results improve as the calculated parameters of algorithms approach the optimal solution. For example, when $\|\mathbf{x}-\mathbf{x}^*\| = O\left(\frac{1}{\sqrt{n}}\right)$, our result is $O(\sqrt{d}/n)$. In some optimal scenario, when $\|\mathbf{x}-\mathbf{x}^*\| \le \frac{1}{n}$, we can attain the best results.

Remark 4. In fact, [Zhang *et al.*, 2022] derived an expectation generalization error for primal functions in minimax problems using complexity. Naturally, we want to use local methods to introduce variance information and obtain a tighter upper bound. A straightforward idea is that we can continue with the traditional localized approach and solve the problem with covering numbers [Bartlett *et al.*, 2002]. However, these technologies require additional bounded assumptions (Assumption 2), or need certain distributional assumptions for unbounded condition. For example, [Mei *et al.*, 2018] introduced the "Hessian statistical noise" assumption when using covering numbers. Fortunately, [Xu and Zeevi, 2020] developed a novel "uniform localized convergence" framework using generic chaining for the minimization problems and [Li and Liu, 2021b] extended it to analyze stochastic algorithms.

This novel framework can not only relax the bounded (or specific distribution) assumptions but also impose fewer restrictions on the surrogate function for the localized method, enabling us to design the measurement functional to achieve a sharper bound. Consequently, we introduce this remarkable framework into minimax problems. Our generalization bound in Theorem 1 uses weaker assumptions comparing with [Zhang *et al.*, 2022] and is sharper in some conditions due to our utilization of variance information.

Introducing this new framework into minimax problems is not straightforward. [Zhang *et al.*, 2022] indeed established a connection between inner and outer layers with the loss of primal functions, but we need do this with a new generic chaining approach. Furthermore, it is noteworthy that the optimal point $\mathbf{y}^*(\mathbf{x}) := \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y})$ for a given $\mathbf{x}$ differs from $\mathbf{y}^*_S(\mathbf{x}) := \arg\max_{\mathbf{y}\in\mathcal{Y}} F_S(\mathbf{x},\mathbf{y})$, thus introducing an additional error term $\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*_S(\mathbf{x})\|$. Compared to [Zhang *et al.*, 2022], they only need to bound this term with $O(1/\sqrt{n})$. But we need to reach the upper bound to $O(1/n)$ under certain assumptions.

Next, we provide a dimension-free uniform convergence of gradients for primal functions when the PL condition is satisfied. Firstly, we introduce the extension of the PL condition to minimax problems used in [Guo *et al.*, 2020; Yang *et al.*, 2020].

Assumption 4 ($\mathbf{x}$-side $\mu_{\mathbf{x}}$-Polyak-Lojasiewicz condition)**.** *For any $\mathbf{y} \in \mathcal{Y}$, the function $F(\mathbf{x},\mathbf{y})$ satisfies the $\mathbf{x}$-side $\mu_{\mathbf{x}}$-Polyak-Lojasiewicz (PL) condition with parameter $\mu_{\mathbf{x}} > 0$ on all $\mathbf{x} \in \mathcal{X}$ if*

$$F(\mathbf{x},\mathbf{y}) - \inf_{\mathbf{x}'} F(\mathbf{x}',\mathbf{y}) \le \frac{1}{2\mu_{\mathbf{x}}}\|\nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{y})\|^2.$$

Remark 5. Numerous studies have been conducted on deep learning to provide evidence for the validity of the PL condition in risk minimization problems. This condition has been demonstrated to hold either globally or locally in certain networks with specific structural, activation, or loss function characteristics [Hardt and Ma, 2016; Li and Liang, 2018; Arora *et al.*, 2018; Charles and Papailiopoulos, 2018; Du *et al.*, 2018; Allen-Zhu *et al.*, 2019]. For instance, [Du *et al.*, 2018] has exhibited that if a two-layer neural network possesses a sufficiently wide width, the PL condition is upheld within a ball centered at the initial solution, and the global optimum is situated within this same ball. Additionally, [Allen-Zhu *et al.*, 2019] has further demonstrated that in overparameterized deep neural networks utilizing ReLU activation, the PL condition is applicable to a global optimum located in the vicinity of a random initial solution.

Theorem 3. *Under Assumption 1 and 3, assume that the population risk $F(\mathbf{x},\mathbf{y})$ satisfies Assumption 4 with parameter $\mu_{\mathbf{x}}$ and let $c = \max\{16C^2, 1\}$. For any $\delta \in (0,1)$*

when $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^4(d+\log\frac{16\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^4\mu_{\mathbf{x}}^2}$*, with probability at least* $1-\delta$*, it holds for all* $\mathbf{x}\in\mathcal{X}$ *that*

$$\begin{aligned}&\|\nabla\Phi(\mathbf{x})-\nabla\Phi_S(\mathbf{x})\|\\ \leq&\|\nabla\Phi_S(\mathbf{x})\|+2\sqrt{\frac{2\mathbb{E}\left[\|\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]\log\frac{8}{\delta}}{n}}+\frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n}\\ &+\frac{\mu_{\mathbf{x}}}{n}+\frac{2\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}\left[\|\nabla_{\mathbf{y}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]\log\frac{8}{\delta}}{n}}+\frac{B_{\mathbf{y}^*}\log\frac{8}{\delta}}{n}\right).\end{aligned}$$

Remark 6. The following inequality can be easily derived using triangle inequality and Cauchy–Bunyakovsky–Schwarz inequality.

$$\begin{aligned}&\Phi(\mathbf{x})-\Phi(\mathbf{x}^*)\\ \leq&\frac{8\|\nabla\Phi_S(\mathbf{x})\|^2}{\mu_{\mathbf{x}}}+\frac{16\mathbb{E}\left[\|\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]\log\frac{8}{\delta}}{\mu_{\mathbf{x}}n}\\ &+\frac{16\beta^2\mathbb{E}\left[\|\nabla_{\mathbf{y}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n}\\ &+\frac{2\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta}+2B_{\mathbf{x}^*}\log\frac{8}{\delta}+\mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}}n^2}.\end{aligned}\tag{4}$$

We can easily derive (4) from Theorem 3 to gain the excess primal risk bound, where $\|\nabla\Phi_S(\mathbf{x})\|$ is the empirical optimization error of the primal function. In Theorem 3 and (4), $\|\nabla\Phi_S(\mathbf{x})\|$ can be very tiny since most famous optimization algorithms such as GDA and SGDA, can optimize it small enough. The term $\mathbb{E}\left[\|\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]$ and $\mathbb{E}\left[\|\nabla_{\mathbf{y}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]$ can be also tiny since they only depend on the the gradient of the optima $(\mathbf{x}^*,\mathbf{y}^*)$ w.r.t $\mathbf{x}$ and $\mathbf{y}$. We further analyze these two terms $\mathbb{E}\left[\|\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]$ and $\mathbb{E}\left[\|\nabla_{\mathbf{y}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]$ using "self-bounding" property for smooth functions [Srebro *et al.*, 2010] and considering specific algorithms in Section 5, which can derive to $O(1/n^2)$ bounds. Thus, comparing with Theorem 2 in [Zhang *et al.*, 2022], this uniform localized convergence bound is clearly tighter when relaxing Lipschitz continuity (Assumption 2) and considering PL condition (Assumption 4). Additionally, uniform convergence often implies results with a square-root dependence on the dimension d such as Theorem 1 and [Zhang *et al.*, 2022]. Another distinctive improvement of Theorem 3 is that we remove the dimension d when the population risk $F(\mathbf{x},\mathbf{y})$ satisfies the $\mathbf{x}$-side PL condition and the sample size n is large enough.

5 Application

5.1 Empirical Saddle Point

Empirical saddle point (ESP) algorithm, which is also known as sample average approximation (SAA) [Zhang *et al.*, 2021] refers to (2). We denote $(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*)$ as one of the ESP solution to (2). Then we can provide some important theorems in this subsection.

Theorem 4. *Suppose the empirical saddle point* $(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*)$ *exists and Assumption 1 and 3 hold, for any* $\delta\in(0,1)$*, with probability at least* $1-\delta$*, we have*

$$\|\nabla\Phi(\hat{\mathbf{x}}^*)\|=O\left(\sqrt{\frac{d+\log\frac{\log n}{\delta}}{n}}\right)$$

Remark 7. When Assumption 1 and 3 hold, Theorem 4 shows that the *population* optimization error $\|\nabla\Phi(\hat{\mathbf{x}}^*)\|$ is $O\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}\right)$ ($\log n$ is small and can be ignored typically). Note that this result doesn't require the Lipschitz continuity assumption (Assumption 2). Although it may be hard to find $(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*)$ in NC-SC minimax problems, it is still meaningful when assuming the ESP $(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*)$ has been found.

Theorem 5. *Suppose Assumption 1 and 3 hold. Assume that the population risk* $F(\mathbf{x},\mathbf{y})$ *satisfies Assumption 4 with parameter* $\mu_{\mathbf{x}}$*. For any* $\delta\in(0,1)$*, with probability at least* $1-\delta$*, when* $n\geq\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^4(d+\log\frac{16\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^4\mu_{\mathbf{x}}^2}$*, where* c *is an absolute constant, we have*

$$\begin{aligned}\Phi(\hat{\mathbf{x}}^*)-\Phi(\mathbf{x}^*)\leq&\frac{12\beta^2\mathbb{E}\|\nabla_{\mathbf{y}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n}\\ &+\frac{12\mathbb{E}\|\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\log\frac{8}{\delta}}{\mu_{\mathbf{x}}n}\\ &+\frac{3\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta}+2B_{\mathbf{x}^*}\log\frac{8}{\delta}+\mu_{\mathbf{x}}\right)^2}{2\mu_{\mathbf{x}}n^2}.\end{aligned}$$

Furthermore, if we assume the function $f(\mathbf{x},\mathbf{y};\mathbf{z})$ *is non-negative, we have*

$$\Phi(\hat{\mathbf{x}}^*)-\Phi(\mathbf{x}^*)=O\left(\frac{\Phi(\mathbf{x}^*)\log\frac{1}{\delta}}{n}+\frac{\log^2\frac{1}{\delta}}{n^2}\right).$$

When $\Phi(\mathbf{x}^*)=O\left(\frac{1}{n}\right)$*, we have*

$$\Phi(\hat{\mathbf{x}}^*)-\Phi(\mathbf{x}^*)=O\left(\frac{\log^2\frac{1}{\delta}}{n^2}\right).$$

Remark 8. Theorem 5 shows that when the population minimax risk $F(\mathbf{x},\mathbf{y})$ satisfies $\mathbf{x}$-side PL condition, we can provide a sharper excess risk bound for primal function, which can be $O(1/n^2)$. Note that the optimal population primal function $\Phi(\mathbf{x}^*)=O(1/n)$ is a very common assumption in many researches such as [Srebro *et al.*, 2010; Zhang *et al.*, 2017; Liu *et al.*, 2018; Zhang and Zhou, 2019; Lei and Ying, 2020], which is natural because $F(\mathbf{x}^*,\mathbf{y}^*)$ is the minimal population risk. Now we compare our results with recent related work [Li and Liu, 2021b], which studied the general machine learning settings for $f(\mathbf{w})$ under PL condition. Their empirical risk minimizer (ERM) excess risk bound provided $O\left(1/n^2\right)$ order rates. We analyze the excess risk with primal functions in minimax problems and our result for ESP is $O\left(1/n^2\right)$, which is the same order as theirs.

5.2 Gradient Descent Ascent

Gradient descent ascent (GDA) presented in Algorithm 1 is one of the most popular algorithms and has been widely used

Algorithm 1 Two-timescale GDA for minimax problem

Input: $(\mathbf{x}_1, \mathbf{y}_1) = (\mathbf{0}, \mathbf{0})$, step sizes $\eta_{\mathbf{x}} > 0, \eta_{\mathbf{y}} > 0$ and dataset $S = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$
for $t = 1, \dots, T$ **do**
  update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)$
  update $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta_{\mathbf{y}} \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)$

Algorithm 2 Two-timescale SGDA for minimax problem

Input: $(\mathbf{x}_1, \mathbf{y}_1) = (\mathbf{0}, \mathbf{0})$, step sizes $\{\eta_{\mathbf{x}_t}\}_t > 0$, $\{\eta_{\mathbf{y}_t}\}_t > 0$ and dataset $S = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$
for $t = 1, \dots, T$ **do**
  update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_{\mathbf{x}_t} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})$
  update $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta_{\mathbf{y}_t} \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})$

in minimax problems. In this subsection, we provide the population optimization error bound and the excess risk bounds of primal functions with the two-timescale GDA algorithm which is harder to analyze compared to GDMax and multi-step GDA [Lin *et al.*, 2020].

Theorem 6. *Suppose Assumption 1 and 2 hold. Let $\{\mathbf{x}_t\}_t$ be the sequence produced by Algorithm 1 with the step sizes chosen as $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ and $\eta_{\mathbf{y}} = \frac{1}{\beta}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \Phi(\mathbf{x}_t)\|^2 \leq O\left(\frac{1}{T}\right) + O\left(\frac{d + \log \frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n} T\right).$$

Furthermore, when $T \asymp O\left(\sqrt{\frac{n}{d}}\right)$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \Phi(\mathbf{x}_t)\|^2 \leq O\left(\frac{d + \log \frac{\log n}{\delta}}{\sqrt{nd}}\right).$$

Remark 9. Theorem 6 also gives the *population* optimization error which reveals that we need to balance the empirical optimization error and the generalization error for GDA. According to the results, the iterative complexity of Algorithm 1 should be chosen as $T \asymp O\left(\sqrt{\frac{n}{d}}\right)$, which achieves the optimal population optimization error of primal function.

In comparison to Theorem 4, Theorem 6 derives into population optimization error w.r.t GDA, which is much more difficult. To establish population optimization error, we need to bound the empirical optimization error, an area where no research has been conducted in NC-SC settings with high probability. One possible approach is to construct the martingale difference sequence of step T for primal functions, yet this constitutes a separate topic warranting further exploration. Theorem 6 aims to directly apply Theorem 1 to GDA. Comparing with Theorem 3, Theorem 6 only necessitates smooth and Lipschitz conditions (Assumption 1 and 2) and doesn't require PL conditions.

Next, we provide the excess risk of primal functions $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$ for Algorithm 1, where $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$. We need to know the empirical optimization error $\|\nabla \Phi_S(\bar{\mathbf{x}}_T)\|$ firstly.

Unfortunately, although the generalization bounds we proved are in NC-SC settings, we require the SC-SC assumptions to derive the empirical optimization error bound of primal functions, to gain the high probability bound. We relax this SC-SC assumption in Appendix E using existing optimization error bound with expectation format.

Definition 5. *A function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is $\mu_{\mathbf{x}}$-strongly-convex-$\mu_{\mathbf{y}}$-strongly-concave if $g(\cdot, \mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly-convex for any $\mathbf{y} \in \mathcal{Y}$ and $g(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly-concave for any $\mathbf{x} \in \mathcal{X}$.*

Assumption 5 (Strongly-convex-strongly-concave minimax problem)**.** *Assume Assumption 1 holds and let $\mu_{\mathbf{x}} > 0, \mu_{\mathbf{y}} > 0$. The function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is $\mu_{\mathbf{x}}$-strongly-convex-$\mu_{\mathbf{y}}$-strongly-concave in $\mathbf{y} \in \mathcal{Y}$ for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$.*

Remark 10. Assumption 5 is commonly used in SC-SC problems [Zhang *et al.*, 2021; Li and Liu, 2021a]. We require this assumption to derive the empirical **optimization error** bound of primal functions. The detailed proofs of the optimization error bound $\|\nabla \Phi_S(\bar{\mathbf{x}}_T)\|$ are given in Section D.2 for GDA and in Section D.3 for SGDA.

Theorem 7. *Suppose Assumption 3 and 5 hold. Let $\{\mathbf{x}_t\}_t$ be the sequence produced by Algorithm 1 and $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$ with the step sizes chosen as $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ and $\eta_{\mathbf{y}} = \frac{1}{\beta}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, when $T \asymp n$ and $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^4(d+\log\frac{16\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^4 \mu_{\mathbf{x}}^2}$, where c is an absolute constant, we have*

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\Bigg(\frac{\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2\right] \log\frac{1}{\delta}}{n} + \frac{\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2\right] \log\frac{1}{\delta}}{n} + \frac{\log^2\frac{1}{\delta}}{n^2}\Bigg).$$

Furthermore, Let $T \asymp n^2$. Assume the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is non-negative, we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\Phi(\mathbf{x}^*)\log\frac{1}{\delta}}{n} + \frac{\log^2\frac{1}{\delta}}{n^2}\right).$$

When $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\log^2\frac{1}{\delta}}{n^2}\right).$$

Remark 11. Theorem 7 shows that the excess risk for primal functions can be bound by $O\left(1/n^2\right)$ comparing with the optimal result $O(1/n)$ given in [Li and Liu, 2021a] when n is large enough. Note that we require the SC-SC assumption to derive the empirical optimization error. If we give this bound in expectation, we can relax the SC-SC assumption with $\mathbf{x}$-side PL-strongly-concave assumption instead.

Reference	Algorithm	Assumption	Measure	Bounds
[Lei *et al.*, 2021]	SGDA	C-SC, Lip, S	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(1/\sqrt{n})$
		C-SC, Lip, S	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(\log n/\sqrt{n})$
	AGDA	PL-SC, Lip, S	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O\left(n^{-\frac{c\beta+1}{2c\beta+1}}\right)$
[Li and Liu, 2021a]	ESP	SC-SC, Lip, S, LN	(HP.) $\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*)$	$O(\log n/n)$
	GDA	SC-SC, Lip, S, LN	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(\log n/n)$
	SGDA	SC-SC, Lip, S, LN	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(\log n/n)$
This work	ESP	PL-SC, B, S, LN	(HP.) $\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*)$	$O(1/n^2)$
	GDA	PL-SC, B, S, LN	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(1/n^2)$
		SC-SC, B, S, LN	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(1/n^2)$
	SGDA	PL-SC, B, S, LN	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(1/n^2)$
		SC-SC, Lip, S, LN	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(1/n^2)$
	AGDA	PL-SC, B, S, LN	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(1/n^2)$

Table 1: Summary of the results. These bounds are established by choosing optimal iterate number T.

5.3 Stochastic Gradient Descent Ascent

We now analyze the excess risk bound of primal functions for stochastic gradient descent ascent (SGDA). The algorithmic scheme that we study is two-timescale SGDA ($\eta_{\mathbf{x}} \neq \eta_{\mathbf{y}}$) with variable step sizes, presented in Algorithm 2 which is more nature in real problems.

Theorem 8. *Suppose Assumption 2 and 5 hold, let $\{\mathbf{x}_t\}_t$ be the sequence produced by Algorithm 2 and $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{x}_t$ with the step sizes chosen as $\eta_{\mathbf{x}_t} = \frac{1}{\mu_{\mathbf{x}}(t+t_0)}$ and $\eta_{\mathbf{y}_t} = \frac{1}{\mu_{\mathbf{y}}(t+t_0)}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, when $T \asymp n^2$ and $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^4(d+\log\frac{16\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^4\mu_{\mathbf{x}}^2}$, where c is an absolute constant, we have*

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\Bigg(\frac{\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]\log\frac{1}{\delta}}{n} + \frac{\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2\right]\log\frac{1}{\delta}}{n} + \frac{\log^2\frac{1}{\delta}}{n^2}\Bigg).$$

Furthermore, Let $T \asymp n^4$. Assume the function $f(\mathbf{x},\mathbf{y};\mathbf{z})$ is non-negative, we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\Phi(\mathbf{x}^*)\log\frac{1}{\delta}}{n} + \frac{\log^2\frac{1}{\delta}}{n^2}\right).$$

When $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\log^2\frac{1}{\delta}}{n^2}\right).$$

Remark 12. Theorem 8 reveals that under the SC-SC settings, the excess risk bound can be $O(1/n^2)$ comparing with the optimal result $O(1/n)$ given in [Li and Liu, 2021a]. Similarly, since the SC-SC assumption is required to derive the empirical optimization bound, we can relax the assumptions when we only need expectation bounds instead of high probability bounds.

Table 1 gives the summary of existing results. AGDA is alternating gradient descent ascent algorithm proposed in [Yang *et al.*, 2020]. Lip means Lipschitz continuity. S means smoothness. B means Bernstein condition. LN means low noise condition. PL-SC means $\mathbf{x}$-side PL condition strongly concave settings. E. means expectation results. HP. means high probability results. Since most of existing works on optimization errors are the expectation format, and our high probability results of generalization error can be transformed into the expectation results. so we give the proofs of the expectation result to relax some assumptions such as SC-SC condition in Appendix E.

6 Conclusion

In this paper, we provide the improved generalization bounds for minimax problems with uniform localized convergence. We firstly provide a sharper bound measured by the gradients of primal functions with weaker assumptions in NC-SC settings. Then we provide dimension-independent results under PL-SC condition. Finally we extend our main theorems into various algorithms to reach the optimal excess primal risk bounds. Our excess primal risk bounds are $O(1/n^2)$ in SC-SC settings with high probability format and $O(1/n^2)$ in PL-SC settings with expectation version. We notice that most optimization works focused on the gradient complexity with expectation results. It would be interesting to give the optimization error of $\bar{\mathbf{x}}_T$ or even $\mathbf{x}_T$ with high probability under weaker conditions (for example in PL-SC and even NC-SC settings). Combining with our generalization work, we can get a tighter excess primal risk bound with weaker conditions.

Ethical Statement

Here are no ethical issues.

Acknowledgments

We thank the anonymous reviewers for their valuable and constructive suggestions and comments. This work is supported by the Beijing Natural Science Foundation

(No.4222029); the National Natural Science Foundation of China (N0.62076234); the National Key Research and Development Project (No.2022YFB2703102); the "Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China"; the Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098); the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.2021030199), the Huawei-Renmin University joint program on Information Retrieval: the Unicom Innovation Ecological Cooperation Plan; the CCF-Huawei Populus Grove Fund; and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.23XNH030).

References

[Allen-Zhu *et al.*, 2019] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019.

[Arora *et al.*, 2018] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[Bartlett *et al.*, 2002] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *International Conference on Computational Learning Theory*, pages 44–58. Springer, 2002.

[Bassily *et al.*, 2020] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *NeurIPS*, 2020.

[Boucheron *et al.*, 2013] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

[Bousquet *et al.*, 2020] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, 2020.

[Charles and Papailiopoulos, 2018] Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *ICML*, 2018.

[Chen *et al.*, 2017] Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. *NeurIPS*, 2017.

[Dai *et al.*, 2018] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *ICML*, 2018.

[Davis and Drusvyatskiy, 2022] Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *Mathematics of Operations Research*, 47(1):209–231, 2022.

[Dirksen, 2015] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1–29, 2015.

[Du *et al.*, 2017] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *ICML*, 2017.

[Du *et al.*, 2018] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[Farnia and Ozdaglar, 2021] Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *ICML*, 2021.

[Feldman and Vondrak, 2018] Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *NeurIPS*, 31, 2018.

[Feldman and Vondrak, 2019] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, 2019.

[Fisher, 1922] Ronald 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.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

[Guo *et al.*, 2020] Zhishuai Guo, Yan Yan, Zhuoning Yuan, and Tianbao Yang. Fast objective & duality gap convergence for nonconvex-strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

[Hardt and Ma, 2016] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[Hsieh *et al.*, 2019] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *NeurIPS*, 2019.

[Karimi *et al.*, 2016] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML*, pages 795–811. Springer, 2016.

[Kleywegt *et al.*, 2002] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12(2):479–502, 2002.

[Klochkov and Zhivotovskiy, 2021] Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *NeurIPS*, 2021.

[Kuzborskij and Lampert, 2018] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *ICML*, 2018.

[Lei and Ying, 2020] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *ICML*, 2020.

[Lei *et al.*, 2021] Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *ICML*, 2021.

[Li and Liang, 2018] Yuanzhi Li and Yingyu Liang. Learning over-parameterized neural networks via stochastic gradient descent on structured data. *NeurIPS*, 2018.

[Li and Liu, 2021a] Shaojie Li and Yong Liu. High probability generalization bounds with fast rates for minimax problems. In *ICLR*, 2021.

[Li and Liu, 2021b] Shaojie Li and Yong Liu. Improved learning rates for stochastic optimization: Two theoretical viewpoints. *arXiv preprint arXiv:2107.08686*, 2021.

[Lin *et al.*, 2020] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020.

[Liu *et al.*, 2017] Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *ICML*, 2017.

[Liu *et al.*, 2018] Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, and Tianbao Yang. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *NeurIPS*, 2018.

[Liu *et al.*, 2021] Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *The Journal of Machine Learning Research*, 22(1):7651–7684, 2021.

[Loizou *et al.*, 2020] Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *ICML*, 2020.

[London *et al.*, 2016] Ben London, Bert Huang, and Lise Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

[Luo *et al.*, 2020] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *NeurIPS*, 2020.

[Mei *et al.*, 2018] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[Namkoong and Duchi, 2017] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *NeurIPS*, 2017.

[Nedić and Ozdaglar, 2009] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142:205–228, 2009.

[Nemirovski *et al.*, 2009] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[Nesterov, 2003] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[Nouiehed *et al.*, 2019] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *NeurIPS*, 32, 2019.

[Ozdaglar *et al.*, 2022] Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. What is a good metric to study generalization of minimax learners? *arXiv preprint arXiv:2206.04502*, 2022.

[Palaniappan and Bach, 2016] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *NeurIPS*, 2016.

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

[Rafique *et al.*, 2022] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

[Rogers and Wagner, 1978] William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

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

[Smale and Zhou, 2007] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[Srebro *et al.*, 2010] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

[Talagrand, 1996] Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

[Tarres and Yao, 2014] Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

[Thekumparampil *et al.*, 2019] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *NeurIPS*, 2019.

[Van der Vaart, 2000] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

[Wainwright, 2019] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[Xu and Zeevi, 2020] Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020.

[Yan *et al.*, 2020] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *NeurIPS*, 2020.

[Yang *et al.*, 2020] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *NeurIPS*, 33:1153–1165, 2020.

[Zhang and Zhou, 2019] Lijun Zhang and Zhi-Hua Zhou. Stochastic approximation of smooth and strongly convex functions: Beyond the $o(1/t)$ convergence rate. In *Conference on Learning Theory*, 2019.

[Zhang *et al.*, 2017] Lijun Zhang, Tianbao Yang, and Rong 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*, 2017.

[Zhang *et al.*, 2021] Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR, 2021.

[Zhang *et al.*, 2022] Siqi Zhang, Yifan Hu, Liang Zhang, and Niao He. Uniform convergence and generalization for nonconvex stochastic minimax problems. *arXiv preprint arXiv:2205.14278*, 2022.