Towards Sharper Risk Bounds for Minimax Problems

Bowie 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

\{bowie.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 non-convex-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 mimimax 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 $O(\sqrt{d/n})$ and [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 $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^d$ be two nonempty closed convex parameters spaces. Let $\mathbb{P}$ be a probability measure defined on a sample space $Z$. We consider the following minimax optimization problem

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

(1)

where $f : X \times Y \times Z \to \mathbb{R}$ is continuously differentiable and Lipschitz smooth jointly in $x$ and $y$ for any $z$. This above minimax objective called as the population minimax problem represents an expectation of a cost function $f(x, y; z)$ for minimization variable $x$, maximization variable $y$ and data variable $z$. In this paper, we focus on the NC-SC problem which means that $f$ is nonconvex in $x$ and strongly concave in $y$. Obviously, our goal is to gain the optimal solution $\langle x^*, y^* \rangle$ to (1). Since the distribution $\mathbb{P}$ is unavailable, we can only gain a dataset $S = \{z_1, \ldots, z_n\}$ drawn $n$ times independently from $\mathbb{P}$. Therefore, we solve the following empirical minimax problem instead

$$\min_{x \in X} \max_{y \in Y} F_S(x, y) := \frac{1}{n} \sum_{i=1}^{n} f(x, y; z_i).$$

(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

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

Since $F_S$ and $F$ are nonconvex in $x$, it is difficult to find the global optimal solution in general. In practice, we design
an algorithm $A$ that finds an $\epsilon$-stationary point
\begin{equation}
\|\nabla \Phi(A_x(S))\| \leq \epsilon, \tag{3}
\end{equation}
where $A_x(S)$ is the $x$-component of the output using any algorithm $A(S) = (A_x(S), A_y(S))$ for solving (2). Then the optimization error for solving the population minimax problem (1) can be decomposed into two terms:

$$
\|\nabla \Phi(A_x(S))\| \leq \|\nabla \Phi_S(A_x(S))\| + \|\nabla \Phi(A_x(S)) - \nabla \Phi_S(A_x(S))\|,
$$

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 $B(x_0, R) := \{x \in \mathbb{R}^d : \|x - x_0\| \leq R\}$ denotes a ball with center $x_0 \in \mathbb{R}^d$ and radius $R$. For the closed convex set $X$, we assume that there is a radius $R_1$ such that $X \in B(x^*, R_1)$. Let $A(S) := (A_x(S), A_y(S))$ denote the output of an algorithm $A$ for solving the empirical minimax problem (2) with dataset $S$ and $\nabla f = (\nabla_x f, \nabla_y f)$ denote the gradient of a function $f$.

**Definition 2** (Strongly convex function). Let $\mu_y > 0$. A differentiable function $g : \mathcal{W} \to \mathbb{R}$ is called $\mu_y$-strongly-convex in $w$ if the following inequality holds for every $w_1, w_2$:

$$
g(w_1) - g(w_2) \geq \nabla g(w_2), w_1 - w_2 + \frac{\mu_y}{2} \|w_1 - w_2\|^2,
$$

we say $g$ is $\mu_y$-strongly-convex if $-g$ is $\mu_y$-strongly-convex.

**Definition 3** (Smooth function). Let $\beta > 0$. A function $f : X \times Y \times Z \to \mathbb{R}$ is $\beta$-smooth in $(x, y)$ if the function is continuously differentiable and for any $x_1, x_2 \in X, y_1, y_2 \in Y$ and $z \in Z$, $f(x, y, z)$ satisfies

$$
\left\| \nabla_x f(x_1, y_1, z) - \nabla_x f(x_2, y_2, z) \right\| \leq \beta \left\| (x_1 - x_2) \right\|, \\
\left\| \nabla_y f(x_1, y_1, z) - \nabla_y f(x_2, y_2, z) \right\| \leq \beta \left\| (y_1 - y_2) \right\|.
$$

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

- Let $\mu_y > 0$. The function $f(x, y, z)$ is $\mu_y$-strongly-concave in $y$ for any $x \in X$ and $z \in Z$.
- The function $f(x, y, z)$ is $\beta$-smooth in $(x, y)$ if the function is convex compact sets $X \times Y$ and $z$.
- $X$ and $Y$ are convex compact sets, which means that there exist constants $D_x, D_y > 0$ such that for any $x \in X$, $\|x\|^2 \leq D_x$ for any $y \in Y$, $\|y\|^2 \leq D_y$.

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 $x \in X$ and any $y \in Y$ respectively for any $z$, the function $f(x, y, z)$ satisfies

$$
\|\nabla_x f(x, y, z)\| \leq L \quad \text{and} \quad \|\nabla_y f(x, y, z)\| \leq L.
$$

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 $\mu = E[X]$ and variance $\sigma^2 = E[X^2] - \mu^2$, we say that Bernstein’s condition holds if there exists $B > 0$ such that for all $k \geq 2, k \in \mathbb{N}$,

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

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

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

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

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

**Remark 2.** We can easily obtain that Assumption 2 can derive Assumption 3. For example, if function $f$ is $L$-Lipschitz continuous, then $\|\nabla_x f(x, y, z)\| \leq L$. Thus for any $x \in X, y \in Y$ and for all $k \geq 2, k \in \mathbb{N}$, we have $E\left[\|\nabla_x f(x, y, z)\|^k\right] \leq \frac{k!}{2}E\left[\|\nabla_x f(x, y, z)\|^2\right] L^{k-2}$, which means that the function $f$ satisfies Bernstein condition for any $x, y$. Similarly, $E\left[\|\nabla_y f(x, y, z)\|^k\right] \leq \frac{k!}{2}E\left[\|\nabla_y f(x, y, z)\|^2\right] L^{k-2}$ 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-exponential if it satisfies Bernstein condition [Wainwright, 2019]. Please refer to [Wainwright, 2019] for more discussions. Furthermore, Bernstein condition assumption is pretty mild since $B_{y^*}$ and $B_{x^*}$ only depends on gradients at $(x^*, y^*)$.

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 $\nabla \Phi(x)$ and the gradients of the primal empirical function $\nabla \Phi_S(x)$. 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 \( x \in X \) that

\[
\|\nabla \Phi(x) - \nabla \Phi_S(x)\| \leq \frac{\beta}{\mu_y} \left( \frac{2E[\|\nabla_x f(x^*, y^*; x)\|^2 \log \frac{2}{n}]}{n} + \frac{B_{\gamma} \log \frac{2}{n}}{n} \right) + \sqrt{\frac{2E[\|\nabla_x f(x^*, y^*; x)\|^4 \log^2 \frac{2}{n}]}{n}} + \frac{B_{\gamma} \log^2 \frac{2}{n}}{n} + \frac{C(\mu_y + \beta)(\mu_y + \beta)}{\mu_y} \max \left\{ \|x - x^*\|, \frac{1}{n} \right\} \times \left( \frac{d + \log \frac{16\log_2(\sqrt{\log n} + 1)}{n}}{n} + \frac{d + \log \frac{16\log_2(\sqrt{\log n} + 1)}{n}}{n} \right),
\]

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

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

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 \( \frac{L(\mu_y + \beta)}{\mu_y} \), while our result in Theorem 1 is \( \frac{C(\mu_y + \beta)(\mu_y + \beta)}{\mu_y} \max \left\{ \|x - x^*\|, \frac{1}{n} \right\} \), 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 \( \max\{\|x - x^*\|, \frac{1}{n}\} \) 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 \( \|x - x^*\| = O \left( \frac{1}{\sqrt{n}} \right) \), our result is \( O(\sqrt{d/n}) \). In some optimal scenario, when \( \|x - x^*\| \leq \frac{1}{n} \), 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 \( y^*(x) := \arg\max_{y \in Y} F(x, y) \) for a given \( x \) differs from \( y^*_S(x) := \arg\max_{y \in Y} F_S(x, y) \), thus introducing an additional error term \( \|y^*(x) - y^*_S(x)\| \). 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 (x-side \( \mu_x \)-Polyak-Lojasiewicz condition). For any \( y \in Y \), the function \( F(x, y) \) satisfies the x-side \( \mu_x \)-Polyak-Lojasiewicz (PL) condition with parameter \( \mu_x > 0 \) on all \( x \in X \) if

\[
F(x, y) - \inf_{x'} F(x', y) \leq \frac{1}{2\mu_x} \|\nabla_x F(x, y)\|^2.
\]

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 \( F(x, y) \) satisfies Assumption 4 with parameter \( \mu_x \) and let \( c = \max\{16C^2, 1\} \). For any \( \delta \in (0, 1) \)
when \( n \geq \frac{c^{2} (\mu_{x} + \beta)^{4} (d + \log n)^{\frac{1}{2}}}{\mu_{x}^2 n^2} \), with probability at least \( 1 - \delta \), it holds for all \( x \in X \) that
\[
\| \nabla \Phi(x) - \nabla \Phi_{S}(x) \| \leq 2 \mathbb{E} \left[ \| \nabla f(x^*, y^*; z) \|^{2} \right] \frac{\log \frac{8}{\delta}}{n} + \frac{2B_{x^*} \log \frac{8}{\delta}}{n} \frac{\mu_{x}}{n} + \frac{2B_{x^*} \log \frac{8}{\delta}}{n} \mu_{x} n^{2}.
\]

Remark 6. The following inequality can be easily derived using triangle inequality and Cauchy–Bunyakovsky–Schwarz inequality.
\[
\Phi(x) - \Phi(x^*) \leq 8 \left\| \nabla \Phi_{S}(x) \right\|^{2} + 16 \mathbb{E} \left[ \| \nabla f(x^*, y^*; z) \|^{2} \right] \frac{\log \frac{8}{\delta}}{n} + \frac{16B_{x^*} \log \frac{8}{\delta}}{n} \mu_{x} n^{2} + 2 \left( \frac{2B_{x^*} \log \frac{8}{\delta}}{n} + \frac{2B_{x^*} \log \frac{8}{\delta}}{n} + \frac{\mu_{x}}{n} \right)^{2}.
\]

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

Theorem 5. Suppose Assumption 1 and 3 hold. Assume that the population risk \( F(x, y) \) satisfies Assumption 4 with parameter \( \mu_{x} \). For any \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \), when \( n \geq \frac{c^{2} (\mu_{x} + \beta)^{4} (d + \log n)^{\frac{1}{2}}}{\mu_{x}^2 n^2} \), where \( c \) is an absolute constant, we have
\[
\Phi(x^*) - \Phi(x^*) \leq 12 \mathbb{E} \left[ \| \nabla f(x^*, y^*; z) \|^{2} \right] \frac{\log \frac{8}{\delta}}{n} + \frac{12B_{x^*} \log \frac{8}{\delta}}{n} \mu_{x} n^{2} + 3 \left( \frac{2B_{x^*} \log \frac{8}{\delta}}{n} + \frac{2B_{x^*} \log \frac{8}{\delta}}{n} + \frac{\mu_{x}}{n} \right)^{2}.
\]
Furthermore, if we assume the function \( f(x, y; z) \) is non-negative, we have
\[
\Phi(x^*) - \Phi(x^*) = O \left( \frac{\Phi(x^*) \log \frac{8}{\delta}}{n} + \frac{\log \frac{8}{n}}{n^2} \right).
\]
When \( \Phi(x^*) = O \left( \frac{1}{n} \right) \), we have
\[
\Phi(x^*) - \Phi(x^*) = O \left( \frac{\log \frac{2}{n}}{n^2} \right).
\]

Remark 8. Theorem 5 shows that when the population minimax risk \( F(x, y) \) satisfies 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 \( \Phi(x^*) = O(1/n) \) 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 \( F(x^*, y^*) \) is the minimal population risk. Now we compare our results with recent related work [Li and Liu, 2011b], which studied the general machine learning settings for \( f(w) \) under PL condition. Their empirical risk minimizer (ERM) excess risk bound provided \( O(1/n^2) \) order rates. We analyze the excess risk with primal functions in minimax problems and our result for ESP is \( O(1/n^2) \), which is the same order as theirs.

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 \((x^*, y^*)\) 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 \((x^*, y^*)\) exists and Assumption 1 and 3 hold, for any \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \), we have
\[
\| \nabla \Phi(x^*) \| = O \left( \sqrt{\frac{d + \log n}{n}} \right).
\]

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
1: Input: \((x_1, y_1) = (0, 0)\), step sizes \(\eta_x > 0, \eta_y > 0\) and dataset \(S = \{z_1, \ldots, z_n\}\)
2: for \(t = 1, \ldots, T\) do
3: update \(x_{t+1} = x_t - \eta_x \nabla_x F_S(x_t, y_t)\)
4: update \(y_{t+1} = y_t + \eta_y \nabla_y F_S(x_t, y_t)\)

Algorithm 2 Two-timescale SGDA for minimax problem
1: Input: \((x_1, y_1) = (0, 0)\), step sizes \(\eta_x > 0, \eta_y > 0\) and dataset \(S = \{z_1, \ldots, z_n\}\)
2: for \(t = 1, \ldots, T\) do
3: update \(x_{t+1} = x_t - \eta_x \nabla_x f(x_t, y_t; z_t)\)
4: update \(y_{t+1} = y_t + \eta_y \nabla_y f(x_t, y_t; z_t)\)

in minimax problems. In this subsection, we provide the empirical 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 multistep GDA [Lin et al., 2020].

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

\[
\frac{1}{T} \sum_{i=1}^{T} \|\nabla \Phi(x_t)\|^2 \leq O \left( \frac{1}{T} \right) + O \left( \frac{d + \log \frac{16 \log(\sqrt{2}R_n + 1)}{\delta}}{\frac{\delta n}{2}} \right).
\]

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

\[
\frac{1}{T} \sum_{i=1}^{T} \|\nabla \Phi(x_t)\|^2 \leq O \left( \frac{d + \log \frac{n}{\delta}}{\sqrt{nd}} \right).
\]

**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 \(\Phi(\bar{x}_T) - \Phi(x^*)\) for Algorithm 1, where \(\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t\). We need to know the empirical optimization error \(\|\nabla \Phi_S(x_T)\|\) 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 \(g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}\) is \(\mu_x\)-strongly-convex-\(\mu_y\)-strongly-concave if \(g(. , y)\) is \(\mu_x\)-strongly-convex for any \(y \in \mathcal{Y}\) and \(g(x, .)\) is \(\mu_y\)-strongly-concave for any \(x \in \mathcal{X}\).

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

**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 \(\|\nabla \Phi_S(x_T)\|\) are given in Section D.2 for GDA and in Section D.3 for SGDA.

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

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

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

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

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

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

**Remark 11.** Theorem 7 shows that the excess risk for primal functions can be bound by \(O \left( \frac{1}{n^2} \right)\) 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 x-side PL-strongly-concave assumption instead.
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 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 $\hat{x}_T$ or even $x_T$ 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 (No.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.BJWZYH1012019100020098); 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-Comfy 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


