UNIFORM LOCALIZED CONVERGENCE AND SHARPER GENERALIZATION BOUNDS FOR MINIMAX PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Minimax problems have achieved widely success in machine learning such as adversarial training, robust optimization, reinforcement learning. Existing studies focus on minimax problems with specific algorithms in stochastic optimization, with only a few work on generalization performance. Current generalization bounds almost all depend on stability, which need case-by-case analyses for specific algorithms. Additionally, recent work provides the $O(\sqrt{d/n})$ generalization bound in expectation based on uniform convergence. In this paper, we study the generalization bounds measured by the gradients of primal functions using the uniform localized convergence. We relax the Lipschitz continuity assumption and give a sharper high probability generalization bound for nonconvex-strongly-concave (NC-SC) stochastic minimax problems considering the localized information. Furthermore, we provide dimension-independent results under Polyak-Lojasiewicz condition for the outer layer. Based on the uniform localized convergence, we analyze some popular algorithms such as the empirical saddle point (ESP), gradient descent ascent (GDA) and stochastic gradient descent ascent (SGDA) and improve the generalization bounds for primal functions. We can even gain approximate $O(1/n^2)$ excess primal risk bounds with further assumptions that the optimal population risks are small, which, to the best of our knowledge, are the sharpest 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 & 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. Most of them were focused on the iteration complexity, 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.

Recently, Zhang et al. (2022) introduced an expectation generalization error for primal functions in minimax problems using complexity. Naturally, we want to create a high-probability version, preferably using 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 & Zeevi (2020) developed a novel "uniform localized convergence" framework using generic chaining for the minimization problems and Li & 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 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, while Zhang et al. (2022) only needed to bound the error caused by the connection between inner and outer layer with $O(1/\sqrt{n})$. We need to introduce the variance for a sharper bound, to bound the error involved by the two layers.

Next, we turn to applications. Firstly, for a sharper excess risk bound, we need to introduce the PL-SC condition to establish a connection between excess risk and the gradient of primal functions, leading to our results for ESP. Unfortunately, for GDA and SGDA algorithms, we found that all the related optimization papers in minimax problems focused on the iteration complexity (or gradient complexity). As a result, they only require $\mathbb{E}[\|\nabla\Phi(\mathbf{x}_T)\|]$ for average of round T with an expected outcome. We were compelled to derive the high probability empirical optimization bound ourselves using classical optimization methods under SC-SC conditions.

Notice that even under SC-SC settings, achieving this for SGDA remains difficult. Drawing inspiration from Lei et al. (2021)'s proof of Primal-Dual Risk optimization bound, we eventually derive the optimization bound for primal risk. The proofs for excess risks in applications differ from minimization problems Li & Liu (2021b) and pose challenges. These challenges stem from errors in the inner and outer layers of minimax problems. Consequently, we can only achieve a result close to $O(1/n^2)$. Our contributions are summarized as follows:

1. We introduce local uniform comvergence 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.

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

3. We extend our main theorems into various algorithms such as ESP, GDA, SGDA. We establish faster $O(1/n)$ order bounds for excess primal risk. We can even gain approximate $O(1/n^2)$ bounds with further assumptions that the optimal population risk is small. To our best knowledge, it is the first time to gain approximate $O(1/n^2)$ for NC-SC minimax problems in expectation and the first result nearly to $O(1/n^2)$ high probability bound 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 main results. Then we apply our main theorems into various algorithms and give the 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 & Bach (2016); Hsieh et al. (2019) for SC-SC cases, Nedić & 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 & 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, which was presented by Rogers & Wagner (1978). It gives 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 & Elisseeff (2002), where they present an important concept called uniform stability. Since then, a lot of works based on uniform stability have emerged. On the one hand, the generalization bound with algorithmic stability have been significantly improved by Bousquet et al. (2020); Feldman & Vondrak (2018; 2019); Klochkov & 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 & 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 & Liu, 2021a), and uniform stability (Lei et al., 2021; Li & Liu, 2021a; Zhang et al., 2021; Farnia & Ozdaglar, 2021; Ozdaglar et al., 2022). Most of them focused on the expectation generalization bounds and only Lei et al. (2021); Li & 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); Davis & Drusvyatskiy (2022) showed that the empirical of generalization error is $O(\sqrt{d/n})$. Xu & Zeevi (2020) developed a novel "uniform localized convergence" framework using generic chaining for the minimization problems and Li & 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 according to $\mathbb{P}$. 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 function (empirical/population))**.** *The primal population function and the primal empirical function are given by*

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

Since F and F_S 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

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

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:

$$\|\nabla\Phi(\mathcal{A}_{\mathbf{x}}(S))\| \le \|\nabla\Phi_S(\mathcal{A}_{\mathbf{x}}(S))\| + \|\nabla\Phi(\mathcal{A}_{\mathbf{x}}(S)) - \nabla\Phi_S(\mathcal{A}_{\mathbf{x}}(S))\|, \tag{4}$$

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 of primal function. The above inequality satisfies from the triangle inequality.

Let $\|\cdot\|$ be the Euclidean norm for simplicity and $B(\mathbf{x}_0, R) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}_0\| \leq R\}$ denote 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 & Ozdaglar, 2021; Lei et al., 2021; Li & Liu, 2021a) and the last one in Assumption 1 is widely used in uniform convergence analysis (Kleywegt et al., 2002; Davis & 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*

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

Lipschitz assmuption is also the standard assumption and widely used in literature such as Zhang et al. (2021); Farnia & Ozdaglar (2021); Lei et al. (2021); Li & 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 $2 \leq k \leq n$,*

$$|\mathbb{E}[(X - \mu)^k]| \leq \frac{1}{2} k! \sigma^2 B^{k-2}. \tag{5}$$

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 the Bernstein condition. Moreover, the 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 the Bernstein condition (Wainwright, 2019). Please refer to Wainwright (2019) for more discussions. Next, we introduce a straightforward generalization of the 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 $2 \leq k \leq n$,*

$$\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^k] \leq \frac{1}{2} k! \mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2] B_{\mathbf{x}^*}^{k-2}. \tag{6}$$

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 $2 \leq k \leq n$,

$$\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^k] \leq \frac{1}{2} k! \mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^2] B_{\mathbf{y}^*}^{k-2}, \tag{7}$$

where $\mathbf{y}^*(\mathbf{x}) := \arg\max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y})$.

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 $2 \leq k \leq n$, we have $\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^k] \leq \frac{1}{2} k! \mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^2] L^{k-2}$, which means that the function f satisfies Bernstein condition for any $\mathbf{x}, \mathbf{y}$. Similarly, $\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^k] \leq \frac{1}{2} k! \mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{z})\|^2] L^{k-2}$ can be easily derived. Furthermore, Bernstein condition assumption is pretty mild since $B_{\mathbf{y}^*}$ only depends on gradients at $(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ for any $\mathbf{x} \in \mathcal{X}$ 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 the 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 the primal functions comparing with Zhang et al. (2022).

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

$$\begin{aligned}
\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\| \leq{}& \frac{\beta}{\mu_{\mathbf{y}}} \left(\sqrt{\frac{2\mathbb{E}\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^2 \log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*} \log\frac{4}{\delta}}{n} \right) \\
&+ \sqrt{\frac{2\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2 \log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*} \log\frac{8}{\delta}}{n} + \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{ \|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n} \right\} \\
&\times \left(\sqrt{\frac{d + \log\frac{16\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}} + \frac{d + \log\frac{16\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n} \right),
\end{aligned}$$

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 $\frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{ \|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n} \right\}$, 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. In the optimal scenario, when* $\|\mathbf{x}-\mathbf{x}^*\| \leq \frac{1}{n}$*, we can attain the best (sharper) results.*

Next, we privide a dimension-free uniform convergence of gradients for the primal functions when the PL condition is satisfied. Firstly, we introduce the extension of the PL condition to the minimax problem 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*

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

Remark 4. *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 & Ma, 2016; Li & Yuan, 2017; Zhou & Liang, 2017; Li & Liang, 2018; Arora et al., 2018; Charles & 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) have 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* $c = \max\{16C^2, 1\}$*. We have that for all* $\mathbf{x} \in \mathcal{X}$*, when* $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$ *with probability at least* $1-\delta$

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

Remark 5. *The following inequality can be easily derived using the norm triangle inequality and Cauchy–Bunyakovsky–Schwarz inequality.*

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

We can easily derive (8) 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 (8),* $\|\nabla\Phi_S(\mathbf{x})\|$ *can be very tiny since most famous optimization algorithms such as GDA and SGDA, can optimize it small enough. The term* $\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2$ *and* $\mathbb{E}\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^2$ *can be also tiny since they only depend on the the gradient of the optima* $\mathbf{x}^*$ *w.r.t* $\mathbf{x}$ *and the gradient of the optima* $\mathbf{y}^*(\mathbf{x})$ *w.r.t.* $\mathbf{y}$*. 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). We further analyze these two terms* $\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2$ *and* $\mathbb{E}\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^2$ *using "Self-bounding" property for smooth function (Srebro et al., 2010) and considering specific algorithms in Section 5, which can derive to almost* $O(1/n^2)$ *bounds. 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.*

Remark 6. *There are two mainly challenges in our work for minimax problems. On one hand, comparing with uniform convergence in Zhang et al. (2022), we use a novel uniform localized convergence techniques (Xu & Zeevi, 2020) to construct a functional w.r.t. loss functions on minimax problems. This two layer structure involves difficulties. On the other hand, 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* $O(\frac{1}{\sqrt{n}})$*. But we need to reach the upper bound of order* $O(\frac{1}{n})$ *under certain assumptions.*

5 APPLICATION

5.1 EMPIRICAL SADDLE POINT

Empirical saddle point (ESP) problem, 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 gives that the population optimization error* $\|\nabla\Phi(\hat{\mathbf{x}}^*)\|$ *is of order* $O\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}\right)$ *(*$\log n$ *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* $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$*, where c is an absolute constant, we have*

$$\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*) \leq \frac{12\beta^2\mathbb{E}\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*, \mathbf{y}^*(\hat{\mathbf{x}}^*); \mathbf{z})\|^2 \log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{12\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2 \log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} + \frac{3\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{2\mu_{\mathbf{x}} n^2}.$$

Furthermore, if we let $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \frac{48\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right\}$ *and assume the function* $f(\mathbf{x},\mathbf{y};\mathbf{z})$ *is non-negative and* $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$*, we have*

$$\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*) = O\left(\frac{\log^2\frac{1}{\delta}}{n\left(n - \frac{48\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)}\right).$$

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 almost* $O(1/n^2)$*. 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 & Zhou (2019); Lei & 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 & Liu, 2021b), which studied the general machine learning settings for* $f(\mathbf{w})$ *under PL condition. Their empirical risk minimizer (ERM) excess risk bounds provided* $O\left(1/n^2\right)$ *order rates. We analyze the excess risk with primal functions which involve an additional error term. In consequence, our result for ESP just approximate* $O\left(1/n^2\right)$ *order rate.*

Algorithm 1 Two-timescale GDA for minimax problem

1: **Input:** $(\mathbf{x}_1, \mathbf{y}_1)$, step sizes $\eta_{\mathbf{x}} > 0, \eta_{\mathbf{y}} > 0$ and dataset $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$
2: **for** $t = 1, \ldots, T$ **do**
3: update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)$
4: 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

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

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 in minimax problems. In this subsection, we provide the generalization and 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 $\{\mathbf{x}_t\}_t$ 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 reveals that we need to balance the 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 SGD, 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 SGD. Comparing with Theorem 3, Theorem 6 only necessitates smooth and Lipschitz conditions (Assumption 1 and 2) and doesn't require PL conditions. In fact, Theorem 7 constitutes a further extension of Theorem 3.

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 $\|\nabla \Phi_S(\bar{\mathbf{x}}_T)\|$. 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 & 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 $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, where c is an absolute constant, we have*

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

Furthermore, Let $T \asymp n^2$ and $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \frac{128\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right\}$. Assume the function $f(\mathbf{x},\mathbf{y};\mathbf{z})$ is non-negative and $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\log^2\frac{1}{\delta}}{n\left(n - \frac{128\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)}\right).$$

Remark 11. *Theorem 7 shows that the excess risk for primal functions can be bound almost to $O\left(1/n^2\right)$ comparing with the optimal result $O(1/n)$ given in Li & 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.*

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 stepsizes, presented in Algorithm 2 which is more nature in the 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 $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, where c is an absolute constant, we have*

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

Furthermore, let $T \asymp n^4$ and $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \frac{128\beta^3\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right\}$. Assume the function $f(\mathbf{x},\mathbf{y};\mathbf{z})$ is non-negative and $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\log^2\frac{1}{\delta}}{n\left(n - \frac{128\beta^3\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)}\right).$$

Remark 12. *Theorem 8 reveals that under the SC-SC settings, the excess risk bound can be approximate $O(1/n^2)$ comparing with the optimal result $O(1/n)$ given in (Li & 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.*

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. Then we provide dimension-independent results under PL condition. Finally we extend our main theorems into various algorithms to reach the optimal excess primal risk bounds. 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. Combining with our generalization work, we can get a tighter excess primal risk bound with weaker conditions.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

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.

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

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4381–4391, 2020.

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

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

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pp. 745–754. PMLR, 2018.

Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

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 *International Conference on Machine Learning*, pp. 1125–1134. PMLR, 2018.

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

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

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058. PMLR, 2017.

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

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 31, 2018.

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

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.

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

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.

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

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

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.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, 34:5065–5076, 2021.

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

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

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pp. 6175–6186. PMLR, 2021.

Shaojie Li and Yong Liu. High probability generalization bounds with fast rates for minimax problems. In *International Conference on Learning Representations*, 2021a.

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

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, 30, 2017.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, and Tianbao Yang. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pp. 2159–2167. PMLR, 2017.

Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pp. 6370–6381. PMLR, 2020.

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 Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

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

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

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

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

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. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

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

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.

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

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.

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

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

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

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.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

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

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800, 2020.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

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

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*, pp. 3160–3179. PMLR, 2019.

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*, pp. 1954–1979. PMLR, 2017.

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

Yi Zhou and Yingbin Liang. Characterization of gradient dominance and regularity conditions for neural networks. *arXiv preprint arXiv:1710.06910*, 2017.

A ADDITIONAL DEFINITIONS AND LEMMATA

Lemma 1 (Bernstein's inequality (Dirksen, 2015)). *Let $X_1, \dots, X_n$ be real-valued, independent, mean-zero random variables and suppose that for some constants $\sigma, B > 0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|X_i|^k \le \frac{k!}{2}\sigma^2 B^{k-2}, \quad k = 2, 3, \dots$$

Then, $\forall \delta \in (0,1)$, with probability at least $1-\delta$

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \sqrt{\frac{2\sigma^2 \log\frac{2}{\delta}}{n}} + \frac{B\log\frac{2}{\delta}}{n}. \tag{9}$$

Lemma 2 (A variant of the "uniform localized convergence" argument (Xu & Zeevi, 2020)). *Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z}$ and $\mathbb{P}_n$ be the corresponding empirical probability measure. For a function class $\mathcal{H} = \{h_f : f \in \mathcal{F}\}$ and functional $T : \mathcal{F} \to [0, R]$, assume there is a function $\psi(r;\delta)$ (possibly depending on the samples), which is non-decreasing with respect to r and satisfies that $\forall \delta \in (0,1)$, $\forall r \in [0, R]$, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}:T(f)\le r} (\mathbb{P}-\mathbb{P}_n)h_f \le \psi(r;\delta).$$

Then, given any $\delta \in (0,1)$ and $r_0 \in (0, R]$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$,

$$(\mathbb{P}-\mathbb{P}_n)h_f \le \psi\left(2T(f) \vee r_0; \frac{\delta}{C_{r_0}}\right),$$

where $C_{r_0} = 2\log_2 \frac{2R}{r_0}$.

Definition 6 (Orlicz$_\alpha$ norm (Dirksen, 2015)). *For every $\alpha \in (0, +\infty)$ we define the Orlicz$_\alpha$ norm of a random u:*

$$\|u\|_{\mathit{Orlicz}_\alpha} = \inf\{K > 0 : \mathbb{E}\exp((|u|/K)^\alpha) \le 2\}.$$

A random variable (or vector) $X \in \mathbb{R}^d$ is K-sub-Gaussian if $\forall \boldsymbol{\lambda} \in \mathbb{R}^d$, we have

$$\|\boldsymbol{\lambda}^{\mathrm{T}} X\|_{\mathit{Orlicz}_2} = K\|\boldsymbol{\lambda}\|_2.$$

A random variable (or vector) $X \in \mathbb{R}^d$ is K-sub-exponential if $\forall \boldsymbol{\lambda} \in \mathbb{R}^d$, we have

$$\|\boldsymbol{\lambda}^{\mathrm{T}} X\|_{\mathit{Orlicz}_1} = K\|\boldsymbol{\lambda}\|_2.$$

Definition 7 (Orlicz$_\alpha$ processes (Dirksen, 2015)). *Let $\{X_f\}_{f\in\mathcal{F}}$ be a sequence of random variables. $\{X_f\}_{f\in\mathcal{F}}$ is called an Orlicz$_\alpha$ process for a metric $\mathsf{metr}(\cdot,\cdot)$ on $\mathcal{F}$ if*

$$\|X_{f_1} - X_{f_2}\|_{\mathit{Orlicz}_\alpha} \le \mathsf{metr}(f_1, f_2), \quad \forall f_1, f_2 \in \mathcal{F}.$$

Typically, the Orlicz$_2$ process is called "process with sub-Gaussian increments" and the Orlicz$_1$ process is called "process with sub-exponential increments".

Definition 8 (Mixed sub-Gaussian-sub-exponential increments (Dirksen, 2015)). *We say a process $(X_\theta)_{\theta\in\Theta}$ has mixed sub-Gaussian-sub-exponential increments with respect to the pair $(\mathsf{metr}_1, \mathsf{metr}_2)$ if for all $\theta_1, \theta_2 \in \Theta$,*

$$Prob(\|X_{\theta_1} - X_{\theta_2}\| \ge \sqrt{u}\cdot \mathsf{metr}_2(\theta_1,\theta_2) + u\cdot \mathsf{metr}_1(\theta_1,\theta_2)) \le 2e^{-u}, \forall u \ge 0,$$

where "$Prob$" means probability.

Definition 9 (Talagrand's γ_α-functional (Dirksen, 2015)). *A sequence $F = (\mathcal{F}_n)_{n\ge 0}$ of subsets of $\mathcal{F}$ is called admissible if $|\mathcal{F}_0| = 1$ and $|\mathcal{F}_n| \le 2^{2^n}$ for all $n \ge 1$. For any $0 < \alpha < \infty$, the γ_α-functional of $(\mathcal{F}, \mathsf{metr})$ is defined by*

$$\gamma_\alpha(F, d) = \inf_F \sup_{f\in\mathcal{F}} \sum_{n=0}^{\infty} 2^{\frac{n}{\alpha}} \mathsf{metr}(f, \mathcal{F}_n),$$

where the infimum is taken over all admissible sequences and we write $\mathsf{metr}(f, \mathcal{F}_n) = \inf_{s\in\mathcal{F}_n} \mathsf{metr}(f, s)$.

Lemma 3 (Bernstein's inequality for sub-exponential random variables (Wainwright, 2019)). *If $X_1, \dots X_n$ are sub-exponential random variables, the Bernstein's inequality in Lemma 1 holds with*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2_{Orlicz_1}, \quad B = \max_{1\le i\le n} \|X_i\|_{Orlicz_1}.$$

Lemma 4 (Vector Bernstein's inequality (Pinelis, 1994; Smale & Zhou, 2007; Xu & Zeevi, 2020)). *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables taking values in a real separable Hilbert space. Assume that $\mathbb{E}[X_i] = \mu$, $\mathbb{E}[\|X_i - \mu\|^2] = \sigma^2$, $\forall 1 \le i \le n$, we say that vector Bernstein's condition with parameter B holds if for all $1 \le i \le n$,*

$$\mathbb{E}[\|X_i - \mu\|^k] \le \frac{1}{2}k!\sigma^2 B^{k-2}, \quad \forall 2 \le k \le n. \tag{10}$$

If this condition holds, then for all $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right\| \le \sqrt{\frac{2\sigma^2 \log(\frac{2}{\delta})}{n}} + \frac{B\log\frac{2}{\delta}}{n}. \tag{11}$$

Definition 10 (Covering number (Wainwright, 2019)). *Assume $(\mathcal{M}, \mathsf{metr})$ is a metric space and $\mathcal{F} \subseteq \mathcal{M}$. For any $\epsilon > 0$, a set $\mathcal{F}_c$ is called an ϵ-conver of $\mathcal{F}$ if for any $f \in \mathcal{F}$ we have an element $g \in \mathcal{F}_c$ such that $\mathsf{metr}(f,g) \le \epsilon$. We denote $N(\mathcal{F}, \mathsf{metr}, \epsilon)$ the covering number as the cardinality of the minimal ϵ-cover of $\mathcal{F}$:*

$$N(\mathcal{F}, \mathsf{metr}, \epsilon) = \min\{|\mathcal{F}_c| : \mathcal{F}_c \textit{is an } \epsilon\textit{-cover of } \mathcal{F}\}.$$

Lemma 5 (Dudley's integral bound for γ_α functional (Talagrand, 1996)). *There exist a constant C_α depending only on α such that*

$$\gamma_\alpha(\mathcal{F}, \mathsf{metr}) \le C_\alpha \int_0^{+\infty} (\log N(\mathcal{F}, \mathsf{metr}, \epsilon))^{\frac{1}{\alpha}} d\epsilon$$

Lemma 6 (Generic chaining for a process with mixed tail increments in (Dirksen, 2015)). *If $(X_f)_{f\in\mathcal{F}}$ has mixed sub-Gaussian-sub-exponential increments with respect to the pair $(\mathsf{metr}_1, \mathsf{metr}_2)$, there are absolute constants $c, C > 0$ such that for $\forall \delta \in (0,1)$, with probability at least $1-\delta$*

$$\sup_{\theta\in\Theta} \|X_f - X_{f_0}\| \le C(\gamma_2(\mathcal{F}, \mathsf{metr}_2) + \gamma_1(\mathcal{F}, \mathsf{metr}_1)) +$$
$$c\left(\sqrt{\log\frac{1}{\delta}} + \sup_{f_1,f_2\in\mathcal{F}}[\mathsf{metr}_2(f_1,f_2)] + \log\frac{1}{\delta}\sup_{f_1,f_2\in\mathcal{F}}[\mathsf{metr}_1(f_1,f_2)]\right).$$

Lemma 7 ("Self-bounding" property for smooth function (Srebro et al., 2010)). *For a β-smooth and non-negative function $f : \mathbf{w} \to \mathbb{R}$, for all $\mathbf{w} \in \mathcal{W}$:*

$$\|\nabla f(\mathbf{w})\| \le \sqrt{4\beta f(\mathbf{w})}$$

B SOME BASIC LEMMATA IN MINIMAX PROBLEMS

Lemma 8 (Smoothness for primal function (Nouiehed et al., 2019)). *Suppose Assumption 1 holds, then the function $\Phi(\mathbf{x})$ and $\Phi_S(\mathbf{x})$ is $\beta + \frac{\beta^2}{\mu_\mathbf{y}}$-smooth.*

Lemma 9 (PL condition for primal function (Yang et al., 2020)). *For NC-SC setting, suppose Assumption 1 holds. Assume that the population risk $F(\mathbf{x},\mathbf{y})$ satisfies Assumption 4 with parameter $\mu_\mathbf{x}$, then function $\Phi(\mathbf{x})$ satisfies the PL condition with $\mu_\mathbf{x}$, which means that for all $\mathbf{x} \in \mathcal{X}$*

$$\Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \le \frac{1}{2\mu_\mathbf{x}}\|\nabla\Phi(\mathbf{x})\|^2.$$

For SC-SC setting, suppose Assumption 5 holds, then function $\Phi(\mathbf{x})$ and $\Phi_S(\mathbf{x})$ satisfy the PL condition with $\mu_\mathbf{x}$, which means that for all $\mathbf{x} \in \mathcal{X}$

$$\Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \le \frac{1}{2\mu_\mathbf{x}}\|\nabla\Phi(\mathbf{x})\|^2 \quad \textit{and} \quad \Phi_S(\mathbf{x}) - \Phi_S(\mathbf{x}^*) \le \frac{1}{2\mu_\mathbf{x}}\|\nabla\Phi_S(\mathbf{x})\|^2.$$

Definition 11. *For a given* $\mathbf{x}$*, the empirical optimal point* $\mathbf{y}^*(\mathbf{x})$ *and the population optimal point* $\mathbf{y}_S^*(\mathbf{x})$ *are given as follows,*

$$\mathbf{y}^*(\mathbf{x}) := \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) \quad \textit{and} \quad \mathbf{y}_S^*(\mathbf{x}) := \arg\max_{\mathbf{y}\in\mathcal{Y}} F_S(\mathbf{x},\mathbf{y}).$$

Remark 13. *According to the definition, we can easily derive the following equations that* $\mathbf{y}^* = \mathbf{y}^*(\mathbf{x}^*) = \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}^*,\mathbf{y})$ *if* $(\mathbf{x}^*,\mathbf{y}^*)$ *is the solution to (1) and* $\hat{\mathbf{y}}^* = \mathbf{y}_S^*(\hat{\mathbf{x}}^*) = \arg\max_{\mathbf{y}\in\mathcal{Y}} F_S(\hat{\mathbf{x}}^*,\mathbf{y})$ *if* $(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*)$ *is the solution to (2).*

Lemma 10 (Concentration of $\mathbf{y}^*$ (Zhang et al., 2022))**.** *For* $\mathbf{y}^*(\mathbf{x})$ *and* $\mathbf{y}_S^*(\mathbf{x})$ *defined in Definition 11, with Assumption 1, we have* $\forall \mathbf{x}\in\mathcal{X}$*,*

$$\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\| \le \frac{1}{\mu_{\mathbf{y}}} \|\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{y}} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\|.$$

Lemma 11. *Suppose Assumption 1 and 3 hold, For any* $\mathbf{x}\in\mathcal{X}$*, With probability at least* $1-\delta$ *we have the following inequalities*

$$\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\| \le \frac{1}{\mu_{\mathbf{y}}} \left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]\log\frac{2}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{2}{\delta}}{n} \right).$$

Proof of Lemma 11. From Bernstein's inequality for vectors (Lemma 4), applying Lemma 10, For any $\mathbf{x}\in\mathcal{X}$, With probability at least $1-\delta$ we have the following inequalities

$$\begin{aligned}\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\| &\le \frac{1}{\mu_{\mathbf{y}}} \|\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{y}} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\| \\ &\le \frac{1}{\mu_{\mathbf{y}}} \left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]\log\frac{2}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{2}{\delta}}{n} \right).\end{aligned}$$

Notice that Bernstein inequality for vectors (Lemma 4) under Assumption 3 can be used here because for all $\mathbf{x}\in\mathcal{X}$, $\mathbf{y}^*(x)$ and $F_S(\cdot)$ are independent.

Next, we need to point out that we only need pointwise convergence for any fixed $\mathbf{x}$ instead of unifrom convergence on $\mathcal{X}$. We take the standard minimization problem as an example. For population risk $R(\theta) = \mathbb{E}_z \ell(\theta, z)$, empirical risk $r(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, z_i)$ and fixed θ, we can directly apply concentration inequality. But we can't apply them to θ_{ERM} as it is a function of the dataset. We need to establish the sup that $R(\theta_{\mathrm{ERM}}) - r(\theta_{\mathrm{ERM}}) \le \sup_{\theta_{sup}}[R(\theta_{sup}) - r(\theta_{sup})]$, where $\theta_{sup} = \arg\min_\theta R(\theta) - r(\theta)$. Yet, as the dataset changes, the parameters θ_{sup} change accordingly. Thus, we require uniform convergence for function $R - r$. In Lemma 11 of our proof, for any $\mathbf{x}$, when the dataset changes, $\mathbf{x}$ and $\mathbf{y}^*(\mathbf{x})$ remain unchanged, and $\nabla F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))$ are the only altered random variables/vectors. Consequently, we can directly apply the Bernstein inequality, thus we only need pointwise convergence for the function $y^*(x) - y_S^*(x)$ w.r.t. x, which suffices.

□

Lemma 12 (Zhang et al. (2021))**.** *For* $\mathbf{y}^*$ *and* $\mathbf{y}_S^*$ *defined in Definition 11, with Assumption 1, then for* $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$*,*

$$\|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)\| \le \frac{\beta}{\mu_{\mathbf{y}}} \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Lemma 13. *Suppose a function* $f : \mathcal{X}\times\mathcal{Y}\times\mathcal{Z} \to \mathbb{R}$ *is* β*-smooth in* $(\mathbf{x},\mathbf{y})$ *and the function* f *is* $\mu_{\mathbf{y}}$*-strongly concave in* $\mathbf{y}\in\mathcal{Y}$ *for any* $\mathbf{x}\in\mathcal{X}$ *and* $\mathbf{z}\in\mathcal{Z}$*. Then we have* $\frac{\mathbf{u}^{\mathrm{T}}(\nabla_{\mathbf{x}} f(\mathbf{x}_1,\mathbf{y}^*(\mathbf{x}_1);\mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2,\mathbf{y}^*(\mathbf{x}_2);\mathbf{z}))}{\|\mathbf{x}_1-\mathbf{x}_2\|}$ *is a* $\frac{\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}$*-sub-exponential random vector. That is for any unit vector* $\mathbf{u}\in B(0,1)$ *and* $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}$*,*

$$\mathbb{E}\left\{ \exp\left(\frac{\left|\mathbf{u}^{\mathrm{T}}(\nabla_{\mathbf{x}} f(\mathbf{x}_1,\mathbf{y}^*(\mathbf{x}_1);\mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2,\mathbf{y}^*(\mathbf{x}_2);\mathbf{z}))\right|}{\frac{\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\|\mathbf{x}_1-\mathbf{x}_2\|} \right) \right\} \le 2.$$

Proof of Lemma 13. According to Definition 3, for any sample $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z})\| \\
\leq &\beta \|\mathbf{x}_1 - \mathbf{x}_2\| + \beta \|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)\| \\
\leq &\beta \|\mathbf{x}_1 - \mathbf{x}_2\| + \frac{\beta^2}{\mu_{\mathbf{y}}} \|\mathbf{x}_1 - \mathbf{x}_2\| \\
= &\frac{\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \|\mathbf{x}_1 - \mathbf{x}_2\|,
\end{aligned}$$

where the first inequality uses the smoothness and the second inequality applies Lemma 12.

Then, for any unit vector $\mathbf{u} \in B(0, 1)$, we have

$$\begin{aligned}
&\left|\mathbf{u}^{\mathrm{T}} (\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}))\right| \\
\leq &\|\mathbf{u}^{\mathrm{T}}\| \|\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z})\| \\
\leq &\frac{\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \|\mathbf{x}_1 - \mathbf{x}_2\|,
\end{aligned}$$

which implies

$$\frac{\left|\mathbf{u}^{\mathrm{T}} (\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}))\right|}{\frac{\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \|\mathbf{x}_1 - \mathbf{x}_2\|} \leq 1.$$

Then we get

$$\mathbb{E}\left\{\exp\left(\frac{\left|\mathbf{u}^{\mathrm{T}} (\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}))\right|}{\frac{\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}} \ln 2} \|\mathbf{x}_1 - \mathbf{x}_2\|}\right)\right\} \leq 2.$$

The proof is complete. □

Lemma 14. *Suppose Assumption 1 holds, we have the following inequality that for all $\mathbf{x} \in \mathcal{X}$ and for any $\delta \in (0, 1)$, with probability at lest $1 - \delta$,*

$$\begin{aligned}
&\|(\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))) - (\nabla_{\mathbf{x}} F(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^*, \mathbf{y}^*))\| \\
\leq &\frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log \frac{4 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{4 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right),
\end{aligned}$$

where C is an absolute constant.

Proof of Lemma 14. We define $\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| \leq \max\{R_1, \frac{1}{n}\}\}$. For all $(\mathbf{x}, \mathbf{v}) \in \mathcal{X} \times \mathcal{V}$, let $g_{(\mathbf{x}, \mathbf{v})}(\mathbf{z}) = \left(\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} \mathbf{v}$. Then for any $(\mathbf{x}_1, \mathbf{v}_1)$ and $(\mathbf{x}_2, \mathbf{v}_2) \in \mathcal{X} \times \mathcal{V}$, we define the following norm on the product space $\mathcal{X} \times \mathcal{V}$

$$\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X} \times \mathcal{V}} = \left(\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{v}_1 - \mathbf{v}_2\|^2\right)^{\frac{1}{2}}.$$

Then we define a ball on the product space $\mathcal{X} \times \mathcal{V}$ that $B(\sqrt{r}) = \{(\mathbf{x}, \mathbf{v}) \in \mathcal{X} \times \mathcal{V} : \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{v}\|^2 \leq r\}$. Given any $(\mathbf{x}_1, \mathbf{v}_1)$ and $(\mathbf{x}_2, \mathbf{v}_2) \in B(\sqrt{r})$, we have the following decomposition

$$\begin{aligned}
&g_{(\mathbf{x}_1, \mathbf{v}_1)}(\mathbf{z}) - g_{(\mathbf{x}_2, \mathbf{v}_2)}(\mathbf{z}) \\
= &\left(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} \mathbf{v}_1 \\
&- \left(\nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} \mathbf{v}_2 \\
= &\left(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} (\mathbf{v}_1 - \mathbf{v}_2) \\
&+ \left(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} \mathbf{v}_2 \\
&- \left(\nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} \mathbf{v}_2 \\
= &\left(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\right)^{\mathrm{T}} (\mathbf{v}_1 - \mathbf{v}_2) \\
&+ \left(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z})\right)^{\mathrm{T}} \mathbf{v}_2.
\end{aligned}$$

Since $(\mathbf{x}_1, \mathbf{v}_1)$ and $(\mathbf{x}_2, \mathbf{v}_2) \in B(\sqrt{r})$, we have

$$\|\mathbf{x}_1 - \mathbf{x}^*\|\|\mathbf{v}_1 - \mathbf{v}_2\| \leq \sqrt{r}\|\mathbf{v}_1 - \mathbf{v}_2\| \leq \sqrt{r}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}. \tag{12}$$

Next, from Assumption 1 and Lemma 13, we know that $\frac{\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z})}{\|\mathbf{x}_1 - \mathbf{x}_2\|}$ is a $\frac{\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}$-sub-exponential random vector for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, which means that

$$\mathbb{E}\left\{\exp\left(\frac{(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*(\mathbf{x}^*); \mathbf{z}))^{\mathrm{T}}(\mathbf{v}_1 - \mathbf{v}_2)}{\frac{\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|\mathbf{x}_1 - \mathbf{x}^*\|\|\mathbf{v}_1 - \mathbf{v}_2\|}\right)\right\} \leq 2. \tag{13}$$

We combine (12) and (13), according to Definition 6, we can easily deduce that $(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z}))^{\mathrm{T}}(\mathbf{v}_1 - \mathbf{v}_2)$ is $\frac{\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}$-sub-exponential. Similarly, we can derive that

$$\|\mathbf{x}_1 - \mathbf{x}_2\|\|\mathbf{v}_2\| \leq \sqrt{r}\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \sqrt{r}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}.$$

Also, there holds that

$$\mathbb{E}\left\{\exp\left(\frac{(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}))^{\mathrm{T}}(\mathbf{v}_2)}{\frac{\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|\mathbf{x}_1 - \mathbf{x}_2\|\|\mathbf{v}_2\|}\right)\right\} \leq 2.$$

Thus, we have that $(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}))^{\mathrm{T}}(\mathbf{v}_2)$ is also $\frac{\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}$-sub-exponential. Now for any $(\mathbf{x}_1, \mathbf{v}_1)$ and $(\mathbf{x}_2, \mathbf{v}_2) \in B(\sqrt{r})$, we know

$$\begin{aligned}
&\mathbb{E}\left\{\exp\left(\frac{g_{(\mathbf{x}_1,\mathbf{v}_1)}(\mathbf{z}) - g_{(\mathbf{x}_2,\mathbf{v}_2)}(\mathbf{z})}{\frac{2\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}}\right)\right\} \\
\leq &\mathbb{E}\left\{\frac{1}{2}\exp\left(\frac{(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z}))^{\mathrm{T}}(\mathbf{v}_1 - \mathbf{v}_2)}{\frac{\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}}\right)\right\} \\
+&\mathbb{E}\left\{\frac{1}{2}\exp\left(\frac{(\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}^*(\mathbf{x}_1); \mathbf{z}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}^*(\mathbf{x}_2); \mathbf{z}))^{\mathrm{T}}(\mathbf{v}_2)}{\frac{\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}}\right)\right\} \leq 2,
\end{aligned} \tag{14}$$

where the first inequality follows from Jensen's inequality. (14) implies that $g_{(\mathbf{x}_1,\mathbf{v}_1)}(\mathbf{z}) - g_{(\mathbf{x}_2,\mathbf{v}_2)}(\mathbf{z})$ is a $\frac{2\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}$-sub-exponential random variable, for which we have

$$\|g_{(\mathbf{x}_1,\mathbf{v}_1)}(\mathbf{z}) - g_{(\mathbf{x}_2,\mathbf{v}_2)}(\mathbf{z})\|_{\text{Orlicz}_1} \leq \frac{2\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}. \tag{15}$$

From Bernstein inequality for sub-exponential variables (Lemma 3), for any fixed $u \leq 0$ and $(\mathbf{x}_1.\mathbf{v}_1), (\mathbf{x}_2, \mathbf{v}_2) \in \mathcal{X} \times \mathcal{V}$, we have

$$\begin{aligned}
Pr\Bigg(|(\mathbb{P} - \mathbb{P}_n)[g_{(\mathbf{x}_1,\mathbf{v}_1)}(\mathbf{z}) - g_{(\mathbf{x}_2,\mathbf{v}_2)}(\mathbf{z})]| \geq &\frac{2\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}\sqrt{\frac{2u}{n}} \\
&+\frac{2u\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{n\mu_{\mathbf{y}} \ln 2}\|(\mathbf{x}_1, \mathbf{v}_1) - (\mathbf{x}_2, \mathbf{v}_2)\|_{\mathcal{X}\times\mathcal{V}}\Bigg) \leq 2e^{-u}.
\end{aligned} \tag{16}$$

According to Definition 8, we know that $(\mathbb{P} - \mathbb{P}_n)g_{(\mathbf{x},\mathbf{v})}(\mathbf{z})$ is a mixed sub-Gaussian-sub-exponential increments w.r.t. the metrics $(\frac{2\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{n\mu_{\mathbf{y}} \ln 2}\|\cdot\|_{\mathcal{X}\times\mathcal{V}}, \frac{2\beta\sqrt{2r}(\mu_{\mathbf{y}}+\beta)}{\sqrt{n}\mu_{\mathbf{y}} \ln 2}\|\cdot\|_{\mathcal{X}\times\mathcal{V}})$ from (16).

Then from the generic chaining for a process with mixed tail increments in Lemma 6, we have the following inequality that for all $\delta \in (0,1)$, with probability at least $1-\delta$

$$\begin{aligned}
&\sup_{\|\mathbf{x}_1-\mathbf{x}^*\|^2+\|\mathbf{v}\|^2\leq r} |(\mathbb{P}-\mathbb{P}_n)g_{(\mathbf{x},\mathbf{v})}(\mathbf{z})| \\
\leq& C\left(\gamma_2\left(B(\sqrt{r}), \frac{2\beta\sqrt{2r}(\mu_{\mathbf{y}}+\beta)}{\sqrt{n}\mu_{\mathbf{y}}\ln 2}\|\cdot\|_{\mathcal{X}\times\mathcal{V}}\right)+\gamma_1\left(B(\sqrt{r}), \frac{2\beta\sqrt{r}(\mu_{\mathbf{y}}+\beta)}{n\mu_{\mathbf{y}}\ln 2}\|\cdot\|_{\mathcal{X}\times\mathcal{V}}\right)\right. \\
&\left.+\frac{r\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\sqrt{\frac{\log\frac{1}{\delta}}{n}}+\frac{r\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\frac{\log\frac{1}{\delta}}{n}\right).
\end{aligned}$$

Next we use the Dudley's integral (Lemma 5) to bound the γ_1 and γ_2 functional. So there exists an absolute constant C such that for any $\delta \in (0,1)$, with probability at least $1-\delta$

$$\sup_{\|\mathbf{x}_1-\mathbf{x}^*\|^2+\|\mathbf{v}\|^2\leq r} |(\mathbb{P}-\mathbb{P}_n)g_{(\mathbf{x},\mathbf{v})}(\mathbf{z})| \leq \frac{rC\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}+\frac{d+\log\frac{1}{\delta}}{n}\right). \tag{17}$$

Then, the next step is to apply Lemma 2 to (17). We denote $T(f)=\|\mathbf{x}-\mathbf{x}^*\|^2+\|\mathbf{v}\|^2$, $\psi(r;\delta)=\frac{rC\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}+\frac{d+\log\frac{1}{\delta}}{n}\right)$. Since $\|\mathbf{x}-\mathbf{x}^*\|^2+\|\mathbf{v}\|^2 \leq R_1^2+R_1^2+\frac{1}{n^2}$, we set $R^2=2R_1^2+\frac{1}{n^2}$ and $r_0=\frac{2}{n^2}$. According to Lemma 2, we have the following inequality that for any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $\mathbf{x}\in\mathcal{X}$ and $\mathbf{v}\in\mathcal{V}$,

$$\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)g_{(\mathbf{x},\mathbf{v})}(\mathbf{z}) \\
=&(\mathbb{P}-\mathbb{P}_n)[(\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})-\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z}))^{\mathrm{T}}\mathbf{v}] \\
\leq&\psi\left(\max\left\{\|\mathbf{x}-\mathbf{x}^*\|^2+\|\mathbf{v}\|^2, \frac{2}{n^2}\right\}; \frac{\delta}{2\log_2(Rn^2)}\right) \\
\leq&\frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|^2+\|\mathbf{v}\|^2, \frac{2}{n^2}\right\}\times\left(\sqrt{\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{n}}+\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{n}\right).
\end{aligned} \tag{18}$$

Finally, we choose $\mathbf{v}=\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\}\frac{(\mathbb{P}-\mathbb{P}_n)(\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})-\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z}))}{\|(\mathbb{P}-\mathbb{P}_n)(\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})-\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z}))\|}$. It is clear that $\|\mathbf{v}\|=\max\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\}\leq\max\{R_1, \frac{1}{n}\}$, which belongs to the space $\mathcal{V}$. Plugging this $\mathbf{v}$ into (18), we have the following inequality that for any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $\mathbf{x}\in\mathcal{X}$,

$$\begin{aligned}
&\|(\mathbb{P}-\mathbb{P}_n)(\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})-\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z}))\| \\
\leq&\frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\}\times\left(\sqrt{\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{n}}+\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{n}\right) \\
=&\frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}\ln 2}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\}\times\left(\sqrt{\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}}+\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right).
\end{aligned} \tag{19}$$

Finally we have

$$\begin{aligned}
&\|(\nabla_{\mathbf{x}}F(\mathbf{x},\mathbf{y}^*(\mathbf{x}))-\nabla_{\mathbf{x}}F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x})))-(\nabla_{\mathbf{x}}F(\mathbf{x}^*,\mathbf{y}^*)-\nabla_{\mathbf{x}}F_S(\mathbf{x}^*,\mathbf{y}^*))\| \\
=&\|(\mathbb{P}-\mathbb{P}_n)(\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})-\nabla_{\mathbf{x}}f(\mathbf{x}^*,\mathbf{y}^*(\mathbf{x}^*);\mathbf{z}))\| \\
\leq&\frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\}\times\left(\sqrt{\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}}+\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right),
\end{aligned}$$

where C is an absolute constant.

The proof is complete. □

Lemma 15. *Suppose Assumption 1 holds, we have the following inequality that for all $\mathbf{x} \in \mathcal{X}$ and for any $\delta \in (0,1)$, with probability at lest $1-\delta$,*

$$\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \sqrt{\frac{2\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*(\mathbf{x}^*); \mathbf{z})\|^2 \log \frac{4}{\delta}}{n}} + \frac{B_{\mathbf{x}^*} \log \frac{4}{\delta}}{n}$$
$$+ \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log \frac{8 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{8 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right).$$

Proof of Lemma 15. According to Lemma 14, we have the following inequality that for any $\delta \in (0,1)$, with probability at least $1 - \frac{\delta}{2}$

$$\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \|\nabla_{\mathbf{x}} F(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^*, \mathbf{y}^*)\|$$
$$+ \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log \frac{8 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{8 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right), \tag{20}$$

where this inequality applies norm triangle inequality. Then we need to bound $\|\nabla_{\mathbf{x}} F(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^*, \mathbf{y}^*)\|$.

According to Lemma 4, with probability at least $1 - \frac{\delta}{2}$

$$\|\nabla_{\mathbf{x}} F(\mathbf{x}^*, \mathbf{y}^*) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^*, \mathbf{y}^*)\| \leq \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2] \log \frac{4}{\delta}}{n}} + \frac{B_{\mathbf{x}^*} \log \frac{4}{\delta}}{n}. \tag{21}$$

Combining (20) and (21), for any $\delta \in (0,1)$, with probability at least $1 - \delta$, we have

$$\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2] \log \frac{4}{\delta}}{n}} + \frac{B_{\mathbf{x}^*} \log \frac{4}{\delta}}{n}$$
$$+ \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log \frac{8 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{8 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right).$$

The proof is complete. □

C PROOFS IN SECTION 4

Proof of Theorem 1. Firstly, for all $\mathbf{x} \in \mathcal{X}$, we divide $\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\|$ into two terms

$$
\begin{aligned}
&\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\| \\
=&\left\|\mathbb{E}_{\mathbf{z}}\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}) - \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_S^*(\mathbf{x}); \mathbf{z}_i)\right\| \\
=&\left\|\mathbb{E}_{\mathbf{z}}\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}) - \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}_i)\right. \\
&\qquad \left. + \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_S^*(\mathbf{x}); \mathbf{z}_i)\right\| \\
\leq&\left\|\mathbb{E}_{\mathbf{z}}\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}) - \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}_i)\right\| \\
&\qquad + \left\|\frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_S^*(\mathbf{x}); \mathbf{z}_i)\right\| \\
\leq&\left\|\mathbb{E}_{\mathbf{z}}\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}) - \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z}_i)\right\| + \beta\left\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\right\| \\
=&\left\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right\| + \beta\left\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\right\|,
\end{aligned}
\tag{22}
$$

where the first inequality satisfies from the triangle inequality, the second inequality holds by the smoothness of f. Next we need to upper bound these two terms respectively.

Firstly we need to upper bound $\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\|$. According to Lemma 15, for all $\mathbf{x} \in \mathcal{X}$ and for any $\delta \in (0, 1)$, with probability at least $1 - \frac{\delta}{2}$, we have

$$
\begin{aligned}
&\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} \\
&+ \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log\frac{16\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}} + \frac{d + \log\frac{16\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}\right).
\end{aligned}
\tag{23}
$$

Next we will bound $\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\|$. According to Lemma 11, with probability at least $1 - \frac{\delta}{2}$ we have the following inequalities

$$
\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\| \leq \frac{1}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right). \tag{24}
$$

Finally, we plug (23) and (24) into (22), we obtain the result that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$
\begin{aligned}
&\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\| \leq \frac{\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}); \mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right) \\
&+ \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \\
&\left(\sqrt{\frac{d + \log\frac{16\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}} + \frac{d + \log\frac{16\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}\right).
\end{aligned}
$$

The proof is complete. □

Proof of Theorem 3. Firstly, for all $\mathbf{x} \in \mathcal{X}$, we divide $\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\|$ into two terms which is the same as (22)

$$\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\| \leq \|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| + \beta \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\|. \tag{25}$$

Firstly, we start to bound $\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\|$. According to Lemma 15, applying norm triangle inequality, for any $\delta \in (0, 1)$, with probability at least $1 - \frac{\delta}{2}$, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| - \|\nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2] \log \frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*} \log \frac{8}{\delta}}{n} \\
&+ \frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log \frac{16 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{16 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right).
\end{aligned} \tag{26}$$

Since the population risk $F(\mathbf{x}, \mathbf{y})$ satisfies Assumption 4 with parameter $\mu_{\mathbf{x}}$, according to Lemma 9, $\Phi(\mathbf{x})$ satisfies the PL condition with $\mu_{\mathbf{x}}$, there holds the error bound property (refer to Theorem 2 in (Karimi et al., 2016))

$$\mu_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}^*\| \leq \|\Phi(\mathbf{x})\| = \|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\|. \tag{27}$$

Thus, combing (26) and (27) we have

$$\begin{aligned}
&\mu_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}^*\| \leq \|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
\leq &\frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \max\left\{\|\mathbf{x} - \mathbf{x}^*\|, \frac{1}{n}\right\} \times \left(\sqrt{\frac{d + \log \frac{16 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{16 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right) \\
&+ \|\nabla_{\mathbf{x}} F_S(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| + \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2] \log \frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*} \log \frac{8}{\delta}}{n}.
\end{aligned} \tag{28}$$

On the other hand, according to Lemma 8 with Assumption 1, the function $\Phi(\mathbf{x})$ is $\beta + \frac{\beta^2}{\mu_{\mathbf{y}}}$-smooth in $\mathbf{x} \in \mathcal{X}$. According to (Nesterov, 2003), $\Phi(\mathbf{x})$ holds the following property

$$\frac{1}{2(\beta + \frac{\beta^2}{\mu_{\mathbf{y}}})} \|\nabla\Phi(\mathbf{x})\|^2 \leq \Phi(\mathbf{x}) - \Phi(\mathbf{x}^*).$$

We know that $\Phi(\mathbf{x})$ satisfies the PL condition with $\mu_{\mathbf{x}}$. Thus, we have

$$\frac{1}{2(\beta + \frac{\beta^2}{\mu_{\mathbf{y}}})} \|\nabla\Phi(\mathbf{x})\| \leq \Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \leq \frac{1}{2\mu_{\mathbf{x}}} \|\nabla\Phi(\mathbf{x})\|^2,$$

which means that $\frac{\mu_{\mathbf{x}}\mu_{\mathbf{y}}}{\beta(\mu_{\mathbf{y}} + \beta)} \leq 1$. Then let $c = \max\{16C^2, 1\}$, when

$$n \geq \frac{c\beta^2(\mu_{\mathbf{y}} + \beta)^2 (d + \log \frac{8 \log_2 \sqrt{2} R_1 n + 1}{\delta})}{\mu_{\mathbf{y}}^2 \mu_{\mathbf{x}}^2},$$

we have

$$\frac{C\beta(\mu_{\mathbf{y}} + \beta)}{\mu_{\mathbf{y}}} \left(\sqrt{\frac{d + \log \frac{16 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}} + \frac{d + \log \frac{16 \log_2(\sqrt{2} R_1 n + 1)}{\delta}}{n}\right) \leq \frac{\mu_{\mathbf{x}}}{2}, \tag{29}$$

with the fact that $\frac{\mu_{\mathbf{x}}\mu_{\mathbf{y}}}{\beta(\mu_{\mathbf{y}} + \beta)} \leq 1$.

Next, plugging (29) into (28), we obtain that

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

Then plugging (30) into (26), we derive that for all $\mathbf{x} \in \mathcal{X}$, when $n \geq \frac{c\beta^2(\mu_\mathbf{y}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_\mathbf{y}^2\mu_\mathbf{x}^2}$, with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}\|\nabla_\mathbf{x} F(\mathbf{x},\mathbf{y}^*(\mathbf{x})) - \nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\| &\leq \|\nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\| \\ &+ 2\sqrt{\frac{2\mathbb{E}[\|\nabla_\mathbf{x} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{\mu_\mathbf{x}}{n}.\end{aligned} \tag{30}$$

Next, we need to bound $\|\nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\|$. We make the following decomposition

$$\begin{aligned}&\|\nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\| \\ =&\|\nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x})) - \nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}_S^*(\mathbf{x})) + \nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}_S^*(\mathbf{x}))\| \\ \leq&\|\nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}^*(\mathbf{x})) - \nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}_S^*(\mathbf{x}))\| + \|\nabla_\mathbf{x} F_S(\mathbf{x},\mathbf{y}_S^*(\mathbf{x}))\| \\ \leq&\beta\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\| + \|\nabla\Phi_S(\mathbf{x})\|,\end{aligned} \tag{31}$$

where the first inequality holds with the norm triangle inequality and the second inequality holds with the fact that f is β-smooth.

According to Lemma 11, for any $\mathbf{x} \in \mathcal{X}$, With probability at least $1-\frac{\delta}{2}$ we have the following inequalities

$$\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}_S^*(\mathbf{x})\| \leq \frac{1}{\mu_\mathbf{y}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_\mathbf{y} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right). \tag{32}$$

Combing (30), (31), (25) and (32), we have that for all $\mathbf{x} \in \mathcal{X}$, let $c = \max\{16C^2, 1\}$, when $n \geq \frac{c\beta^2(\mu_\mathbf{y}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_\mathbf{y}^2\mu_\mathbf{x}^2}$, with probability at least $1-\delta$

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

The proof is complete. □

Proof of Remark 5. Here we briefly prove the results given in Remark 5. Since $\Phi(\mathbf{x})$ satisfies the PL condition with $\mu_\mathbf{x}$, we have we have

$$\Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \leq \frac{\|\nabla\Phi(\mathbf{x})\|^2}{2\mu_\mathbf{x}}. \tag{33}$$

Therefore, we need to bound $\|\nabla\Phi(\mathbf{x})\|^2$. According to Theorem 3, for any $\delta \in (0,1)$, when $n \geq \frac{c\beta^2(\mu_\mathbf{y}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_\mathbf{y}^2\mu_\mathbf{x}^2}$, with probability at least $1-\delta$,

$$\begin{aligned}\|\nabla\Phi(\mathbf{x})\| &\leq 2\|\nabla\Phi_S(\mathbf{x})\| + 2\sqrt{\frac{2\mathbb{E}[\|\nabla_\mathbf{x} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} \\ &+ \frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{\mu_\mathbf{x}}{n} + \frac{2\beta}{\mu_\mathbf{y}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_\mathbf{y} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right).\end{aligned} \tag{34}$$

Then, substituting (34) into (33), we have

$$\begin{aligned}
&\Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \\
\leq &\frac{\|\nabla\Phi(\mathbf{x})\|^2}{2\mu_{\mathbf{x}}} \\
\leq &\frac{1}{2\mu_{\mathbf{x}}}\Bigg\{2\|\nabla\Phi_S(\mathbf{x})\| + \frac{2\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right) \\
&\quad + 2\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{\mu_{\mathbf{x}}}{n}\Bigg\}^2 \\
\leq &\frac{8\|\nabla\Phi_S(\mathbf{x})\|^2}{\mu_{\mathbf{x}}} + \frac{16\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} \\
&\quad + \frac{16\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} + \frac{2\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2},
\end{aligned}$$

where the second inequality holds with Cauchy–Bunyakovsky–Schwarz inequality.

The proof is complete. □

D Application

D.1 Empirical saddle point

Proof of Theorem 4. Plugging $\hat{\mathbf{x}}^*$ into Theorem 1, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\|\nabla\Phi(\hat{\mathbf{x}}^*)\| - \|\nabla\Phi_S(\hat{\mathbf{x}}^*)\| \leq \frac{\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right) \\
&+ \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x}-\mathbf{x}^*\|, \frac{1}{n}\right\} \\
&\times\left(\sqrt{\frac{d+\log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}} + \frac{d+\log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}\right).
\end{aligned}$$

Since $\hat{\mathbf{x}}^*$ is the solution of (2), there holds that $\|\nabla\Phi_S(\hat{\mathbf{x}}^*)\| = 0$. Thus, we can derive that

$$\begin{aligned}
\|\nabla\Phi(\hat{\mathbf{x}}^*)\| \leq &\frac{\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right) \\
&+ \sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\left(R_1 + \frac{1}{n}\right) \\
&\times\left(\sqrt{\frac{d+\log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}} + \frac{d+\log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}\right) \\
= &\, O\left(\sqrt{\frac{d+\log\frac{\log n}{\delta}}{n}}\right).
\end{aligned}$$

The proof is complete. □

Proof of Theorem 5. According to Lemma 9, $\Phi(\mathbf{x})$ satisfies the PL condition with $\mu_{\mathbf{x}}$, we have

$$\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*) \leq \frac{\|\nabla\Phi(\hat{\mathbf{x}}^*)\|^2}{2\mu_{\mathbf{x}}}. \tag{35}$$

Therefore, we need to bound $\|\nabla\Phi(\hat{\mathbf{x}}^*)\|^2$. Plugging $\hat{\mathbf{x}}^*$ into Theorem 3, for any $\delta \in (0,1)$, when $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, with probability at least $1-\delta$

$$\begin{aligned}\|\nabla\Phi(\hat{\mathbf{x}}^*)\| \leq\ & 2\|\nabla\Phi_S(\hat{\mathbf{x}}^*)\| + 2\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} \\ &+ \frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{\mu_{\mathbf{x}}}{n} + \frac{2\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right).\end{aligned} \tag{36}$$

Since $\nabla\Phi_S(\hat{\mathbf{x}}^*) = \nabla_{\mathbf{x}} F_S(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*) = \mathbf{0}$, we have $\|\nabla\Phi_S(\hat{\mathbf{x}}^*)\| = \|\nabla_{\mathbf{x}} F_S(\hat{\mathbf{x}}^*,\hat{\mathbf{y}}^*)\| = 0$. By plugging (36) into (35), we have

$$\begin{aligned}&\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*) \\ \leq&\frac{\|\nabla\Phi(\hat{\mathbf{x}}^*)\|^2}{2\mu_{\mathbf{x}}} \\ \leq&\frac{1}{2\mu_{\mathbf{x}}}\left\{\frac{2\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right)\right. \\ &\left. + 2\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{\mu_{\mathbf{x}}}{n}\right\}^2 \\ \leq&\frac{12\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})\|^2]\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{12\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} \\ &+ \frac{3\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{2\mu_{\mathbf{x}} n^2},\end{aligned}$$

where the second inequality holds with Cauchy–Bunyakovsky–Schwarz inequality.

Next, if we further assume $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, according to Lemma 7, we have $\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2 \leq 4\beta\mathbb{E}[f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})]$ and $\mathbb{E}\|\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})\|^2 \leq 4\beta\mathbb{E}[f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})]$. Then we have

$$\begin{aligned}\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*) \leq&\frac{48\beta^3\mathbb{E}[f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})]\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{48\beta\mathbb{E}[f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})]\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} \\ &+ \frac{3\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{2\mu_{\mathbf{x}} n^2} \\ =&\frac{48\beta^3\Phi(\hat{\mathbf{x}}^*)\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{48\beta\Phi(\mathbf{x}^*)\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} + \frac{3\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{2\mu_{\mathbf{x}} n^2},\end{aligned}$$

which implies that

$$\begin{aligned}&\left(1 - \frac{48\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n}\right)(\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*)) \\ \leq&\frac{48\beta\Phi(\mathbf{x}^*)\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} + \frac{48\beta^3\Phi(\mathbf{x}^*)\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{3\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{2\mu_{\mathbf{x}} n^2}.\end{aligned}$$

When $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \frac{48\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right\}$, we have

$$\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*) = O\left(\frac{\Phi(\mathbf{x}^*)\log\frac{1}{\delta}}{n - \frac{48\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}} + \frac{\log^2\frac{1}{\delta}}{n\left(n - \frac{48\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)}\right).$$

The proof is complete. □

D.2 GRADIENT DESCENT ASCENT

Firstly, we introduce the optimization error bound for GDA given in (Lin et al., 2020).

Lemma 16 (Optimization error bound for NC-SC minimax problems (Lin et al., 2020))**.** *under Assumption 1, and letting the step sizes be chosen as* $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ *and* $\eta_{\mathbf{y}} = \frac{1}{\beta}$*, then the optimization error bound of Algorithm 1 can be bounded by*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\Phi_S(\mathbf{x}_t)\|^2 \leq \frac{128\beta^3\Delta_\Phi}{\mu_{\mathbf{y}}^2T} + \frac{5\beta^3D_{\mathcal{Y}}}{\mu_{\mathbf{y}}T},$$

where $\Delta_\Phi = \Phi_S(\mathbf{x}_0) - \min_{\mathbf{x}}\Phi_S(\mathbf{x})$.

Proof of Theorem 6. Firstly, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2 \leq& \frac{2}{T}\sum_{t=1}^{T}\|\nabla\Phi(\mathbf{x}_t) - \nabla_{\mathbf{x}}\Phi_S(\mathbf{x}_t)\|^2 + \frac{2}{T}\sum_{t=1}^{T}\|\nabla_{\mathbf{x}}\Phi_S(\mathbf{x}_t)\|^2 \\
\leq& \frac{2}{T}\sum_{t=1}^{T}\max_{t=1,\cdots,T}\|\nabla\Phi(\mathbf{x}_t) - \nabla_{\mathbf{x}}\Phi_S(\mathbf{x}_t)\|^2 + \frac{2}{T}\sum_{t=1}^{T}\|\nabla_{\mathbf{x}}\Phi_S(\mathbf{x}_t)\|^2 \\
\leq& \frac{2}{T}\sum_{t=1}^{T}\max_{t=1,\cdots,T}\|\nabla\Phi(\mathbf{x}_t) - \nabla_{\mathbf{x}}\Phi_S(\mathbf{x}_t)\|^2 + \frac{256\beta^3\Delta_\Phi}{\mu_{\mathbf{y}}^2T} + \frac{10\beta^3D_{\mathcal{Y}}}{\mu_{\mathbf{y}}T},
\end{aligned} \tag{37}$$

Thus, with probability at least $1-\delta$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2 \leq& O\left(\frac{1}{T}\right) + O\Bigg(\max_{t=1,\cdots,T}\Bigg[\frac{C\beta(\mu_{\mathbf{y}}+\beta)}{\mu_{\mathbf{y}}}\max\left\{\|\mathbf{x}_t - \mathbf{x}^*\|, \frac{1}{n}\right\} \\
&\times\left(\sqrt{\frac{d+\log\frac{16\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{16\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right)\Bigg]^2\Bigg).
\end{aligned} \tag{38}$$

where the inequality holds according to Theorem 1.

Next, we need to bound $\|\mathbf{x}_t - \mathbf{x}^*\|$. Since we assume that $\mathbf{x}_1 = 0$, and $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_{\mathbf{x}}\nabla_{\mathbf{x}}F_S(\mathbf{x}_t, \mathbf{y}_t)$, we have $\mathbf{x}_{t+1} = -\eta_{\mathbf{x}}\sum_{k=1}^{t}\nabla_{\mathbf{x}}F_S(\mathbf{x}_t, \mathbf{y}_t)$. then we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}^*\| \leq \|\mathbf{x}_{t+1}\| + \|\mathbf{x}^*\| \\
\leq& \left\|\eta_{\mathbf{x}}\sum_{k=1}^{t}\nabla_{\mathbf{x}}F_S(\mathbf{x}_t, \mathbf{y}_t)\right\| + \|\mathbf{x}^*\| = O\left(\eta_{\mathbf{x}}Lt\right).
\end{aligned} \tag{39}$$

Then plugging (39) into (38), with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\Phi(\mathbf{x}_t)\|^2 \leq O\left(\frac{1}{T}\right) + O\left(\frac{d+\log\frac{16\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}T\right).$$

Let $T \asymp O\left(\sqrt{\frac{n}{d}}\right)$, we can derive that

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

The proof is complete. □

Proof of Theorem 7. According to Lemma 9, $\Phi(\mathbf{x})$ satisfies the PL assumption with parameter $\mu_{\mathbf{x}}$, we have

$$\Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \leq \frac{\|\nabla\Phi(\mathbf{x})\|^2}{2\mu_{\mathbf{x}}}. \tag{40}$$

To bound $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$, we firstly need to bound the term $\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2$. There holds that

$$\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2 \leq 2\|\nabla\Phi(\bar{\mathbf{x}}_T) - \nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2 + 2\|\nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2. \tag{41}$$

From Theorem 3, under Assumption 3 and 5, plugging $\bar{\mathbf{x}}_T$ into Theorem 3, for any $\delta \in (0,1)$, when $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, with probability at least $1-\delta$

$$\begin{aligned}\|\nabla\Phi(\bar{\mathbf{x}}_T) - \nabla\Phi_S(\bar{\mathbf{x}}_T)\| \leq \|\nabla\Phi_S(\bar{\mathbf{x}}_T)\| + 2\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*(\mathbf{x}^*); \mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} \\ + \frac{2B_{\mathbf{x}^*}\log\frac{8}{\delta}}{n} + \frac{\mu_{\mathbf{x}}}{n} + \frac{2\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T, \mathbf{y}^*(\bar{\mathbf{x}}_T); \mathbf{z})\|^2]\log\frac{4}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{4}{\delta}}{n}\right).\end{aligned} \tag{42}$$

Next, we need to bound the optimization error $\|\nabla\Phi_S(\bar{\mathbf{x}}_T)\|$. According to Lemma 9, $\Phi_S(\mathbf{x})$ satisfies the PL assumption with parameter $\mu_{\mathbf{x}}$, then using Lemma 16 we have

$$\frac{1}{T}\sum_{t=1}^{T}\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}^*) \leq \frac{1}{2\mu_{\mathbf{x}}T}\sum_{t=1}^{T}\|\nabla\Phi_S(\mathbf{x}_t)\|^2 \leq \frac{64\beta^3\Delta_\Phi}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 T} + \frac{5\beta^3 D_{\mathcal{Y}}}{2\mu_{\mathbf{x}}\mu_{\mathbf{y}}T}.$$

From the convexity of $F_S(\cdot, \mathbf{y})$, we get

$$\Phi_S(\bar{\mathbf{x}}_T) - \Phi_S(\mathbf{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\Phi_S(\mathbf{x}_t) - \Phi_S(\mathbf{x}^*) \leq \frac{64\beta^3\Delta_\Phi}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 T} + \frac{5\beta^3 D_{\mathcal{Y}}}{2\mu_{\mathbf{x}}\mu_{\mathbf{y}}T}.$$

According to (Nesterov, 2003) and Lemma 8, there holds the following property for $\beta + \frac{\beta^2}{\mu_{\mathbf{y}}}$ function $\Phi_S(\mathbf{x})$, we have

$$\frac{1}{2\left(\beta + \frac{\beta^2}{\mu_{\mathbf{y}}}\right)}\|\nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2 \leq \Phi_S(\bar{\mathbf{x}}_T) - \Phi_S(\mathbf{x}^*) \leq \frac{64\beta^3\Delta_\Phi}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 T} + \frac{5\beta^3 D_{\mathcal{Y}}}{2\mu_{\mathbf{x}}\mu_{\mathbf{y}}T}. \tag{43}$$

Plugging (43) into (42), according to Cauchy–Bunyakovsky–Schwarz inequality, we can derive that

$$\begin{aligned}\|\nabla\Phi(\bar{\mathbf{x}}_T) - \nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2 \leq O\left(\frac{1}{T}\right) + \frac{32\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T, \mathbf{y}^*(\bar{\mathbf{x}}_T); \mathbf{z})\|^2]\log\frac{4}{\delta}}{\mu_{\mathbf{y}}^2 n} \\ + \frac{32\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2]\log\frac{8}{\delta}}{n} + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{n^2}.\end{aligned} \tag{44}$$

Then substituting (44), (43) into (41), we derive that

$$\begin{aligned}\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2 \leq O\left(\frac{1}{T}\right) + \frac{64\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T, \mathbf{y}^*(\bar{\mathbf{x}}_T); \mathbf{z})\|^2]\log\frac{4}{\delta}}{\mu_{\mathbf{y}}^2 n} \\ + \frac{64\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2]\log\frac{8}{\delta}}{n} + \frac{8\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{n^2}.\end{aligned} \tag{45}$$

Finally, we plug (45) into (40) and choose $T \asymp O(n)$, with probability at least $1-\delta$

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

Next, if we further assume $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, According to Lemma 7, we have $\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2 \leq 4\beta\mathbb{E}[f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})]$ and $\mathbb{E}\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T, \mathbf{y}^*(\bar{\mathbf{x}}_T); \mathbf{z})\|^2 \leq 4\beta\mathbb{E}[f(\bar{\mathbf{x}}_T, \mathbf{y}^*(\bar{\mathbf{x}}_T); \mathbf{z})]$. Plugging (45) into (40), we have

$$
\begin{aligned}
&\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) \\
\leq &\frac{\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2}{2\mu_{\mathbf{x}}} \\
\leq &O\left(\frac{1}{T}\right) + \frac{128\beta^3\mathbb{E}[f(\hat{\mathbf{x}}^*, \mathbf{y}^*(\hat{\mathbf{x}}^*); \mathbf{z})]\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{128\beta\mathbb{E}[f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})]\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} \\
&+ \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2} \\
= &O\left(\frac{1}{T}\right) + \frac{128\beta^3\Phi(\hat{\mathbf{x}}^*)\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{128\beta\Phi(\mathbf{x}^*)\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2},
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&\left(1 - \frac{128\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n}\right)(\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)) \\
= &O\left(\frac{1}{T}\right) + \frac{128\beta\Phi(\mathbf{x}^*)\log\frac{8}{\delta}}{\mu_{\mathbf{x}} n} + \frac{128\beta^3\Phi(\mathbf{x}^*)\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{4}{\delta} + 2B_{\mathbf{x}^*}\log\frac{8}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2}.
\end{aligned}
$$

When $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \frac{128\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right\}$, and $T \asymp n^2$ we have

$$
\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\Phi(\mathbf{x}^*)\log\frac{1}{\delta}}{n - \frac{128\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}} + \frac{\log^2\frac{1}{\delta}}{n\left(n - \frac{128\beta^3\log\frac{4}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)}\right).
$$

The proof is complete. □

D.3 Stochastic gradient descent ascent

In this subsection, we present empirical optimization error bounds of primal functions for SGDA, which are motivated by (Lei et al., 2021; Li & Liu, 2021a). The proofs are standard in the literature (Nedić & Ozdaglar, 2009; Nemirovski et al., 2009) and we give the optimization error bounds with high probability. Firstly, we introduce two concentration inequalities for martingales.

Lemma 17 ((Boucheron et al., 2013)). *Let $z_1, \ldots, z_n$ be a sequence of random variables such that z_k may depend the previous variables $z_1, \ldots, z_{k-1}$ for all $k = 1, \ldots, n$. Consider a sequence of functionals $\xi_k(z_1, \ldots, z_k), k = 1, \ldots, n$. Assume $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b_k$ for each k. Let $\delta \in (0, 1)$. With probability at least $1 - \delta$*

$$
\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \leq \left(2\sum_{k=1}^n b_k^2 \log\frac{1}{\delta}\right)^{\frac{1}{2}}.
$$

Lemma 18 ((Tarres & Yao, 2014)). *Let $\{\xi_k\}_{k\in\mathbb{N}}$ be a martingale difference sequence in $\mathbb{R}^d$. Suppose that almost surely $\|\xi_k\| \leq D$ and $\sum_{k=1}^t \mathbb{E}[\|\xi_k\|^2|\xi_1, \ldots, \xi_{k-1}] \leq \sigma_t^2$. Then, for any $\delta \in (0, 1)$, the following inequality holds with probability at least $1 - \delta$*

$$
\max_{1\leq j\leq t}\left\|\sum_{k=1}^j \xi_k\right\| \leq 2\left(\frac{D}{3} + \sigma_t\right)\log\frac{2}{\delta}.
$$

The following lemma shows the optimization error bounds of primal function for SGDA.

Lemma 19. *Suppose Assumption 5 and 2 hold and let the stepsizes be 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* $1-\delta$*, then the optimization error of Algorithm 2 can be bounded by*

$$\Phi_S(\bar{\mathbf{x}}_T) - \Phi_S(\hat{\mathbf{x}}^*) \leq \frac{t_0(\mu_{\mathbf{x}} D_{\mathbf{x}} + \mu_{\mathbf{y}} D_{\mathbf{y}})}{2T} + \frac{L^2 \log(eT)}{2T}\left(\frac{1}{\mu_{\mathbf{x}}} + \frac{1}{\mu_{\mathbf{y}}}\right) + \frac{2(\sqrt{D_{\mathcal{X}}} + \sqrt{D_{\mathcal{Y}}})}{T}\left(\frac{2L}{3} + 2L\sqrt{T}\right)\log\frac{6}{\delta} + \frac{2L(\sqrt{D_{\mathcal{X}}} + \sqrt{D_{\mathcal{Y}}})\left(2T\log\frac{6}{\delta}\right)^{\frac{1}{2}}}{T}.$$

Proof of Lemma 19. This proof mainly follows from (Lei et al., 2021; Li & Liu, 2021a). Firstly, we have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 &= \|\mathbf{x}_t - \eta_{\mathbf{x}_t}\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \mathbf{x}\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 + \eta_{\mathbf{x}_t}^2 \|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\|^2 + 2\eta_{\mathbf{x}_t}\langle \mathbf{x} - \mathbf{x}_t, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle \\
&\leq \|\mathbf{x}_t - \mathbf{x}\|^2 + \eta_{\mathbf{x}_t}^2 L^2 + 2\eta_{\mathbf{x}_t}\langle \mathbf{x} - \mathbf{x}_t, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle \\
&\quad + 2\eta_{\mathbf{x}_t}\langle \mathbf{x} - \mathbf{x}_t, \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle,
\end{aligned}$$

where the first inequality holds because of Assumption 2. According to the strong convexity of $F_S(\cdot, \mathbf{y}_t)$, we have

$$\begin{aligned}
2\eta_{\mathbf{x}_t}(F_S(\mathbf{x}_t, \mathbf{y}_t) - F_S(\mathbf{x}, \mathbf{y}_t)) \leq (1 - \eta_{\mathbf{x}_t}\mu_{\mathbf{x}})\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 + \eta_{\mathbf{x}_t}^2 L^2 \\
+ 2\eta_{\mathbf{x}_t}\langle \mathbf{x} - \mathbf{x}_t, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle.
\end{aligned}$$

Let $\eta_{\mathbf{x}_t} = \frac{1}{\mu_{\mathbf{x}}(t+t_0)}$, we further get

$$\begin{aligned}
\frac{2}{\mu_{\mathbf{x}}(t+t_0)}(F_S(\mathbf{x}_t, \mathbf{y}_t) - F_S(\mathbf{x}, \mathbf{y}_t)) &\leq \left(1 - \frac{1}{t+t_0}\right)\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \\
&\quad + \left(\frac{L}{\mu_{\mathbf{x}}(t+t_0)}\right)^2 + \frac{2}{\mu_{\mathbf{x}}(t+t_0)}\langle \mathbf{x} - \mathbf{x}_t, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle.
\end{aligned}$$

Multiplying both sides by $t + t_0$, we have

$$\begin{aligned}
\frac{2}{\mu_{\mathbf{x}}}(F_S(\mathbf{x}_t, \mathbf{y}_t) - F_S(\mathbf{x}, \mathbf{y}_t)) &\leq (t + t_0 - 1)\|\mathbf{x}_t - \mathbf{x}\|^2 - (t+t_0)\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \\
&\quad + \frac{L^2}{\mu_{\mathbf{x}}^2(t+t_0)} + \frac{2}{\mu_{\mathbf{x}}}\langle \mathbf{x} - \mathbf{x}_t, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle.
\end{aligned}$$

Since $\mathbf{x}_1 = 0$ and $\sum_{t=1}^T t^{-1} \leq \log(eT)$, by taking a summation of the above inequality from $t = 1$ to T, we have

$$\begin{aligned}
\sum_{t=1}^T (F_S(\mathbf{x}_t, \mathbf{y}_t) - F_S(\mathbf{x}, \mathbf{y}_t)) &\leq \frac{\mu_{\mathbf{x}}}{2} t_0 D_{\mathcal{X}} + \frac{L^2 \log(eT)}{2\mu_{\mathbf{x}}} \\
&+ \sum_{t=1}^T \langle \mathbf{x}, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + \sum_{t=1}^T \langle \mathbf{x}_t, \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle.
\end{aligned}$$

Since $\mathbf{y}_S^*(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} F_S(\mathbf{x}, \mathbf{y})$, for any $\mathbf{x} \in \mathcal{X}$, we obtain that $F_S(\mathbf{x}, \mathbf{y}_t) \leq F_S(\mathbf{x}, \mathbf{y}_S^*(\mathbf{x}))$. Then we have

$$\begin{aligned}
\sum_{t=1}^T (F_S(\mathbf{x}_t, \mathbf{y}_t) - F_S(\mathbf{x}, \mathbf{y}_S^*(\mathbf{x}))) &\leq \frac{\mu_{\mathbf{x}}}{2} t_0 D_{\mathcal{X}} + \frac{L^2 \log(eT)}{2\mu_{\mathbf{x}}} \\
&+ \sum_{t=1}^T \langle \mathbf{x}, \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + \sum_{t=1}^T \langle \mathbf{x}_t, \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle.
\end{aligned}$$

Since this inequality holds for any $\mathbf{x}$, we get

$$\sum_{t=1}^{T}(F_S(\mathbf{x}_t,\mathbf{y}_t)-\inf_{\mathbf{x}\in\mathcal{X}}F_S(\mathbf{x},\mathbf{y}_S^*(\mathbf{x})))\leq\frac{\mu_{\mathbf{x}}}{2}t_0D_{\mathcal{X}}+\frac{L^2\log(eT)}{2\mu_{\mathbf{x}}}$$
$$+\sum_{t=1}^{T}\sup_{x\in\mathcal{X}}\langle\mathbf{x},\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})-\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)\rangle+\sum_{t=1}^{T}\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})\rangle,$$

which implies that

$$\sum_{t=1}^{T}(F_S(\mathbf{x}_t,\mathbf{y}_t)-\Phi_S(\hat{\mathbf{x}}^*))\leq\frac{\mu_{\mathbf{x}}}{2}t_0D_{\mathcal{X}}+\frac{L^2\log(eT)}{2\mu_{\mathbf{x}}}$$
$$+\sum_{t=1}^{T}\sup_{x\in\mathcal{X}}\langle\mathbf{x},\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})-\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)\rangle+\sum_{t=1}^{T}\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})\rangle.$$

By Schwarz's inequality, we have

$$\sum_{t=1}^{T}(F_S(\mathbf{x}_t,\mathbf{y}_t)-\Phi_S(\hat{\mathbf{x}}^*))\leq\frac{\mu_{\mathbf{x}}}{2}t_0D_{\mathcal{X}}+\frac{L^2\log(eT)}{2\mu_{\mathbf{x}}}$$
$$+\sqrt{D_{\mathcal{X}}}\left\|\sum_{t=1}^{T}(\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})-\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t))\right\|+\sum_{t=1}^{T}\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})\rangle.$$

Denote that $\xi_t=\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{x}_t;\mathbf{z}_{i_t})\rangle$. Since $\mathbb{E}_{i_t}[\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})\rangle]=0$, so $\{\xi_t|t=1,\ldots,T\}$ is a martingale difference sequence. By Schwarz's inequality and Assumption 2, we know that $|\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})\rangle|\leq 2L\sqrt{D_{\mathcal{X}}}$. Then according to Lemma 17, we have the following inequality with probability at least $1-\frac{\delta}{6}$

$$\sum_{t=1}^{T}\langle\mathbf{x}_t,\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})\rangle\leq 2L\sqrt{D_{\mathcal{X}}}\left(2T\log\frac{6}{\delta}\right)^{\frac{1}{2}}.$$

Define $\xi_t'=\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})-\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t)$. Then we get $\|\xi_t'\|\leq 2L$ and

$$\sum_{t=1}^{T}\mathbb{E}[\|\xi_t'\|^2|\xi_1',\ldots,\xi_{t-1}']\leq 4TL^2.$$

Applying Lemma 18 to the martingale difference sequence $\{\xi_t'\}$, we have the following inequality with probability at least $1-\frac{\delta}{3}$

$$\left\|\sum_{t=1}^{T}\xi_t'\right\|\leq 2\left(\frac{2L}{3}+2L\sqrt{T}\right)\log\frac{6}{\delta}.$$

This implies that with probability at least $1-\frac{\delta}{3}$

$$\left\|\sum_{t=1}^{T}(\nabla_{\mathbf{x}}f(\mathbf{x}_t,\mathbf{y}_t;\mathbf{z}_{i_t})-\nabla_{\mathbf{x}}F_S(\mathbf{x}_t,\mathbf{y}_t))\right\|\leq 2\left(\frac{2L}{3}+2L\sqrt{T}\right)\log\frac{6}{\delta}.$$

Combined with the above results, we finally have the following inequality with probability at least $1-\frac{\delta}{2}$

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(F_S(\mathbf{x}_t,\mathbf{y}_t)-\Phi_S(\hat{\mathbf{x}}^*))&\leq\frac{\mu_{\mathbf{x}}t_0D_{\mathcal{X}}}{2T}+\frac{L^2\log(eT)}{2\mu_{\mathbf{x}}T}\\&+\frac{2\sqrt{D_{\mathcal{X}}}}{T}\left(\frac{2L}{3}+2L\sqrt{T}\right)\log\frac{6}{\delta}+\frac{2L\sqrt{D_{\mathcal{X}}}\left(2T\log\frac{6}{\delta}\right)^{\frac{1}{2}}}{T}.\end{aligned}\tag{46}$$

Similarly, we can bound $\Phi(\bar{\mathbf{x}}_T) - \frac{1}{T}\sum_{t=1}^{T} F_S(\mathbf{x}_t, \mathbf{y}_t)$. Firstly, we have

$$\begin{aligned}\|\mathbf{y}_{t+1} - \mathbf{y}\|^2 &= \|\mathbf{y}_t + \eta_{\mathbf{y}_t}\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \mathbf{y}\|^2 \\ &= \|\mathbf{y}_t - \mathbf{y}\|^2 + \eta_{\mathbf{y}_t}^2 \|\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\|^2 + 2\eta_{\mathbf{x}_y}\langle \mathbf{y}_t - \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle \\ &\leq \|\mathbf{y}_t - \mathbf{y}\|^2 + \eta_{\mathbf{y}_t}^2 L^2 + 2\eta_{\mathbf{y}_t}\langle \mathbf{y}_t - \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle \\ &\quad + 2\eta_{\mathbf{y}_t}\langle \mathbf{y}_t - \mathbf{y}, \nabla_{\mathbf{x}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle,\end{aligned}$$

where the first inequality holds because of Assumption 2. According to the strong concavity of $F_S(\mathbf{x}_t, \cdot)$, we have

$$\begin{aligned}2\eta_{\mathbf{y}_t}(F_S(\mathbf{x}_t, \mathbf{y}) - F_S(\mathbf{x}_t, \mathbf{y}_t)) \leq (1 - \eta_{\mathbf{y}_t}\mu_{\mathbf{y}})\|\mathbf{y}_t - \mathbf{y}\|^2 - \|\mathbf{y}_{t+1} - \mathbf{y}\|^2 + \eta_{\mathbf{y}_t}^2 L^2 \\ + 2\eta_{\mathbf{y}_t}\langle \mathbf{y}_t - \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle.\end{aligned}$$

Let $\eta_{\mathbf{y}_t} = \frac{1}{\mu_{\mathbf{y}}(t+t_0)}$, we further get

$$\begin{aligned}\frac{2}{\mu_{\mathbf{y}}(t+t_0)}(F_S(\mathbf{x}_t, \mathbf{y}) - F_S(\mathbf{x}_t, \mathbf{y}_t)) &\leq \left(1 - \frac{1}{t+t_0}\right)\|\mathbf{y}_t - \mathbf{y}\|^2 - \|\mathbf{y}_{t+1} - \mathbf{y}\|^2 \\ &\quad + \left(\frac{L}{\mu_{\mathbf{y}}(t+t_0)}\right)^2 + \frac{2}{\mu_{\mathbf{y}}(t+t_0)}\langle \mathbf{y}_t - \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle.\end{aligned}$$

Multiplying both sides by $t + t_0$, we have

$$\begin{aligned}\frac{2}{\mu_{\mathbf{y}}}(F_S(\mathbf{x}_t, \mathbf{y}) - F_S(\mathbf{x}_t, \mathbf{y}_t)) &\leq (t + t_0 - 1)\|\mathbf{y}_t - \mathbf{y}\|^2 - (t+t_0)\|\mathbf{y}_{t+1} - \mathbf{y}\|^2 \\ &\quad + \frac{L^2}{\mu_{\mathbf{y}}^2(t+t_0)} + \frac{2}{\mu_{\mathbf{y}}}\langle \mathbf{y}_t - \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle.\end{aligned}$$

Since $\mathbf{y}_1 = 0$ and $\sum_{t=1}^{T} t^{-1} \leq \log(eT)$, by taking a summation of the above inequality from $t = 1$ to T, we have

$$\begin{aligned}&\sum_{t=1}^{T}(F_S(\mathbf{x}_t, \mathbf{y}) - F_S(\mathbf{x}_t, \mathbf{y}_t)) \leq \frac{\mu_{\mathbf{y}}}{2}t_0 D_{\mathcal{Y}} + \frac{L^2\log(eT)}{2\mu_{\mathbf{y}}} \\ &+ \sum_{t=1}^{T}\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + \sum_{t=1}^{T}\langle \mathbf{y}, \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle.\end{aligned}$$

From the convexity of $F_S(\cdot, \mathbf{y})$, we get

$$\begin{aligned}&\sum_{t=1}^{T}(F_S(\bar{\mathbf{x}}_T, \mathbf{y}) - F_S(\mathbf{x}_t, \mathbf{y}_t)) \leq \frac{\mu_{\mathbf{y}}}{2}t_0 D_{\mathcal{Y}} + \frac{L^2\log(eT)}{2\mu_{\mathbf{y}}} \\ &+ \sum_{t=1}^{T}\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + \sum_{t=1}^{T}\langle \mathbf{y}, \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle.\end{aligned}$$

Since this inequality holds for any $\mathbf{y}$, we get

$$\begin{aligned}&\sum_{t=1}^{T}(\sup_{y\in\mathcal{Y}} F_S(\bar{\mathbf{x}}_T, \mathbf{y}) - F_S(\mathbf{x}_t, \mathbf{y}_t)) \leq \frac{\mu_{\mathbf{y}}}{2}t_0 D_{\mathcal{Y}} + \frac{L^2\log(eT)}{2\mu_{\mathbf{y}}} \\ &+ \sum_{t=1}^{T}\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + \sum_{t=1}^{T}\sup_{y\in\mathcal{Y}}\langle \mathbf{y}, \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle,\end{aligned}$$

which implies that

$$\begin{aligned}&\sum_{t=1}^{T}(\Phi_S(\bar{\mathbf{x}}_T) - F_S(\mathbf{x}_t, \mathbf{y}_t)) \leq \frac{\mu_{\mathbf{y}}}{2}t_0 D_{\mathcal{Y}} + \frac{L^2\log(eT)}{2\mu_{\mathbf{y}}} \\ &+ \sum_{t=1}^{T}\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + \sum_{t=1}^{T}\sup_{y\in\mathcal{Y}}\langle \mathbf{y}, \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})\rangle.\end{aligned}$$

By Schwarz's inequality, we have

$$\sum_{t=1}^{T}(\Phi_S(\bar{\mathbf{x}}_T) - F_S(\mathbf{x}_t, \mathbf{y}_t)) \leq \frac{\mu_{\mathbf{y}}}{2} t_0 D_{\mathcal{Y}} + \frac{L^2 \log(eT)}{2\mu_{\mathbf{y}}}$$
$$+ \sum_{t=1}^{T}\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle + D_{\mathcal{Y}} \left\| \sum_{t=1}^{T}(\nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})) \right\|.$$

Denote that $\tilde{\xi}_t = \langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{x}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle$. Since $\mathbb{E}_{i_t}[\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{x}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle] = 0$, so $\{\tilde{\xi}_t | t = 1, \ldots, T\}$ is a martingale difference sequence. By Schwarz's inequality and Assumption 2, we know that $|\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{x}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle| \leq 2L\sqrt{D_{\mathcal{Y}}}$. Then according to Lemma 17 we have the following inequality with probability at least $1 - \frac{\delta}{6}$

$$\sum_{t=1}^{T}\langle \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{x}_t; \mathbf{z}_{i_t}) - \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t)\rangle \leq 2L\sqrt{D_{\mathcal{Y}}} \left(2T \log \frac{6}{\delta}\right)^{\frac{1}{2}}.$$

Define $\tilde{\xi}'_t = \nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})$. Then we get $\|\tilde{\xi}'_t\| \leq 2L$ and

$$\sum_{t=1}^{T} \mathbb{E}[\|\tilde{\xi}'_t\|^2 | \tilde{\xi}'_1, \ldots, \tilde{\xi}'_{t-1}] \leq 4TL^2.$$

Applying Lemma 18 to the martingale difference sequence $\{\tilde{\xi}'_t\}$, we have the following inequality with probability at least $1 - \frac{\delta}{3}$

$$\left\| \sum_{t=1}^{T} \tilde{\xi}'_t \right\| \leq 2\left(\frac{2L}{3} + 2L\sqrt{T}\right) \log \frac{6}{\delta}.$$

This implies that with probability at least $1 - \frac{\delta}{3}$

$$\left\| \sum_{t=1}^{T} (\nabla_{\mathbf{y}} F_S(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \mathbf{z}_{i_t})) \right\| \leq 2\left(\frac{2L}{3} + 2L\sqrt{T}\right) \log \frac{6}{\delta}.$$

Combined with the above results, we finally have the following inequality with probability at least $1 - \frac{\delta}{2}$

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}(\Phi_S(\bar{\mathbf{x}}_T) - F_S(\mathbf{x}_t, \mathbf{y}_t)) &\leq \frac{\mu_{\mathbf{y}} t_0 D_{\mathcal{Y}}}{2T} + \frac{L^2 \log(eT)}{2\mu_{\mathbf{y}} T} \\ &+ \frac{2\sqrt{D_{\mathcal{Y}}}}{T}\left(\frac{2L}{3} + 2L\sqrt{T}\right)\log\frac{6}{\delta} + \frac{2L\sqrt{D_{\mathcal{Y}}}\left(2T\log\frac{6}{\delta}\right)^{\frac{1}{2}}}{T}. \end{aligned} \tag{47}$$

Combing (46) and (47) together, with probability at least $1 - \delta$ we get the following inequality

$$\begin{aligned} \Phi_S(\bar{\mathbf{x}}_T) - \Phi_S(\hat{\mathbf{x}}^*) &\leq \frac{t_0(\mu_{\mathbf{x}} D_{\mathbf{x}} + \mu_{\mathbf{y}} D_{\mathbf{y}})}{2T} + \frac{L^2 \log(eT)}{2T}\left(\frac{1}{\mu_{\mathbf{x}}} + \frac{1}{\mu_{\mathbf{y}}}\right) \\ &+ \frac{2(\sqrt{D_{\mathcal{X}}} + \sqrt{D_{\mathcal{Y}}})}{T}\left(\frac{2L}{3} + 2L\sqrt{T}\right)\log\frac{6}{\delta} + \frac{2L(\sqrt{D_{\mathcal{X}}} + \sqrt{D_{\mathcal{Y}}})\left(2T\log\frac{6}{\delta}\right)^{\frac{1}{2}}}{T}. \end{aligned}$$

□

Proof of Theorem 8. According to Lemma 9, $\Phi(\mathbf{x})$ satisfies the PL assumption with parameter $\mu_{\mathbf{x}}$, we have

$$\Phi(\mathbf{x}) - \Phi(\mathbf{x}^*) \leq \frac{\|\nabla\Phi(\mathbf{x})\|^2}{2\mu_{\mathbf{x}}}. \tag{48}$$

To bound $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$, we need to bound the term $\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2$. There holds that

$$\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2 \leq 2\|\nabla\Phi(\bar{\mathbf{x}}_T) - \nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2 + 2\|\nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2. \tag{49}$$

From Theorem 3, under Assumption 3 and 5. Plugging $\bar{\mathbf{x}}_T$ into Theorem 3, for any $\delta \in (0,1)$, when $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, with probability at least $1-\frac{\delta}{2}$

$$\begin{aligned}\|\nabla\Phi(\bar{\mathbf{x}}_T)-\nabla\Phi_S(\bar{\mathbf{x}}_T)\| &\leq \|\nabla\Phi_S(\bar{\mathbf{x}}_T)\| + 2\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*(\mathbf{x}^*);\mathbf{z})\|^2]\log\frac{16}{\delta}}{n}} \\ &+ \frac{2B_{\mathbf{x}^*}\log\frac{16}{\delta}}{n} + \frac{\mu_{\mathbf{x}}}{n} + \frac{2\beta}{\mu_{\mathbf{y}}}\left(\sqrt{\frac{2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T,\mathbf{y}^*(\bar{\mathbf{x}}_T);\mathbf{z})\|^2]\log\frac{8}{\delta}}{n}} + \frac{B_{\mathbf{y}^*}\log\frac{8}{\delta}}{n}\right).\end{aligned} \tag{50}$$

Next, we need to bound the optimization error bound $\|\nabla\Phi_S(\bar{\mathbf{x}}_T)\|$. Firstly, according to Lemma 19, with probability at least $1-\delta$, we have

$$\Phi_S(\bar{\mathbf{x}}_T) - \Phi_S(\mathbf{x}^*) = O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right).$$

According to (Nesterov, 2003) and Lemma 8, there holds the following property for $\beta+\frac{\beta^2}{\mu_{\mathbf{y}}}$ function $\Phi_S(\mathbf{x})$, we have

$$\frac{1}{2\left(\beta+\frac{\beta^2}{\mu_{\mathbf{y}}}\right)}\|\nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2 \leq \Phi_S(\bar{\mathbf{x}}_T) - \Phi_S(\mathbf{x}^*) = O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right). \tag{51}$$

Plugging (51) into (50), according to Cauchy–Bunyakovsky–Schwarz inequality, we can derive that

$$\begin{aligned}\|\nabla\Phi(\bar{\mathbf{x}}_T)-\nabla\Phi_S(\bar{\mathbf{x}}_T)\|^2 &\leq O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right) + \frac{32\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T,\mathbf{y}^*(\bar{\mathbf{x}}_T);\mathbf{z})\|^2]\log\frac{8}{\delta}}{\mu_{\mathbf{y}}^2 n} \\ &+ \frac{32\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{16}{\delta}}{n} + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta} + 2B_{\mathbf{x}^*}\log\frac{16}{\delta} + \mu_{\mathbf{x}}\right)^2}{n^2}.\end{aligned} \tag{52}$$

Then substituting (52), (51) into (49), we derive that

$$\begin{aligned}\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2 &\leq O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right) + \frac{64\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T,\mathbf{y}^*(\bar{\mathbf{x}}_T);\mathbf{z})\|^2]\log\frac{8}{\delta}}{\mu_{\mathbf{y}}^2 n} \\ &+ \frac{64\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]\log\frac{16}{\delta}}{n} + \frac{8\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta} + 2B_{\mathbf{x}^*}\log\frac{16}{\delta} + \mu_{\mathbf{x}}\right)^2}{n^2}.\end{aligned} \tag{53}$$

Finally, we plug (53) into (48) and choose $T \asymp O\left(n^2\right)$, with probability at least $1-\delta$

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

Next, if we further assume $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, According to Lemma 7, we have $\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2 \leq 4\beta\mathbb{E}[f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})]$ and $\mathbb{E}\|\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_T,\mathbf{y}^*(\bar{\mathbf{x}}_T);\mathbf{z})\|^2 \leq 4\beta\mathbb{E}[f(\bar{\mathbf{x}}_T,\mathbf{y}^*(\bar{\mathbf{x}}_T);\mathbf{z})]$. Plugging (53) into (48), then we have

$$\begin{aligned}&\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) \\ &\leq \frac{\|\nabla\Phi(\bar{\mathbf{x}}_T)\|^2}{2\mu_{\mathbf{x}}} \\ &\leq O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right) + \frac{128\beta^3\mathbb{E}[f(\hat{\mathbf{x}}^*,\mathbf{y}^*(\hat{\mathbf{x}}^*);\mathbf{z})]\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{128\beta\mathbb{E}[f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})]\log\frac{16}{\delta}}{\mu_{\mathbf{x}} n} \\ &\quad + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta} + 2B_{\mathbf{x}^*}\log\frac{16}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2} \\ &= O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right) + \frac{128\beta^3\Phi(\hat{\mathbf{x}}^*)\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{128\beta\Phi(\mathbf{x}^*)\log\frac{16}{\delta}}{\mu_{\mathbf{x}} n} + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta} + 2B_{\mathbf{x}^*}\log\frac{16}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2},\end{aligned}$$

Reference	Algorithm	Assumption	Measure	Bounds
Lei	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.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(n^{-\frac{c\beta+1}{2c\beta+1}})$
Li	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.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	$O(\log n/n)$
Zhang	Convergence	NC-SC, Lip, S	(E.) $\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\|$	$O(\sqrt{d/n})$
This work	Convergence	NC-SC, B, S	(H.P.) $\|\nabla\Phi(\mathbf{x}) - \nabla\Phi_S(\mathbf{x})\|$	$O(\sqrt{d/n})$
	ESP	PL-SC, B, S, A	(HP.) $\Phi(\hat{\mathbf{x}}^*) - \Phi(\mathbf{x}^*)$	app. $O(1/n^2)$
	GDA	PL-SC, B, S, A	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	app. $O(1/n^2)$
		SC-SC, B, S, A	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	app. $O(1/n^2)$
	SGDA	PL-SC, B, S, A	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	app. $O(1/n^2)$
		SC-SC, Lip, S, A	(HP.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	app. $O(1/n^2)$
	AGDA	PL-SC, B, S, A	(E.) $\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*)$	app. $O(1/n^2)$

Table 1: Summary of the results. Lei means Lei et al. (2021), Li means Li & Liu (2021a), Zhang means Zhang et al. (2022), The bounds are established by choosing optimal iterate number T.

which implies that

$$\left(1 - \frac{128\beta^3 \log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n}\right)(\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*))$$

$$= O\left(\frac{\log\frac{1}{\delta}}{\sqrt{T}}\right) + \frac{128\beta\Phi(\mathbf{x}^*)\log\frac{16}{\delta}}{\mu_{\mathbf{x}} n} + \frac{128\beta^3\Phi(\mathbf{x}^*)\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{4\left(\frac{2\beta B_{\mathbf{y}^*}}{\mu_{\mathbf{y}}}\log\frac{8}{\delta} + 2B_{\mathbf{x}^*}\log\frac{16}{\delta} + \mu_{\mathbf{x}}\right)^2}{\mu_{\mathbf{x}} n^2}.$$

When $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \frac{128\beta^3\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right\}$, and $T \asymp n^4$ we have

$$\Phi(\bar{\mathbf{x}}_T) - \Phi(\mathbf{x}^*) = O\left(\frac{\Phi(\mathbf{x}^*)\log\frac{1}{\delta}}{n - \frac{128\beta^3\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}} + \frac{\log^2\frac{1}{\delta}}{n\left(n - \frac{128\beta^3\log\frac{8}{\delta}}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)}\right).$$

The proof is complete.

□

E SOME IMPROVED BOUNDS WITH EXPECTATION FORMATS

Table 1 gives the summary of the existing results. Convergence means the uniform (localized) convergence 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. A means that we assume $\Phi(\mathbf{x}^*)$ is of order $O\left(\frac{1}{n}\right)$. 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 and app. $O(1/n^2)$ means the result is approximate $O(1/n^2)$ when n is large enough. Since most of the existing work on optimization error is 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 this section.

Firstly, we translate our high probability result of Theorem 3 into an expectation result.

Theorem 9. *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 $c = \max\{16C^2, 1\}$. We have that for all $\mathbf{x} \in \mathcal{X}$, when*

$n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, *the excess risks of primal functions can be bounded by*

$$\mathbb{E}\|\Phi(\mathbf{x})-\Phi(\mathbf{x}^*)\| \leq \frac{8\mathbb{E}[\|\nabla\Phi_S(\mathbf{x})\|^2]}{\mu_{\mathbf{x}}} + O\Bigg(\frac{\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]}{\mu_{\mathbf{x}} n} + \frac{\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{1}{n^2}\Bigg).$$

Proof of Theorem 9. According to Theorem 3, we have that for all $\mathbf{x} \in \mathcal{X}$, when $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$ with probability at least $1-\delta$

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

Thus, with probability at least $1-\delta$, we have

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

According to the standard statistical analysis, Let X be a random variable, if for some $v, c > 0$,$\mathbb{P}\{X > vt + ct^2\} \leq e^{-t}$ for every $t > 0$. Then we can easily derive that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}\{X > x\}dx = v + 2c$. Thus, we have the expectation result

$$\mathbb{E}\|\Phi(\mathbf{x})-\Phi(\mathbf{x}^*)\| \leq \frac{8\mathbb{E}[\|\nabla\Phi_S(\mathbf{x})\|^2]}{\mu_{\mathbf{x}}} + O\Bigg(\frac{\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*;\mathbf{z})\|^2]}{\mu_{\mathbf{x}} n} + \frac{\beta^2\mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\mathbf{z})\|^2]}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2 n} + \frac{1}{n^2}\Bigg).$$

The proof is complete. □

Next, we give the proofs of the expectation result to relax the SC-SC assumptions given in Table 1. The proofs are similar with high probability format since we use the existing results for optimization error bounds with expectation format under NC-SC assumptions.

E.1 GDA

Lemma 20 (Optimization error bound for GDA in NC-SC minimax problems (Lin et al., 2020))**.** *Under Assumption 1, and letting the step sizes be chosen as* $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ *and* $\eta_{\mathbf{y}} = \frac{1}{\beta}$*, then the optimization error bound of Algorithm 1 can be bounded by*

$$\mathbb{E}\|\nabla\Phi_S(\mathbf{x}_T)\|^2 = O\left(\frac{\beta^3\Delta_\Phi}{\mu_{\mathbf{y}}^2 T} + \frac{\beta^3 D_{\mathcal{Y}}}{\mu_{\mathbf{y}} T}\right),$$

where $\Delta_\Phi = \Phi_S(\mathbf{x}_0) - \min_{\mathbf{x}}\Phi_S(\mathbf{x})$.

Using above optimization error bound, we can obtain the following theorem.

Theorem 10. *Suppose Assumption 1 and 3 hold. Assume that the population risk $F(\mathbf{x}, \mathbf{y})$ satisfies Assumption 4 with parameter $\mu_{\mathbf{x}}$. Let the step sizes choose as $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ and $\eta_{\mathbf{y}} = \frac{1}{\beta}$. When $T \asymp n$ and $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, where c is an absolute constant, then the excess risk for primal functions of Algorithm 1 can be bounded by*

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

Furthermore, Let $T \asymp n^2$ and $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \Theta\left(\frac{\beta^3}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)\right\}$. Assume the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is non-negative and $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\mathbb{E}[\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)] \approx O\left(\frac{1}{n^2}\right).$$

E.2 SGDA

For SGDA settings, we introduce a weaker assumption comparing with Assumption 2.

Assumption 6. *Assume the existence of $\sigma > 0$ satisfies*

$$\mathbb{E}[\nabla f(\mathbf{x}, \mathbf{y}; \mathbf{z}) - \nabla F_S(\mathbf{x}, \mathbf{y})] = 0,$$
$$\mathbb{E}[\|\nabla f(\mathbf{x}, \mathbf{y}; \mathbf{z}) - \nabla F_S(\mathbf{x}, \mathbf{y})\|^2] \leq \sigma^2.$$

Lemma 21 (Optimization error bound for SGDA in NC-SC minimax problems (Lin et al., 2020))**.** *Under Assumption 1 and 6, and letting the step sizes be chosen as $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ and $\eta_{\mathbf{y}} = \frac{1}{\beta}$, then the optimization error bound of Algorithm 1 can be bounded by*

$$\mathbb{E}\|\nabla\Phi_S(\mathbf{x}_T)\|^2 = O\left(\sqrt[5]{\frac{\beta^6}{\mu_{\mathbf{y}}^6 T^2}}\right).$$

Using above optimization error bound, we can obtain the following theorem.

Theorem 11. *Suppose Assumption 1, 3 and 6 hold. Assume that the population risk $F(\mathbf{x}, \mathbf{y})$ satisfies Assumption 4 with parameter $\mu_{\mathbf{x}}$. Let the step sizes choose as $\eta_{\mathbf{x}} = \frac{1}{16(\frac{\beta}{\mu}+1)^2\beta}$ and $\eta_{\mathbf{y}} = \frac{1}{\beta}$. When $T \asymp n^{\frac{5}{2}}$ and $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}$, where c is an absolute constant, then the excess risk for primal functions of Algorithm 2 can be bounded by*

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

Furthermore, Let $T \asymp n^5$ and $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \Theta\left(\frac{\beta^3}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)\right\}$. Assume the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is non-negative and $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\mathbb{E}[\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)] \approx O\left(\frac{1}{n^2}\right).$$

E.3 AGDA

Alternating gradient descent ascent presented in Algorithm 3 was proposed recently to optimize nonconvex-nonconcave problems (Yang et al., 2020).

Algorithm 3 Two-timescale AGDA for minimax problem

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

Lemma 22 (Optimization error bound for AGDA in PL-SC minimax problems (Yang et al., 2020; Lei et al., 2021))**.** *Under Assumption 1 and assume that the population risk $F(\mathbf{x}, \mathbf{y})$ satisfies Assumption 4 with parameter $\mu_{\mathbf{x}}$. Let $\{\mathbf{x}_t, \mathbf{y}_t\}_t$ be the sequence produced by Algorithm 3 with the step sizes chosen as $\eta_{\mathbf{x}_t} \asymp \frac{1}{\mu_{\mathbf{x}} t}$ and $\eta_{\mathbf{y}_t} = \frac{1}{\mu_{\mathbf{x}} \mu_{\mathbf{y}}^2 t}$, then the optimization error bound of Algorithm 3 can be bounded by*

$$\mathbb{E}[\|\nabla \Phi_S(\mathbf{x}_T)\|^2] = O\left(\frac{1}{\mu_{\mathbf{x}}^2 \mu_{\mathbf{y}}^4 T}\right).$$

Theorem 12. *Suppose Assumption 1, 3 and 6 hold. Assume that the population risk $F(\mathbf{x}, \mathbf{y})$ satisfies Assumption 4 with parameter $\mu_{\mathbf{x}}$. Let $\{\mathbf{x}_t, \mathbf{y}_t\}_t$ be the sequence produced by Algorithm 3 with the step sizes chosen as $\eta_{\mathbf{x}_t} \asymp \frac{1}{\mu_{\mathbf{x}} t}$ and $\eta_{\mathbf{y}_t} \asymp \frac{1}{\mu_{\mathbf{x}} \mu_{\mathbf{y}}^2 t}$. When $T \asymp n$ and $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2 \mu_{\mathbf{x}}^2}$, where c is an absolute constant, then the excess risk for primal functions of Algorithm 3 can be bounded by*

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

Furthermore, Let $T \asymp n^2$ and $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2 \mu_{\mathbf{x}}^2}, \Theta\left(\frac{\beta^3}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)\right\}$. Assume the function $f(\mathbf{x}, \mathbf{y}; \mathbf{z})$ is non-negative and $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, we have

$$\mathbb{E}[\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)] \approx O\left(\frac{1}{n^2}\right).$$

Proof of Theorem 12. Since the proofs of Theorem 10, 11 and 12 are similar, we only prove Theorem 12 as an example.

From Theorem 9, under Assumption 3 and 1, when we plug $\mathbf{x}_T$ into Theorem 9, when $n \geq \frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2 \mu_{\mathbf{x}}^2}$, we have

$$\begin{aligned}\mathbb{E}\|\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)\| \leq \frac{8\mathbb{E}[\|\nabla \Phi_S(\mathbf{x}_T)\|^2]}{\mu_{\mathbf{x}}} + O\Bigg(&\frac{\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2]}{\mu_{\mathbf{x}} n} \\ &+ \frac{\beta^2 \mathbb{E}[\|\nabla_{\mathbf{y}} f(\mathbf{x}_T, \mathbf{y}^*(\mathbf{x}_T); \mathbf{z})\|^2]}{\mu_{\mathbf{x}} \mu_{\mathbf{y}}^2 n} + \frac{1}{n^2}\Bigg).\end{aligned} \tag{54}$$

According to Lemma 22, we have

$$\mathbb{E}[\|\nabla \Phi_S(\mathbf{x}_T)\|^2] = O\left(\frac{1}{\mu_{\mathbf{x}}^2 \mu_{\mathbf{y}}^4 T}\right). \tag{55}$$

Plugging (55) into (54) and choose $T \asymp O(n)$, with probability at least $1 - \delta$

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

Next, if we further assume $\Phi(\mathbf{x}^*) = O\left(\frac{1}{n}\right)$, According to Lemma 7, we have $\mathbb{E}\|\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})\|^2 \leq 4\beta \mathbb{E}[f(\mathbf{x}^*, \mathbf{y}^*; \mathbf{z})]$ and $\mathbb{E}\|\nabla_{\mathbf{y}} f(\mathbf{x}_T, \mathbf{y}^*(\mathbf{x}_T); \mathbf{z})\|^2 \leq 4\beta \mathbb{E}[f(\mathbf{x}_T, \mathbf{y}^*(\mathbf{x}_T); \mathbf{z})]$, then substituting (55) into (54) and choosing $T \asymp n^2$, we have

$$\mathbb{E}[\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)] = O\left(\frac{\beta^3 \mathbb{E}[\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)]}{\mu_{\mathbf{x}} \mu_{\mathbf{y}}^2 n} + \frac{\beta^3 \mathbb{E}[\Phi(\mathbf{x}^*)]}{\mu_{\mathbf{x}} \mu_{\mathbf{y}}^2 n} + \frac{\beta \mathbb{E}[\Phi(\mathbf{x}^*)]}{\mu_{\mathbf{x}} n} + \frac{1}{n^2}\right).$$

which implies that When $n \geq \max\left\{\frac{c\beta^2(\mu_{\mathbf{y}}+\beta)^2(d+\log\frac{8\log_2\sqrt{2}R_1 n+1}{\delta})}{\mu_{\mathbf{y}}^2\mu_{\mathbf{x}}^2}, \Theta\left(\frac{\beta^3}{\mu_{\mathbf{x}}\mu_{\mathbf{y}}^2}\right)\right\}$, we have

$$\mathbb{E}[\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}^*)] \approx O\left(\frac{1}{n^2}\right).$$

The proof is complete.

□