Generalization Bounds of Nonconvex-(Strongly)-Concave Stochastic Minimax Optimization

Siqi Zhang* Johns Hopkins University

Yifan Hu* EPFL & ETH Zürich

Liang Zhang ETH Zürich

Niao He ETH Zürich

Abstract

This paper studies the generalization performance of algorithms for solving nonconvex-(strongly)-concave (NC-SC / NC-C) stochastic minimax optimization measured by the stationarity of primal functions. We first establish *algorithm-agnostic generalization bounds* via *uniform convergence* between the empirical minimax problem and the population minimax problem. The sample complexities for achieving ϵ-generalization are $\tilde{\mathcal{O}}(d\kappa^2\epsilon^{-2})$ and $\tilde{\mathcal{O}}(d\epsilon^{-4})$ for NC-SC and NC-C settings, respectively, where d is the dimension of the primal variable and κ is the condition number. We further study the *algorithm-dependent generalization bounds* via stability arguments of algorithms. In particular, we introduce a novel stability notion for minimax problems and build a connection between stability and generalization. As a result, we establish *algorithm-dependent generalization bounds* for *stochastic gradient descent ascent (SGDA)* and the more general *sampling-determined algorithms (SDA)*.

1 Introduction

In this paper, we consider stochastic minimax problems:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F(x,y) \triangleq \mathbb{E}_{\xi}\left[f(x,y;\xi)\right], \tag{1}$$

where $\mathcal{X}\subseteq\mathbb{R}^d$ and $\mathcal{Y}\subseteq\mathbb{R}^{d'}$ $(d,d'\in\mathbb{N}_+)$ are two nonempty closed convex sets, $\xi\in\Xi$ is a random variable following an unknown distribution $\mathcal{D}$, and $f:\mathcal{X}\times\mathcal{Y}\times\Xi\to\mathbb{R}$ is continuously differentiable and Lipschitz smooth jointly in x and y for any ξ. We denote the objective (1) as the *population minimax problem*. Throughout the paper, we focus on the case where F is nonconvex in x and (strongly)-concave in y, i.e., *nonconvex-(strongly)-concave (NC-SC / NC-C)*. Such minimax problems appear ubiquitously in practical applications, including adversarial training (Madry et al., 2018; Wang et al., 2019), generative adversarial networks (GANs) (Goodfellow et al., 2014; Sanjabi et al., 2018; Lei et al., 2020), reinforcement learning (Dai et al., 2017, 2018; Huang and Jiang, 2022) and robust training (Sinha et al., 2018).

Although the distribution $\mathcal{D}$ often remains unknown, one generally has access to a dataset $S=\{\xi_1,\cdots,\xi_n\}$ consisting of n independently and identical distributed (i.i.d.) samples from $\mathcal{D}$. Correspondingly, researchers resort to solving an *empirical minimax problem*:

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

A natural question arises: *How does the output of an algorithm $\mathcal{A}$ for solving the empirical minimax problem generalizes on the population minimax problem?*

We first specify the measurement. Since functions F and F_S are nonconvex in x, finding their global optimal solutions is generally intractable. Instead, one aims to design an algorithm $\mathcal{A}$ that finds an *ϵ-stationary point* of the primal function

$$\Phi(x) \triangleq \max_{y\in\mathcal{Y}} F(x,y).$$

It has been shown that $\Phi(x)$ for NC-SC problems is smooth, while it can be nonsmooth for NC-C problems (Thekumparampil et al., 2019; Lin et al., 2020a). So in the NC-SC case, we find the point such that[1]:

$$\|\nabla\Phi(\mathcal{A}_x(S))\|\le\epsilon,$$

and in the NC-C case, it is characterized as

$$\mathbf{dist}\left(0,\partial\Phi(\mathcal{A}_x(S))\right)\le\epsilon,$$

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

[1]For simplicity, here we assume $X=\mathbb{R}^d$, and primal functions $\Phi(x)$ and $\Phi_S(x)$ are differentiable. We will formally introduce the detailed settings in Section 2.

where $\mathcal{A}_x(S)$ is the x-component of the output of any algorithm $\mathcal{A}$ for solving (2), $\mathbf{dist}(y, X) \triangleq \inf_{x\in X} \|y - x\|$ and $\partial\Phi$ is the (Fréchet) subdifferential of Φ. When Φ is nonsmooth, we resort to the gradient norm of its Moreau envelope to measure the first-order stationarity as it provides an upper bound on $\mathbf{dist}(0, \partial\Phi(x))$ (Davis and Drusvyatskiy, 2019).

Taking the gradient norm as an example, the error for solving the population minimax problem (1) via solving its empirical counterpart (2) consists of two terms:

$$\begin{aligned}&\mathbb{E}\,\|\nabla\Phi(\mathcal{A}_x(S))\| \\ \leq\ &\underbrace{\mathbb{E}\,\|\nabla\Phi_S(\mathcal{A}_x(S))\|}_{\text{optimization error}} + \underbrace{\mathbb{E}\,\|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\|}_{\text{generalization error}},\end{aligned} \tag{3}$$

here $\Phi_S(x) \triangleq \max_{y\in\mathcal{Y}} F_S(x, y)$ is the primal function of the empirical function (2). Such decomposition on the gradient norm also appears in nonconvex minimization, e.g., Foster et al. (2018); Mei et al. (2018); Davis and Drusvyatskiy (2022); Lei (2022). The optimization error corresponds to the error of solving the empirical minimax problem (2) which has been widely studied (Luo et al., 2020; Yang et al., 2020b).

On the other hand, the generalization error for minimax problems remains largely unexplored, recently Ozdaglar et al. (2022) found counterexamples to show that the common primal function value gap (Farnia and Ozdaglar, 2021) between the population and empirical objectives may fail to characterize the generalization error. So in this paper, our goal is to characterize the generalization error

$$\mathbb{E}\,\|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\|.$$

It is not easy as both $\Phi_S(\cdot)$ and $\mathcal{A}_x(S)$ depend on the dataset S, which induces correlation issues when taking expectation. To address such dependence issue, one may use *uniform convergence* or *stability arguments*.

By uniform convergence, we characterize the difference between the empirical minimax optimization and the population minimax problem on worst $x \in \mathcal{X}$, i.e.,

$$\mathbb{E}\sup_{x\in\mathcal{X}} \|\nabla\Phi(x) - \nabla\Phi_S(x)\|.$$

Although uniform convergence has been extensively studied for stochastic minimization (Kleywegt et al., 2002; Mei et al., 2018; Davis and Drusvyatskiy, 2022), a key difference for stochastic minimax problems is that the empirical primal function $\Phi_S(x)$ is the maximum of an average over n random functions, which drives existing uniform convergence analysis for stochastic minimization to be inapplicable here. Note that uniform convergence is invariant to the choice of algorithm and provides an upper bound of the generalization error for any $\mathcal{A}_x(S) \in \mathcal{X}$. Thus the derived bound is *algorithm-agnostic* that applies to any algorithms.

Another approach to investigating generalization is stability arguments, which analyze the stability of specific algorithms and build a connection between stability and generalization. It has been extensively studied for stochastic minimization (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016; Klochkov and Zhivotovskiy, 2021) and recently for minimax problems (Farnia and Ozdaglar, 2021; Lei et al., 2021; Boob and Guzmán, 2023; Yang et al., 2022c; Ozdaglar et al., 2022). Yet most of these work use function-value gaps as the measurement. For the measurement of stationarity for nonconvex problems, building up a link between stability and generalization becomes significantly more challenging. Compared to uniform convergence, the stability-based generalization bound is generally independent of the dimension d. As it requires a case-by-case analysis of stability for different algorithms, it is *algorithm-dependent*. We particularly study the generalization of the widely used *stochastic gradient descent ascent (SGDA)* (Farnia and Ozdaglar, 2021) and a broad class of algorithms called *sampling-determined algorithms* (SDA, see Definition 4.7) (Lei, 2022).

1.1 Contributions

In this paper, we provide a systematic study on generalization bounds (see Table 1) for nonconvex stochastic minimax problems from both *uniform convergence* and *stability argument* perspectives. To be more specific:

- We establish the first uniform convergence results between the population and the empirical nonconvex minimax optimization in NC-SC and NC-C settings, measured by stationarity. Our results provide an *algorithm-agnostic* generalization bound for any algorithms that solve empirical nonconvex minimax problems. Specifically, the sample complexities to achieve an ϵ-uniform convergence or an ϵ-generalization error are $\tilde{\mathcal{O}}\big(d\kappa^2\epsilon^{-2}\big)$ and $\tilde{\mathcal{O}}\big(d\epsilon^{-4}\big)$ for the NC-SC and NC-C settings, respectively.
- We introduce a novel stability measurement based on stationarity; then, we establish the connection between the stability and the generalization error of an algorithm in both NC-SC and NC-C settings. We further provide the *algorithm-dependent* generalization error bound measured by the stationarity for the classical SGDA algorithm and sampling-determined algorithms utilizing their stability.
- Regarding the technical novelty compared to existing works, in the NC-SC case, we identified and

characterized the distance between maximizes

$$\operatorname*{argmax}_{y\in\mathcal{Y}} F_S(x,y) \quad \text{and} \quad \operatorname*{argmax}_{y\in\mathcal{Y}} F(x,y),$$

which is the most significant differentiation from prior works; in the NC-C case, given the non-uniqueness of the dual maximizers, we incorporated the ℓ_2-regularized objective into the analysis, which successfully forged a link between the surrogate and the original objective, facilitating the transition of the analysis into the NC-SC domain.

1.2 Literature Review

Nonconvex Minimax Optimization Various algorithms have been proposed to solve NC-SC minimax optimization (Nouiehed et al., 2019; Lin et al., 2020a,b; Luo et al., 2020; Yang et al., 2020a; Boţ and Böhm, 2020; Xu et al., 2023; Lu et al., 2020; Yan et al., 2020; Guo et al., 2021; Sharma et al., 2022; Yang et al., 2022b; Zhang et al., 2022), and the lower bounds are recently studied in several works (Zhang et al., 2021b; Han et al., 2021; Li et al., 2021). Recent years witnessed a surge of algorithms for NC-C problems in both deterministic and stochastic settings, e.g., Zhang et al. (2020); Ostrovskii et al. (2021); Thekumparampil et al. (2019); Zhao (2020); Nouiehed et al. (2019); Yang et al. (2020b); Lin et al. (2020a); Boţ and Böhm (2020); Rafique et al. (2021). These works differ from ours in that we aim to characterize the generalization error of algorithms while they focus mainly on the optimization part.

Uniform Convergence A series of work from stochastic optimization and statistical learning theory studied uniform convergence on the worst-case differences between the population objective and its empirical objective constructed via sample average approximation (SAA, also known as empirical risk minimization). Interested readers may refer to prominent results in the literature (Fisher, 1922; Vapnik, 1999; Van der Vaart, 2000; Kleywegt et al., 2002; Shapiro, 2006; Hu et al., 2020). For finite-dimensional problems, Kleywegt et al. (2002) showed that the sample complexity is $\mathcal{O}(d\epsilon^{-2})$ to achieve an ϵ-uniform convergence of function values in high probability. For nonconvex empirical objectives, Mei et al. (2018) and Davis and Drusvyatskiy (2022) established $\tilde{\mathcal{O}}(d\epsilon^{-2})$ sample complexity to achieve an ϵ-uniform convergence measured by the stationarity for nonconvex smooth and weakly convex functions, respectively. In addition, Wang et al. (2017) used uniform convergence to demonstrate the generalization and the gradient complexity of differentially private algorithms for stochastic optimization. Recently, Amir et al. (2022) demonstrated the generalization error of gradient descent on a generalized linear model using uniform convergence and showed that the stability argument is insufficient to achieve generalization. To the best of our knowledge, our paper is the first work to study uniform convergence for nonconvex minimax optimization.

Stability-Based Generalization Bounds This line of research focuses on analyzing generalization bounds of stochastic optimization via the uniform stability property of specific algorithms, including SAA (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010), stochastic gradient descent (Hardt et al., 2016; Bassily et al., 2020; Lei, 2022), and uniformly stable algorithms (Klochkov and Zhivotovskiy, 2021). Recently, a series of work further studied generalization measured by the function-value gap of various algorithms in minimax problems. For example, Farnia and Ozdaglar (2021) gave the generalization bound for the outputs of gradient-descent-ascent (GDA) and proximal-point algorithm (PPA) for both (strongly)-convex-(strongly)-concave and nonconvex-nonconcave smooth minimax problems. Lei et al. (2021) studied the stability and generalization of GDA in various settings of minimax problems covering both convex and nonconvex scenarios, while their results are sill based on the function-value gaps. Boob and Guzmán (2023) established stability and generalization results of extragradient (EG) in the smooth convex-concave setting. Zhang et al. (2021a) studied the stability and generalization of the empirical minimax problem under the (strongly)-convex-(strongly)-concave setting, assuming access to the optimal solution of the empirical minimax problem. Our work differs from those in that we consider different notions of generalization errors, also we propose a novel stability notion for minimax optimization measured by stationarity, and build up a link between such stability and generalization.

Notations $\|\cdot\|$ is reserved for the ℓ_2-norm. A set $\mathcal{X}$ is compact with a diameter $D_{\mathcal{X}} > 0$ if $\forall x \in \mathcal{X}$, $\|x\|^2 \leq D_{\mathcal{X}}$. $\mathbf{proj}_{\mathcal{X}}(x') \triangleq \operatorname{argmin}_{x\in\mathcal{X}} \|x - x'\|^2$ is the projection of x on a set $\mathcal{X}$. Let $\mathcal{A}(S) \triangleq (\mathcal{A}_x(S), \mathcal{A}_y(S))$ be the output of an algorithm $\mathcal{A}$ on the empirical minimax problem (2) with dataset S. We use $\nabla f = (\nabla_x f, \nabla_y f)$ to denote the gradient of a continuously differentiable function $f(x,y)$. Given $\mu \geq 0$, we say a function $g : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex if $g(x) - (\mu/2)\|x\|^2$ is convex, and μ-strongly concave if $-g$ is μ-strongly convex. The function $g(x)$ is μ-weakly convex if $g(x)+(\mu/2)\|x\|^2$ is convex. We say a function $f(x,y)$ is L-smooth with $L > 0$ if it is continuously differentiable and $\|\nabla f(x_1,y_1) - \nabla f(x_2,y_2)\|^2 \leq L^2(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2)$ for any $(x_1,y_1),(x_2,y_2) \in \mathcal{X}\times\mathcal{Y}$. By definition, it is easy to verify that any L-smooth function is also L-weakly convex.

Table 1: Summary of Generalization Bounds for Nonconvex Stochastic Minimax Optimization

Setting[1] \ Approach	Uniform Convergence	Stability Argument	
		SGDA	Sampling-determined Alg.
NC-SC	$\tilde{\mathcal{O}}\Big(\kappa\sqrt{\frac{d}{n}}\Big)$ Theorem 3.1	$\mathcal{O}\Big(\kappa^{1+\zeta_1}\Big(\frac{T^{1-\zeta_1}}{n}+\frac{1}{\sqrt{n}}\Big)\Big)$ Corollary 4.5	$\mathcal{O}\Big(\kappa\Big(\sqrt{\frac{T}{n}}+\frac{1}{\sqrt{n}}\Big)\Big)$ Corollary 4.8
NC-C	$\tilde{\mathcal{O}}\Big(\big(\frac{d}{n}\big)^{1/4}\Big)$ Theorem 3.2	$\mathcal{O}\Big(\Big(\frac{T^{1-\zeta_2}}{n}\Big)^{1/6}+\big(\frac{1}{n}\big)^{1/8}\Big)$ Corollary 4.6	$\mathcal{O}\Big(\big(\frac{T}{n}\big)^{1/12}+\big(\frac{1}{n}\big)^{1/8}\Big)$ Corollary 4.9

[1] $\tilde{O}(\cdot)$ hides logarithmic factors, d: the dimension of $\mathcal{X}$, n: sample size, κ: condition number $\frac{L}{\mu}$
L: Lipschitz smoothness parameter, μ: strong concavity parameter, T: iteration number of algorithms
$\zeta_1, \zeta_2 \in (0,1)$: constants depending on stepsizes, refer to Corollary 4.5 and 4.6 for details. SGDA has specific requirements on stepsize, while sampling-determined algorithms do not have restrictions on stepsize.

2 Problem Setting

We study the generalization errors of *nonconvex-strongly-concave* (NC-SC) and *nonconvex-concave* (NC-C) minimax problems. We begin with the main assumptions used throughout the paper.

Assumption 2.1 (Main Settings)**.** We assume:

(a) The function $f(x,y;\xi)$ is μ-strongly concave in $y \in \mathcal{Y}$ for any $x \in \mathcal{X}$ and $\xi \in \Xi$ for some $\mu \geq 0$.

(b) The function $f(x,y;\xi)$ is L-smooth jointly in $(x,y) \in \mathcal{X} \times \mathcal{Y}$ for any $\xi \in \Xi$.

(c) The gradient norm $\|\nabla f(x,y;\xi)\|$ is bounded by G for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $\xi \in \Xi$.

(d) Domains $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets with diameters $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$.

Assumption 2.1 appears widely in nonconvex minimax optimization literature (Lin et al., 2020a; Zhang et al., 2021b), and the compact domain assumption is standard for uniform convergence (Kleywegt et al., 2002; Davis and Drusvyatskiy, 2022).

Performance Measurement Next, we demonstrate how to evaluate generalization in nonconvex minimax optimization. In the NC-SC setting ($\mu > 0$), the primal functions Φ and Φ_S are both continuously differentiable and $\tilde{L}$-smooth as presented in the literature.

Lemma 2.2 (Properties of Φ (Davis and Drusvyatskiy, 2019; Lin et al., 2020a))**.** In the NC-SC setting, both $\Phi(x)$ and $\Phi_S(x)$ are $\tilde{L} \triangleq L(1+\kappa)$-smooth with the condition number $\kappa \triangleq L/\mu$. Both $y^*(x)$ and $y_S^*(x)$ are κ-Lipschitz continuous, and $\nabla\Phi(x) = \nabla_x F(x, y^*(x)), \nabla\Phi_S(x) = \nabla_x F_S(x, y_S^*(x))$.

Due to the constraint $\mathcal{X}$, we measure the difference of the stationarity between the population and empirical problems using their *generalized gradients*, i.e., $\mathbb{E}\,\|\mathcal{G}_\Phi(\mathcal{A}_x(S)) - \mathcal{G}_{\Phi_S}(\mathcal{A}_x(S))\|$, where $\mathcal{G}_\Phi(x) \triangleq \tilde{L}(x - \mathbf{proj}_{\mathcal{X}}(x - (1/\tilde{L})\nabla\Phi(x)))$. It is easy to find that

$$\underbrace{\mathbb{E}\,\|\mathcal{G}_\Phi(\mathcal{A}_x(S)) - \mathcal{G}_{\Phi_S}(\mathcal{A}_x(S))\|}_{\text{generalization error of Algorithm } \mathcal{A}} \leq \underbrace{\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi(x) - \nabla\Phi_S(x)\|}_{\text{alg.-agnostic uniform convergence}},$$

where the inequality holds as the projection operation onto a convex set is non-expansive. The term in the left-hand side (LHS) is the generalization error of an algorithm $\mathcal{A}$ in the NC-SC case. As it is always bounded by the difference between gradients in the right-hand side (RHS) above, we can directly analyze the RHS to derive the generalization bounds.

For the NC-C case ($\mu = 0$), the primal function $\Phi(x)$ is L-weakly convex and can be nonsmooth (Lin et al., 2020a). Thus we use the gradient of its Moreau Envelope to characterize the (near)-stationarity (Davis and Drusvyatskiy, 2019).

Definition 2.3 (Moreau Envelope)**.** For an L-weakly convex function Φ and $0 < \lambda < 1/L$, we use $\Phi^\lambda(x)$ and $\mathbf{prox}_{\lambda\Phi}(x)$ to denote the the Moreau envelope of Φ and the proximal point of Φ for a given point x, defined as following:

$$\Phi^\lambda(x) \triangleq \min_{z\in\mathcal{X}}\left\{\Phi(z) + \frac{1}{2\lambda}\|z-x\|^2\right\},$$
$$\mathbf{prox}_{\lambda\Phi}(x) \triangleq \operatorname*{argmin}_{z\in\mathcal{X}}\left\{\Phi(z) + \frac{1}{2\lambda}\|z-x\|^2\right\}.$$

The following lemma illustrates the relationship between the subdifferential of the primal function and the gradient of its Moreau envelope.

Lemma 2.4 (Properties of Φ^λ (Davis and Drusvyatskiy, 2019; Thekumparampil et al., 2019))**.** In the NC-C setting, the primal function Φ is L-weakly

convex. For $\lambda \in (0, 1/L)$, the Moreau envelope $\Phi^\lambda(x)$ is smooth, and its gradient satisfies $\nabla\Phi^\lambda(x) = \lambda^{-1}(x - \hat{x})$ where $\hat{x} = \mathbf{prox}_{\lambda\Phi}(x)$ is the proximal point, and $\mathbf{dist}\,(0, \partial\Phi(\hat{x})) \le \left\|\nabla\Phi^\lambda(x)\right\|$.

Lemma 2.4 indicates that in the NC-C case, we can measure the generalization error via the difference between *the gradients of the Moreau envelopes* from the population and empirical problems, i.e.,

$$\mathbb{E}\,\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\|.$$

3 Uniform Convergence and Generalization

In this section, we discuss the algorithm-agnostic generalization errors of stochastic minimax optimization using uniform convergence, in both NC-SC and NC-C cases. Throughout the section, we measure the performance using the stationarity of primal functions as discussed in the previous section.

3.1 NC-SC Stochastic Minimax Optimization

Under the NC-SC setting, the next theorem demonstrates the uniform convergence between gradients of the population and empirical primal functions in minimax problems.

Theorem 3.1 (Uniform Convergence, NC-SC)**.** Under Assumption 2.1 with $\mu > 0$, we have the uniform convergence:

$$\mathbb{E}\left[\max_{x\in\mathcal{X}}\|\nabla\Phi(x) - \nabla\Phi_S(x)\|\right] = \tilde{\mathcal{O}}\big(d^{1/2}\kappa G n^{-1/2} + \epsilon\big).$$

This means it suffices to have $n = \tilde{\mathcal{O}}\big(d\kappa^2 G^2 \epsilon^{-2}\big)$ to achieve ϵ-generalization error for any algorithm $\mathcal{A}$ such that $\mathbb{E}\,\|\mathcal{G}_\Phi(\mathcal{A}_x(S)) - \mathcal{G}_{\Phi_S}(\mathcal{A}_x(S))\| \le \epsilon$.

To the best of our knowledge, it is the first uniform convergence and algorithm-agnostic generalization error bound result for NC-SC stochastic minimax problems. In comparison, existing works on generalization error analysis of minimax problems (Farnia and Ozdaglar, 2021; Lei et al., 2021) using stability arguments are algorithm-specific and can only handle function-value gap measurement. Zhang et al. (2021a) establish algorithm-agnostic stability and generalization in the strongly-convex-strongly-concave regime, yet their analysis does not extend to the nonconvex regime. Since the above generalization result is algorithm-agnostic, it holds for any algorithms that search for approximate stationary points of empirical minimax problems. This is particularly useful for SOTA algorithms designed for finite-sum minimax optimization, including Catalyst-SVRG (Zhang et al., 2021b) and SREDA (Luo et al., 2020), as these algorithms are too complicated to conduct stability analysis to derive generalization bounds.

Proof Sketch We briefly discuss the proof of Theorem 3.1, we defer the proof to Appendix B.

Step 1: First, we use a ρ-net $\{x_k\}_{k=1}^Q$ (Vapnik, 1999) to decompose the error and handle the dependence issue between $\mathrm{argmax}_{x\in\mathcal{X}}\,\|\nabla\Phi_S(x) - \nabla\Phi(x)\|$ and $\Phi_S(x)$.

Step 2: For any x_k within the ρ-net, we have the error following decomposition

$$\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| \le \mathbb{E}\,\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| + (\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| - \mathbb{E}\,\|\nabla\Phi_S(x) - \nabla\Phi(x_k)\|).$$

When bounding $\mathbb{E}\,\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\|$ in the right-hand side (RHS), we need to characterize the difference between $\mathrm{argmax}_{y\in\mathcal{Y}} F_S(x, y)$ and $\mathrm{argmax}_{y\in\mathcal{Y}} F(x, y)$ using the stability argument of sample average approximation (Shalev-Shwartz et al., 2009). This step appears uniquely for minimax optimization due to the special structure of the primal function $\Phi_S(x) = \max_y \frac{1}{n}\sum_{i=1}^n f(x, y; \xi_i)$, which is not the average over n random functions. Then we utilize the established stability argument to show that the first term in the RHS is sub-Gaussian and apply the concentration inequality, which leads to the result. □

3.2 NC-C Stochastic Minimax Optimization

Recall that, in the NC-C case, the primal function Φ is L-weakly convex (Thekumparampil et al., 2019), and thus that $\nabla\Phi$ is not well-defined. As shown in Lemma 2.4, we use the gradient norm of the Moreau envelope of the primal function as the measurement.

Here we consider the algorithm-agnostic generalization bound obtained via uniform convergence, i.e.,

$$\begin{aligned}&\mathbb{E}\left\|\nabla\Phi^{\frac{1}{2L}}(\mathcal{A}_x(S)) - \nabla\Phi_S^{\frac{1}{2L}}(\mathcal{A}_x(S))\right\| \\ \le\ &\mathbb{E}\left[\max_{x\in\mathcal{X}}\left\|\nabla\Phi^{\frac{1}{2L}}(x) - \nabla\Phi_S^{\frac{1}{2L}}(x)\right\|\right].\end{aligned}$$

The next theorem illustrates the generalization error for NC-C stochastic minimax optimization problems.

Theorem 3.2 (Uniform Convergence, NC-C)**.** Under Assumption 2.1 with $\mu = 0$, we have

$$\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S^{1/(2L)}(x) - \nabla\Phi^{1/(2L)}(x)\| = \tilde{\mathcal{O}}(d^{1/4}n^{-1/4}).$$

Thus it suffices to have $n = \tilde{\mathcal{O}}\big(d\epsilon^{-4}\big)$ to achieve ϵ-generalization error for any algorithm $\mathcal{A}$ such that $\mathbb{E}\,\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\| \le \epsilon$.

To the best of our knowledge, this is the first algorithm-agnostic generalization result for NC-C stochastic min-

imax optimization. Theorem 3.2 with a similar error decomposition as (3) provides guarantees for the population minimax problem for any algorithms that solve the NC-C empirical problem, including the best-known Catalyst algorithm (Yang et al., 2020b). More specifically, if an algorithm finds an ϵ-stationary point of the empirical minimax problem, with sample size $n = \tilde{\mathcal{O}}\big(d\epsilon^{-4}\big)$, the point is also an $\mathcal{O}(\epsilon)$-stationary point of the population minimax problem.

Proof Sketch The analysis of Theorem 3.2 builds up a link between NC-C and NC-SC settings and consists of three parts. We defer the detailed proof to Appendix C, and briefly discuss the main flow here.

Step 1: By the definition of the gradient of the Moreau envelope, it holds that

$$\|\nabla\Phi_S^\lambda(x) - \nabla\Phi^\lambda(x)\| \le \frac{1}{\lambda}\|\mathbf{prox}_{\lambda\Phi}(x) - \mathbf{prox}_{\lambda\Phi_S}(x)\|.$$

We first use a ρ-net $\{x_k\}_{k=1}^Q$ (Vapnik, 1999) to handle the dependence issue between $\tilde{x}^* \in \operatorname{argmax}_{x\in\mathcal{X}} \|\mathbf{prox}_{\lambda\Phi}(x) - \mathbf{prox}_{\lambda\Phi_S}(x)\|$ and Φ_S.

Step 2: We introduce the following ℓ_2-regularized minimax problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\ F(x,y) - \frac{\nu}{2}\|y\|^2.$$

Notice that this problem is NC-SC. We further build a connection between NC-C stochastic minimax optimization problems and the corresponding regularized NC-SC stochastic minimax optimization problems. Then we carefully choose the regularization parameter ν to derive the uniform convergence.

The following lemma characterizes the distance between the proximal points of the primal function from the original NC-C problem and its regularized NC-SC problem. Note that the lemma may be of independent interest for the design and the analysis of gradient-based methods for NC-C problems.

Lemma 3.3. Let $\nu > 0$ and denote

$$\hat{\Phi}(x) = \max_{y\in\mathcal{Y}} F(x,y) - \frac{\nu}{2}\|y\|^2$$

as the primal function of the regularized NC-C problem. It holds for $\lambda \in (0, (L+\nu)^{-1})$ that

$$\|\mathbf{prox}_{\lambda\Phi}(x) - \mathbf{prox}_{\lambda\hat{\Phi}}(x)\|^2 \le \frac{\nu D_{\mathcal{Y}}\lambda}{1-\lambda(L+\nu)}.$$

The above lemma implies that for a small regularization parameter ν, the difference between the proximal point of the primal function Φ of the NC-C objective function and the primal function $\hat{\Phi}$ of the regularized NC-SC problem is small.

Step 3: It remains to characterize the distance between $\mathbf{prox}_{\lambda\hat{\Phi}}(x)$ and $\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)$, where $\hat{\Phi}_S$ is the primal function of the regularized empirical minimax problem. By definition of $\mathbf{prox}_{\lambda\hat{\Phi}}(x)$ and $\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)$, the distance is equivalent to the difference between the optimal solutions of a strongly-convex strongly-concave (SC-SC) population minimax problem and its corresponding empirical problem. We utilize the existing stability-based results for SC-SC minimax optimization Zhang et al. (2021a) to identify the distance. Later we further apply the sub-Gaussian random variable argument and concentration inequality to imply the final uniform convergence conclusion. □

3.3 Discussion: Comparing Minimization and Minimax Optimization

For stochastic nonconvex minimization problems $\min_{x\in\mathcal{X}} \mathbb{E}[f(x;\xi)]$, the sample complexity of achieving ϵ-uniform convergence between the gradients of the population problem and the empirical problem, i.e.,

$$\mathbb{E}\max_{x\in\mathcal{X}}\|\frac{1}{n}\sum_{i=1}^n \nabla f(x;\xi_i) - \mathbb{E}\nabla f(x;\xi)\| \le \epsilon,$$

is $\tilde{\mathcal{O}}(d\epsilon^{-2})$ (Davis and Drusvyatskiy, 2022; Mei et al., 2018). For nonconvex minimax optimization, if we only care about the uniform convergence in terms of the gradient of F, i.e.,

$$\mathbb{E}\max_{x\in\mathcal{X},y\in\mathcal{Y}}\|\frac{1}{n}\sum_{i=1}^n \nabla f(x,y;\xi_i) - \mathbb{E}\nabla f(x,y;\xi)\|,$$

where ∇f denotes the full gradient with respect to x and y, existing analysis in Mei et al. (2018) directly gives a $\tilde{\mathcal{O}}(d\epsilon^{-2})$ sample complexity. However, since we consider the generalization measured by primal gradients here, the analysis becomes more complicated, which we detail in the following.

First, in the NC-SC setting, to establish uniform convergence, we bound

$$\begin{aligned}&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S(x) - \nabla\Phi(x)\| \\ &= \mathbb{E}\max_{x\in\mathcal{X}}\Big\|\frac{1}{n}\sum_{i=1}^n \nabla_x f(x, y_S^*(x);\xi_i) - \mathbb{E}\nabla_x f(x, y^*(x);\xi_i)\Big\|.\end{aligned}$$

The primal function Φ_S is not in the form of averaging over n samples, thus existing analysis for minimization problems is not directly applicable. In addition, the difference between the optimal points $y_S^*(x)$ and $y^*(x)$ brings in an additional error term. In the NC-SC case, the error is upper bounded by $\mathcal{O}(n^{-1/2})$, which is the same scale as the error from establishing uniform convergence on x. Thus, the final uniform convergence established in Theorem 3.1 is of the same order as that

for minimization problem (Mei et al., 2018; Davis and Drusvyatskiy, 2022) except for an additional dependence on the condition number κ.

Moreover, in the NC-C case, since there may exist multiple dual maximizers, the distance between y^* and y_S^* may not be well-defined. Instead, we bound the distance between $\hat{y}_S^*(x)$ and $\hat{y}^*(x)$, where

$$\hat{y}_S^*(x) \triangleq \operatorname*{argmax}_{y\in\mathcal{Y}} F_S(x,y) - \frac{\nu}{2}\|y\|^2,$$
$$\hat{y}^*(x) \triangleq \operatorname*{argmax}_{y\in\mathcal{Y}} F(x,y) - \frac{\nu}{2}\|y\|^2,$$

with a small enough $\nu = \mathcal{O}(n^{-1/2})$. The distance can be controlled by $\mathcal{O}(n^{-1/4})$. Thus, the sample complexity for achieving ϵ-uniform convergence for the NC-C case is larger than that of the NC-SC case. We leave it for future investigation to see if one could achieve smaller sample complexity via a better characterization of the extra error brought in by y in the NC-C setting.

4 Algorithmic Stability and Generalization Bounds

Notice that the uniform convergence in Theorems 3.1 and 3.2 has a dependence on the dimension d, which can be vacuous for high-dimensional problems (Lei, 2022; Feldman and Vondrak, 2019). It remains interesting to build dimension-independent generalization results utilizing the special structure of the algorithms. In this section, we investigate the generalization performance of specific algorithms for nonconvex stochastic minimax optimization problems utilizing stability arguments.

4.1 Stability and Generalization

Existing literature on stability arguments in minimax optimization often rely on stability notions based on function values (Farnia and Ozdaglar, 2021; Lei et al., 2021; Zhang et al., 2021a). In order to derive bounds on the generalization in terms of primal stationarity, we introduce the following novel notions of uniform stability on gradients of the primal function, called *uniform primal stability*.

Definition 4.1 (Uniform Primal Stability)**.** A randomized algorithm $\mathcal{A}$ is δ-uniformly primal stable if for every two neighboring dataset S, S' which differ in only one sample, we have

$$\sup_{\xi} \mathbb{E}_{\mathcal{A}}\|\nabla f(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S)); \xi) - \nabla f(\mathcal{A}_x(S'), y^*(\mathcal{A}_x(S')); \xi)\|^2 \le \delta^2.$$

The following theorem connects stability and generalization in minimax optimization problems. We defer the proof to Appendix D.

Theorem 4.2 (Stability and Generalization, NC-SC)**.** Let $\mathcal{A}$ be a δ-uniformly primal stable algorithm. For any function f satisfying Assumption 2.1 with $\mu > 0$, we have

$$\mathbb{E}_{\mathcal{A},S}\|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\| \le (1+\kappa)\big(4\delta + \frac{G}{\sqrt{n}}\big).$$

To the best of our knowledge, this is the first result that connects uniformly stable algorithms and generalization errors in minimax optimization, while measured by the primal stationarity. As a comparison, in the minimization case, Lei (2022, Theorem 2) proved that the gap between the empirical and population gradients is $\mathcal{O}(\delta + 1/\sqrt{n})$, while Theorem 4.2 has an additional dependence on the condition number κ that comes from the minimax structure.

In the NC-C case, the uniform primal stability above in Definition 4.1 is less meaningful as $y^*(\cdot)$ is not well-defined. Instead, we use the following notion of *uniform primal argument stability*.

Definition 4.3 (Uniform Primal Argument Stability)**.** A randomized algorithm $\mathcal{A}$ is δ-uniformly primal argument stable if for every two dataset S, S' which differ in only one sample, it holds that

$$\mathbb{E}_{\mathcal{A}}\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\|^2 \le \delta^2.$$

It is easy to see that the uniform primal argument stability here in Definition 4.3 implies the uniform primal stability in Definition 4.1 applied in the NC-SC case. The following theorem connects argument stability and generalization in NC-C case, measured by primal Moreau envelope stationarity.

Theorem 4.4 (Stability and Generalization, NC-C)**.** Let $\mathcal{A}$ be a δ-uniformly primal argument stable algorithm. For any function f satisfying Assumption 2.1 with $\mu = 0$, we have

$$\mathbb{E}_{\mathcal{A},S}\Big\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\Big\| \le \mathcal{O}(\delta^{1/6} + n^{-1/8}).$$

We defer the proof to Appendix E. Note that the analysis also leverages the idea of adding regularization to create a surrogate NC-SC problem, as we did in Section 3.2. This result yields the relationship between stability and generalization in NC-C problems measured by primal stationarity. Different from the minimization case, the perturbation on the dataset incurs errors on both the function gradients and the dual maximizers, which requires more careful analysis to derive the final generalization bound.

As a comparison, Ozdaglar et al. (2022) proposed another modified metric based on the primal function value gap to characterize the generalization error in the NC-C case, while they still require that the nonconvex primal function is solved exactly, and our work circumvents the restriction by using the primal stationarity for the measurement. It is an interesting open problem to study the relationship between our stationarity-based generalization measurement and those function-value-based ones in the literature.

With Theorems 4.2 and 4.4, to obtain the generalization bounds of algorithms designed for NC-SC and NC-C minimax optimization problems, it suffices to derive the stability of specific algorithms.

4.2 Generalization of Stochastic Gradient Descent Ascent (SGDA)

In this subsection, we study the generalization bounds of the classical *stochastic gradient descent ascent* (Nemirovski, 2004; Lin et al., 2020a) for minimax optimization problems in both NC-SC and NC-C cases. Recall the procedures of SGDA: in each iteration t,

$$\begin{cases} x_{t+1} = \mathbf{proj}_{\mathcal{X}}(x_t - \alpha_t^x \nabla_x f(x_t, y_t; \xi_t)), \\ y_{t+1} = \mathbf{proj}_{\mathcal{Y}}(y_t + \alpha_t^y \nabla_y f(x_t, y_t; \xi_t)), \end{cases}$$

where (α_t^x, α_t^y) are the stepsizes. Farnia and Ozdaglar (2021) investigated the δ-stability of SGDA. Together with Theorems 4.2 and 4.4, we have the following generalization errors in NC-SC and NC-C cases, respectively.

Corollary 4.5 (Generalization of SGDA, NC-SC)**.** Assume the function f is NC-SC as defined in Assumption 2.1 with $\mu > 0$, then if we run SGDA for T iterations with stepsize $(\alpha_t^x, \alpha_t^y) = \left(\frac{c}{t}, \frac{cr^2}{t}\right)$ for some constant $c > 0$ and $1 \le r < \kappa$, let $\zeta_1 = (cL(r+1)+1)^{-1}$, we have

$$\mathbb{E}_{S,\mathcal{A}} \|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\| \le \mathcal{O}\left(\kappa^{1+\zeta_1}\left(\frac{T^{1-\zeta_1}}{n} + \frac{1}{\sqrt{n}}\right)\right),$$

where $(\mathcal{A}_x(S), \mathcal{A}_y(S)) = (x_T, y_T)$ is the output of SGDA.

Corollary 4.6 (Generalization of SGDA, NC-C)**.** Assume the function f is NC-C as defined in Assumption 2.1 with $\mu = 0$, then if we run SGDA for T iterations with stepsize $\max\{\alpha_t^x, \alpha_t^y\} \le \frac{c}{t}$ for some constant $c > 0$, let $\zeta_2 = (cL+1)^{-1}$ then we have

$$\mathbb{E}_{S,\mathcal{A}} \left\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\right\| \le \mathcal{O}\big((T^{1-\zeta_2}/n)^{1/6} + n^{-1/8}\big),$$

where $(\mathcal{A}_x(S), \mathcal{A}_y(S)) = (x_T, y_T)$ is the output of SGDA.

The proof relies on the stability results in Farnia and Ozdaglar (2021), which we defer to Appendix F. Compared to the generalization bounds in Theorems 3.1 and 3.2 that use uniform convergence, the generalization bounds of SGDA avoid the dependence on the dimension d. However, the dependence on n of generalization bounds of SGDA becomes worse compared to uniform convergence in the NC-C setting. We leave the improvement of the current results as a future direction.

4.3 Generalization of Sampling-determined Algorithms (SDA)

Another class of algorithms we consider is the *sampling-determined algorithm* (SDA) proposed in Lei (2022), which covers a wide range of algorithms including SGDA, stochastic extragradient, and some adaptive variants of SGDA (Yang et al., 2022a). Its definition is presented below for completeness.

Definition 4.7 (Sampling-determined Algorithm (Lei, 2022))**.** Let $\mathcal{A}$ be an algorithm that randomly chooses an index sequence $I(\mathcal{A}) = \{i_t\}$ from the dataset to build stochastic gradients. We say $\mathcal{A}$ is *sampling-determined* if its output is independent of the sample ξ_i for any $i \notin I(\mathcal{A})$.

Note that the SDA algorithm class does not include some sophisticated common algorithms like SVRGDA (Palaniappan and Bach, 2016). In Lei (2022), the δ-stability of SDA is derived using its sampling-determined property. Equipped with Theorem 4.2, we obtain the following generalization error bounds of SDA n NC-SC and NC-C cases, respectively.

Corollary 4.8 (Generalization of SDA, NC-SC)**.** Assume the function f is NC-SC as defined in Assumption 2.1 with $\mu > 0$. If we run a SDA algorithm $\mathcal{A}$ for T iterations, we have

$$\mathbb{E}_{S,\mathcal{A}} \|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\| \le \mathcal{O}\Big(\kappa\big(\sqrt{\frac{T}{n}} + \frac{1}{\sqrt{n}}\big)\Big).$$

Corollary 4.9 (Generalization of SDA, NC-C)**.** Assume the function f is NC-C as defined in Assumption 2.1 with $\mu = 0$. If we run a SDA algorithm $\mathcal{A}$ for T iterations, we have

$$\mathbb{E}_{S,\mathcal{A}} \left\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\right\| \le \mathcal{O}\big((T/n)^{1/12} + n^{-1/8}\big).$$

Compared with Corollary 4.5 and 4.6, the generalization bound of SDA does not require specific stepsizes

and applies to a wider class of algorithms, but the generalization bound of SDA algorithm has a worse dependence on sample size n. We leave lifting the specific stepsize requirements of SGDA in Corollary 4.5 as an interesting future direction.

5 Conclusion and Future Directions

In this work, we take an initial step toward understanding the generalization performances of NC-SC and NC-C stochastic minimax problems, measured by the first-order stationarity of the primal functions. Our study covers both uniform convergence and algorithmic stability argument perspectives.

Several future directions are worth further investigation. First it remains interesting to see whether we can improve the uniform convergence and stability results under the NC-C setting, particularly the dependence on sample size n. Also it is an interesting problem to relax existing assumptions in this work, like the bounded gradient norm and the compact domain, recently there have appeared several ways to relax the bounded gradient assumption, e.g., instance-dependent Lipschitz continuity in Davis and Drusvyatskiy (2022) and Bernstein condition in Klochkov and Zhivotovskiy (2021). Another possible direction is to investigate the generalization performances for specific applications, in fact some recent studies in stochastic minimization show that specific machine learning models (e.g., generalized linear models) enjoy dimension-free uniform convergence bounds (Amir et al., 2022; Davis and Drusvyatskiy, 2022). It would be interesting to see whether such dimension-free uniform convergence property also holds for some minimax applications.

Acknowledgements

S.Z. and Y.H. contributed equally to this work. Y.H. gratefully acknowledges funding support from NCCR Automation in Switzerland, L.Z. gratefully acknowledges funding by the Max Planck ETH Center for Learning Systems (CLS). N.H. is supported by ETH research grant and Swiss National Science Foundation Project Funding No. 200021-207343.

References

Amir, I., Livni, R., and Srebro, N. (2022). Thinking outside the ball: Optimal learning with gradient descent for generalized linear stochastic convex optimization. *Advances in Neural Information Processing Systems*, 35:23539–23550.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. (2020). Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33.

Boob, D. and Guzmán, C. (2023). Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming*, pages 1–43.

Boţ, R. I. and Böhm, A. (2020). Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*.

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

Dai, B., He, N., Pan, Y., Boots, B., and Song, L. (2017). Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pages 1458–1467.

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. (2018). SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134.

Davis, D. and Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239.

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

Farnia, F. and Ozdaglar, A. (2021). Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pages 3174–3185. PMLR.

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

Fisher, R. A. (1922). 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.

Foster, D. J., Sekhari, A., and Sridharan, K. (2018). Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. (2021). A novel convergence analysis for algorithms

of the adam family and beyond. *arXiv preprint arXiv:2104.14840*.

Han, Y., Xie, G., and Zhang, Z. (2021). Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR.

Hu, Y., Chen, X., and He, N. (2020). Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133.

Huang, J. and Jiang, N. (2022). On the convergence rate of off-policy policy optimization methods with density-ratio correction. In *International Conference on Artificial Intelligence and Statistics*, pages 2658–2705. PMLR.

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

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

Lei, Q., Lee, J., Dimakis, A., and Daskalakis, C. (2020). Sgd learns one-layer networks in wgans. In *International Conference on Machine Learning*, pages 5799–5808. PMLR.

Lei, Y. (2022). Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. *arXiv preprint arXiv:2206.07082*.

Lei, Y., Yang, Z., Yang, T., and Ying, Y. (2021). Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR.

Li, H., Tian, Y., Zhang, J., and Jadbabaie, A. (2021). Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34.

Lin, T., Jin, C., and Jordan, M. (2020a). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR.

Lin, T., Jin, C., and Jordan, M. I. (2020b). Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. (2020). Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691.

Luo, L., Ye, H., Huang, Z., and Zhang, T. (2020). Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

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

Nemirovski, A. (2004). Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32:14934–14942.

Ostrovskii, D. M., Lowy, A., and Razaviyayn, M. (2021). Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538.

Ozdaglar, A., Pattathil, S., Zhang, J., and Zhang, K. (2022). What is a good metric to study generalization of minimax learners? *Advances in Neural Information Processing Systems*, 35:38190–38203.

Palaniappan, B. and Bach, F. (2016). Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424.

Rafique, H., Liu, M., Lin, Q., and Yang, T. (2021). Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. (2018). On the convergence and robustness of training gans with regularized optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7091–7101.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. In *COLT*, volume 2, page 5.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform

convergence. *The Journal of Machine Learning Research*, 11:2635–2670.

Shapiro, A. (2006). On complexity of multistage stochastic programs. *Operations Research Letters*, 34(1):1–8.

Sharma, P., Panda, R., Joshi, G., and Varshney, P. (2022). Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR.

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. (2019). Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32:12680–12691.

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

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

Wang, D., Ye, M., and Xu, J. (2017). Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. (2019). On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595. PMLR.

Xu, Z., Zhang, H., Xu, Y., and Lan, G. (2023). A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, pages 1–72.

Yan, Y., Xu, Y., Lin, Q., Liu, W., and Yang, T. (2020). Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800.

Yang, J., Kiyavash, N., and He, N. (2020a). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33.

Yang, J., Li, X., and He, N. (2022a). Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. *Advances in Neural Information Processing Systems*, 35:11202–11216.

Yang, J., Orvieto, A., Lucchi, A., and He, N. (2022b). Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR.

Yang, J., Zhang, S., Kiyavash, N., and He, N. (2020b). A catalyst framework for minimax optimization. In *Advances in Neural Information Processing Systems*.

Yang, Z., Hu, S., Lei, Y., Vashney, K. R., Lyu, S., and Ying, Y. (2022c). Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*, pages 2192–2202. PMLR.

Zhang, J., Hong, M., Wang, M., and Zhang, S. (2021a). Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR.

Zhang, J., Xiao, P., Sun, R., and Luo, Z. (2020). A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389.

Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. (2021b). The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR.

Zhang, X., Aybat, N. S., and Gurbuzbalaban, M. (2022). Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *Advances in Neural Information Processing Systems*, 35:21668–21681.

Zhao, R. (2020). A primal dual smoothing framework for max-structured nonconvex optimization. *arXiv preprint arXiv:2003.04375*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Additional Definitions and Tools

For convenience, we summarize the notations commonly used throughout the paper.

- Population minimax problem and its primal function[2]

$$F(x,y) \triangleq \mathbb{E}_\xi\, f(x,y;\xi), \quad \Phi(x) \triangleq \max_{y\in\mathcal{Y}} F(x,y), \quad y^*(x) \triangleq \operatorname*{argmax}_{y\in\mathcal{Y}} F(x,y).$$

- Empirical minimax problem and its primal function

$$F_S(x,y) \triangleq \frac{1}{n}\sum_{i=1}^n f(x,y;\xi_i), \quad \Phi_S(x) \triangleq \max_{y\in\mathcal{Y}} F_S(x,y), \quad y_S^*(x) \triangleq \operatorname*{argmax}_{y\in\mathcal{Y}} F_S(x,y).$$

- Moreau envelope and corresponding proximal point:

$$\Phi^\lambda(x) \triangleq \min_{z\in\mathcal{X}}\left\{\Phi(z)+\frac{1}{2\lambda}\|z-x\|^2\right\}, \quad \mathbf{prox}_{\lambda\Phi}(x) \triangleq \operatorname*{argmin}_{z\in\mathcal{X}}\left\{\Phi(z)+\frac{1}{2\lambda}\|z-x\|^2\right\},$$
$$\Phi_S^\lambda(x) \triangleq \min_{z\in\mathcal{X}}\left\{\Phi_S(z)+\frac{1}{2\lambda}\|z-x\|^2\right\}, \quad \mathbf{prox}_{\lambda\Phi_S}(x) \triangleq \operatorname*{argmin}_{z\in\mathcal{X}}\left\{\Phi_S(z)+\frac{1}{2\lambda}\|z-x\|^2\right\}.$$

- $\mathcal{G}_\Phi(x)$: gradient mapping (generalized gradient) of a function Φ.
- $\|\cdot\|$: ℓ_2-norm.
- $\nabla f = (\nabla_x f, \nabla_y f)$: the gradient of a function f.
- $\mathbf{proj}_{\mathcal{X}}(x')$: the projection operator.
- $\mathcal{A}(S) \triangleq (\mathcal{A}_x(S), \mathcal{A}_y(S))$: the output of an algorithm $\mathcal{A}$ on the empirical minimax problem (2) with dataset S.
- NC / WC: nonconvex, weakly convex.
- NC-SC / NC-C: nonconvex-(strongly)-concave.
- SOTA: state-of-the-art.
- d: dimension number of $\mathcal{X}$.
- κ: condition number $\frac{L}{\mu}$, L: Lipschitz smoothness parameter, μ: strong concavity parameter.
- $\tilde{O}(\cdot)$ hides poly-logarithmic factors.
- $f = \Omega(g)$ if $f(x) \geq cg(x)$ for some $c > 0$ and nonnegative functions f and g.
- We say a function $g : \mathcal{X} \to \mathbb{R}$ is convex if $\forall\ x_1, x_2 \in \mathcal{X}$ and $p \in [0,1]$, we have $g(px_1 + (1-p)x_2) \geq pg(x_1) + (1-p)g(x_2)$.

For completeness, we introduce the definition of a sub-Gaussian random variable and related lemma, which are important tools in the analysis.

Definition A.1 (Sub-Gaussian Random Variable)**.** A random variable η is a zero-mean sub-Gaussian random

[2]Another commonly used convergence criterion in minimax optimization is the *first-order stationarity of* F, i.e., $\|\nabla_x F\| \leq \epsilon$ and $\|\nabla_y F\| \leq \epsilon$ (or its corresponding gradient mapping) (Lin et al., 2020a; Xu et al., 2023). We refer readers to Lin et al. (2020a); Yang et al. (2022b) for a thorough comparison of these two measurements. In this paper, we always stick to the convergence measured by the stationarity of the primal function.

variable with variance proxy σ_η^2 if $\mathbb{E}\,\eta = 0$ and either of the following two conditions hold:

$$(a)\ \mathbb{E}\left[\exp(s\eta)\right] \le \exp\left(\frac{\sigma_\eta^2 s^2}{2}\right) \text{ for any } s \in \mathbb{R}; \quad (b)\ \mathbb{P}(|\eta| \ge t) \le 2\exp\left(-\frac{t^2}{2\sigma_\eta^2}\right) \text{ for any } t > 0.$$

We use the following McDiarmid's inequality to show that a random variable is sub-Gaussian.

Lemma A.2 (McDiarmid's inequality)**.** Let $\eta_1, \ldots, \eta_n \in \mathbb{R}$ be independent random variables. Let $h : \mathbb{R}^n \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_n)$-bounded differences property: for every $i = 1, \ldots, n$ and every $(\eta_1, \ldots, \eta_n)$, and $(\eta_1', \ldots, \eta_n')$ that differ only in the i-th coordinate ($\eta_j = \eta_j'$ for all $j \ne i$), we have

$$|h(\eta_1, \ldots, \eta_n) - h(\eta_1', \ldots, \eta_n')| \le c_i.$$

For any $t > 0$, it holds that

$$\mathbb{P}(|h(\eta_1, \ldots, \eta_n) - \mathbb{E}\,h(\eta_1, \ldots, \eta_n)| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma A.3 (Properties of Φ and Φ^λ, Full Version)**.** In the NC-SC setting ($\mu > 0$), both $\Phi(x)$ and $\Phi_S(x)$ are $\tilde{L} \triangleq L(1+\kappa)$-smooth with the condition number $\kappa \triangleq L/\mu$, both $y^*(x)$ and $y_S^*(x)$ are κ-Lipschitz continuous and $\nabla\Phi(x) = \nabla_x F(x, y^*(x)), \nabla\Phi_S(x) = \nabla_x F_S(x, y_S^*(x))$. In the NC-C setting ($\mu = 0$), the primal function Φ is L-weakly convex, and its its Moreau envelope $\Phi^\lambda(x)$ is differentiable, Lipschitz smooth, also

$$\nabla\Phi^\lambda(x) = \lambda^{-1}(x - \hat{x}), \quad \left\|\nabla\Phi^\lambda(x)\right\| \ge \mathbf{dist}\left(0, \partial\Phi(\hat{x})\right), \tag{4}$$

where $\hat{x} = \mathbf{prox}_{\lambda\Phi}(x)$ and $0 < \lambda < 1/L$.

For completeness, we formally define the stationary point here. Note that the generalized gradient is defined on $\mathcal{X}$ while the Moreau envelope is defined on the whole domain $\mathbb{R}^d$.

Definition A.4 (Stationary Point)**.** Let $\epsilon > 0$, for an $\tilde{L}$-smooth function $\Phi : \mathcal{X} \to \mathbb{R}$, we call a point x an ϵ-stationary point of Φ if $\|\mathcal{G}_\Phi(x)\| \le \epsilon$, where $\mathcal{G}_\Phi$ is the gradient mapping (or generalized gradient) defined as $\mathcal{G}_\Phi(x) \triangleq \tilde{L}\Big(x - \mathbf{proj}_{\mathcal{X}}\Big(x - (1/\tilde{L})\nabla\Phi(x)\Big)\Big)$; for an L-weakly convex function Φ, we say a point x an ϵ-(nearly)-stationary point of Φ if $\left\|\nabla\Phi^{1/(2L)}(x)\right\| \le \epsilon$.

B Proof of Theorem 3.1

Proof. To derive the desired generalization bounds, we take an ρ-net $\{x_k\}_{k=1}^Q$ on $\mathcal{X}$ so that there exists a $k \in \{1, \cdots, Q\}$ for any $x \in \mathcal{X}$ such that $\|x - x_k\| \le \rho$. Note that such ρ-net exists with $Q = \mathcal{O}(\rho^{-d})$ for compact $\mathcal{X}$ (Kleywegt et al., 2002). Utilizing the definition of the ρ-net, we have

$$\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}} \|\nabla\Phi_S(x) - \nabla\Phi(x)\| \\
\le\ &\mathbb{E}\max_{x\in\mathcal{X}} [\|\nabla\Phi_S(x) - \nabla\Phi_S(x_k)\| + \|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| + \|\nabla\Phi(x_k) - \nabla\Phi(x)\|] \\
\le\ &\mathbb{E}\max_{k\in[Q]} \|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| + 2L(1+\kappa)\rho,
\end{aligned} \tag{5}$$

where the last inequality holds as Φ and Φ_S are $L(1+\kappa)$-smooth following Lemma 2.2. For any $s>0$, we have

$$\begin{aligned}
&\exp\Big(s\,\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S(x)-\nabla\Phi(x)\|\Big)\\
\le\ &\exp\Big(s\Big[\mathbb{E}\max_{k\in[Q]}\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\rho\Big]\Big)\\
\le\ &\mathbb{E}\max_{k\in[Q]}\exp\Big(s\Big[\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\rho\Big]\Big)\\
\le\ &\mathbb{E}\sum_{k\in[Q]}\exp\Big(s\Big[\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\rho\Big]\Big)\\
=\ &\sum_{k\in[Q]}\mathbb{E}\exp\Big(s\Big[\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\rho\Big]\Big),
\end{aligned}\tag{6}$$

where the second inequality uses Jensen's inequality and monotonicity of exponential function, and the third inequality uses summation over $k\in[Q]$ to handle the dependence issue, i.e., the x_k in the last line is independent of S. We use the exponential function as an intermediate step so that the final sample complexity depends on $\log(Q)$ rather than Q, which is of order $\mathcal{O}(\rho^{-d})$. Without loss of generality, selecting ρ such that $2L(1+\kappa)\rho=\frac{\epsilon}{2}$, we have

$$\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S(x)-\nabla\Phi(x)\|\\
\le\ &\frac{1}{s}\log\Bigg(\sum_{k\in[Q]}\mathbb{E}\exp\left(s[\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|-\mathbb{E}\,\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|]\right)\\
&\qquad\cdot\exp\left(s\,\mathbb{E}\,\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|\right)\exp\Big(\frac{s\epsilon}{2}\Big)\Bigg).
\end{aligned}\tag{7}$$

To upper bound $\mathbb{E}\,\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|$, we use the following observation. Define $y^*_{S^{(i)}}(x)\triangleq\operatorname{argmax}_{y\in\mathcal{Y}}F_{S^{(i)}}(x,y)$ where $S=\{\xi_i\}_{i=1}^n$, $S^{(i)}=\{\xi_1,\dots,\xi_{i-1},\xi_i',\xi_{i+1},\dots,\xi_n\}$ and ξ_i' is i.i.d. from ξ_i. Since x is independent of S or $S^{(i)}$ for any i, by Danskin's theorem, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla\Phi(x)-\nabla\Phi_S(x)\|=\mathbb{E}\left\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi)-\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y_S^*(x);\xi_i)\right\|\\
=\ &\mathbb{E}\Bigg\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi)-\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y^*(x);\xi_i)\\
&\qquad+\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y^*(x);\xi_i)-\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y_S^*(x);\xi_i)\Bigg\|\\
\le\ &\mathbb{E}\left\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi)-\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y^*(x);\xi_i)\right\|\\
&\qquad+\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y^*(x);\xi_i)-\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y_S^*(x);\xi_i)\right\|\\
\le\ &\mathbb{E}\left\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi)-\frac{1}{n}\sum_{i=1}^n\nabla_x f(x,y^*(x);\xi_i)\right\|+L\,\mathbb{E}\,\|y^*(x)-y_S^*(x)\|\\
\le\ &\frac{G}{\sqrt{n}}+L\,\mathbb{E}\,\|y^*(x)-y_S^*(x)\|,
\end{aligned}\tag{8}$$

where the second inequality holds by the smoothness of f, and the last inequality holds because the variance is upper bounded by the second moment. To derive an upper bound on $\|y^*(x)-y_S^*(x)\|$, we first bound $\left\|y^*_{S^{(i)}}(x)-y_S^*(x)\right\|$ and utilize the stability argument. Since $f(x,y;\xi)$ is μ-strongly concave in y for any x and ξ and $y_S^*(x)$ is the maximizer of $F_S(x,\cdot)$, we have

$$\big(-F_S\big(x,y^*_{S^{(i)}}(x)\big)\big)-\big(-F_S(x,y_S^*(x))\big)\ge\frac{\mu}{2}\left\|y^*_{S^{(i)}}(x)-y_S^*(x)\right\|^2,\tag{9}$$

On the other hand, we have

$$\begin{aligned}
&F_S(x, y_S^*(x)) - F_S(x, y_{S^{(i)}}^*(x)) \\
=&F_{S^{(i)}}(x, y_S^*(x)) - F_{S^{(i)}}(x, y_{S^{(i)}}^*(x)) \\
&\quad + \frac{1}{n}\Big[f(x, y_S^*(x); \xi_i) - f(x, y_{S^{(i)}}^*(x); \xi_i) + f(x, y_{S^{(i)}}^*(x); \xi_i') - f(x, y_S^*(x); \xi_i')\Big] \\
\leq&F_{S^{(i)}}(x, y_S^*(x)) - F_{S^{(i)}}(x, y_{S^{(i)}}^*(x)) \\
&\quad + \frac{1}{n}\Big|f(x, y_{S^{(i)}}^*(x); \xi_i) - f(x, y_S^*(x); \xi_i)| + \frac{1}{n}|f(x, y_{S^{(i)}}^*(x); \xi_i') - f(x, y_S^*(x); \xi_i')\Big| \\
\leq&\frac{2G}{n}\big\|y_{S^{(i)}}^*(x) - y_S^*(x)\big\|,
\end{aligned}$$

where the last inequality holds by Lipschitz continuity and the optimality of $y_{S^{(i)}}^*(x)$. Combined with (9), it holds that

$$\Big\|y_{S^{(i)}}^*(x) - y_S^*(x)\Big\| \leq \frac{4G}{\mu n}.$$

In addition, we have

$$\begin{aligned}
&\mathbb{E}[F(x, y^*(x)) - F(x, y_S^*(x))] \\
=\ &\mathbb{E}\left[F(x, y^*(x)) - F_S(x, y^*(x))\right] + \mathbb{E}\left[F_S(x, y^*(x)) - F_S(x, y_S^*(x))\right] \\
&\qquad + \mathbb{E}\left[F_S(x, y_S^*(x)) - F(x, y_S^*(x))\right] \\
\leq\ &\mathbb{E}\left[F_S(x, y_S^*(x)) - F(x, y_S^*(x))\right] \\
=\ &\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x, y_S^*(x); \xi_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\xi\, f(x, y_S^*(x); \xi)\right] \\
=\ &\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x, y_S^*(x); \xi_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi_i}\, f(x, y_{S^{(i)}}^*(x); \xi_i)\right] \\
=\ &\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x, y_S^*(x); \xi_i) - \frac{1}{n}\sum_{i=1}^n f(x, y_{S^{(i)}}^*(x); \xi_i)\right] \\
\leq\ &G\,\mathbb{E}\left\|y_S^*(x) - y_{S^{(i)}}^*(x)\right\| \\
\leq\ &\frac{4G^2}{\mu n}
\end{aligned} \tag{10}$$

where the first inequality holds as $y_S^*(x) = \operatorname{argmax}_{y\in\mathcal{Y}} F_S(x, y_S^*(x))$ and $\mathbb{E}[F(x, y^*(x)) - F_S(x, y^*(x))] = 0$, the third equality holds as $y_S^*(x)$ and $y_{S^{(i)}}^*(x)$ are identical distributed and $y_{S^{(i)}}^*(x)$ is independent of ξ by definition, the second inequality holds by Lipschitz continuity of f on y, and the last inequality holds by plugging the upper bound on $\|y_S^*(x) - y_{S^{(i)}}^*(x)\|$. On the other hand, since $F(x, y)$ is strongly concave in y and $y^*(x) = \operatorname{argmax}_{y\in\mathcal{Y}} F(x, y)$, it holds that

$$F(x, y^*(x)) - F(x, y_S^*(x)) \geq \frac{\mu}{2}\|y^*(x) - y_S^*(x)\|^2.$$

Therefore, we have

$$\mathbb{E}\,\|y^*(x) - y_S^*(x)\| \leq \sqrt{\frac{8G^2}{\mu^2 n}}.$$

Plugging into (8), it holds that

$$\mathbb{E}\,\|\nabla\Phi(x_k) - \nabla\Phi_S(x_k)\| \leq L\sqrt{\frac{8G^2}{\mu^2 n}} + \frac{G}{\sqrt{n}}. \tag{11}$$

Next we show that $\|\nabla\Phi(x) - \nabla\Phi_S(x)\| - \mathbb{E}\,\|\nabla\Phi(x) - \nabla\Phi_S(x)\|$ is zero-mean sub-Gaussian. Notice that for any ξ_i', we have

$$\begin{aligned}
&\|\nabla\Phi(x) - \nabla\Phi_S(x)\| - \|\nabla\Phi(x) - \nabla\Phi_{S^{(i)}}(x)\| \\
\le\ &\|\nabla\Phi_S(x) - \nabla\Phi_{S^{(i)}}(x)\| \\
=\ &\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla_x f(x, y_S^*(x), \xi_j) - \frac{1}{n}\sum_{j\neq i}^{n}\nabla_x f\big(x, y_{S^{(i)}}^*(x), \xi_j\big) - \frac{1}{n}\nabla_x f\big(x, y_{S^{(i)}}^*(x), \xi_i'\big)\right\| \\
\le\ &L\big\|y_{S^{(i)}}^*(x) - y_S^*(x)\big\| + \frac{1}{n}\big\|\nabla_x f(x, y_{S^{(i)}}^*(x); \xi_i') - \nabla_x f(x, y_{S^{(i)}}^*(x); \xi_i)\big\| \\
\le\ &\frac{4LG/\mu + 2G}{n},
\end{aligned} \tag{12}$$

where the first inequality uses triangle inequality, the first equality uses the definition of Φ_S and $\Phi_{S^{(i)}}$, the third inequality uses the assumption that G is the uniform upper bound of $\nabla f(x, y; \xi)$ on $\mathcal{X}\times\mathcal{Y}$ for any ξ. By McDiarmid's inequality (Lemma A.2) and the definition of sub-Gaussian random variables, it holds that $\|\nabla\Phi(x_k) - \nabla\Phi_S(x_k)\| - \mathbb{E}\,\|\nabla\Phi(x_k) - \nabla\Phi_S(x_k)\|$ is a zero-mean sub-Gaussian random variable with variance proxy $\sigma^2 \triangleq (2LG/\mu + G)^2/n$. By the definition of zero-mean sub-Gaussian random variables, it holds that

$$\mathbb{E}\exp(s[\|\nabla\Phi(x_k) - \nabla\Phi_S(x_k)\| - \mathbb{E}\,\|\nabla\Phi(x_k) - \nabla\Phi_S(x_k)\|]) \le \exp\left(\frac{s^2\sigma^2}{2}\right). \tag{13}$$

Plugging (11) and (13) into (7), we have

$$\mathbb{E}\,\|\nabla\Phi_S(x) - \nabla\Phi(x)\| \le \frac{log(Q)}{s} + \frac{s\sigma^2}{2} + L\sqrt{\frac{8G^2}{\mu^2 n}} + \frac{G}{\sqrt{n}} + \frac{\epsilon}{2} \tag{14}$$

Minimizing the right-hand side over s, we have

$$\begin{aligned}
\mathbb{E}\,\|\nabla\Phi_S(x) - \nabla\Phi(x)\| &\le 2\sqrt{\frac{log(Q)\sigma^2}{2}} + L\sqrt{\frac{8G^2}{\mu^2 n}} + \frac{G}{\sqrt{n}} + \frac{\epsilon}{2} \\
&= \sqrt{\frac{2log(Q)(2LG/\mu + G)^2}{n}} + L\sqrt{\frac{8G^2}{\mu^2 n}} + \frac{G}{\sqrt{n}} + \frac{\epsilon}{2}.
\end{aligned} \tag{15}$$

Recall that $Q = \mathcal{O}(\rho^{-d})$ with $\rho = \epsilon/(4L(1+\kappa))$, thus $\log(Q) = \mathcal{O}(d\log(4L(1+\kappa)\epsilon^{-1}))$, which verifies the first statement in the theorem. For the sample complexity, following the discussion on the performance measurement in Section 2, it is easy to derive that it requires

$$n = \mathcal{O}\left(2d\epsilon^{-2}(2LG/\mu + G)^2\log(4L(1+\kappa)\epsilon^{-1})\right) = \tilde{\mathcal{O}}(d\kappa^2\epsilon^{-2}) \tag{16}$$

to guarantee that $\mathbb{E}\,\|\nabla\Phi_S(x) - \nabla\Phi(x)\| \le \epsilon$ for any $x \in \mathcal{X}$, which concludes the proof. □

C Proof of Theorem 3.2

We first provide the proof of Lemma 3.3.

Proof. Since $F(x, y)$ is L-smooth, it is obvious that $F(x, y) - \frac{\nu}{2}\|y\|^2$ is $(L+\nu)$-smooth. By Thekumparampil et al. (2019, Lemma 3), $\hat{\Phi}(x)$ is $(L+\nu)$-weakly convex in x. Therefore, $\hat{\Phi}(x) + \frac{1}{2\lambda}\|x - x'\|^2$ is $\big(\frac{1}{\lambda} - (L+\nu)\big)$-strongly convex in x for any fixed x'. Denote $\hat{y}(x) \triangleq \operatorname{argmax}_{y\in\mathcal{Y}} F(x, y) - \frac{\nu}{2}\|y\|^2$, $y^*(x) \triangleq \operatorname{argmax}_{y\in\mathcal{Y}} F(x, y)$. It holds

that

$$
\begin{aligned}
&\frac{1}{2}(1/\lambda-(L+\nu))\|\mathbf{prox}_{\lambda\Phi}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|^2\\
&\leq \hat{\Phi}(\mathbf{prox}_{\lambda\Phi}(x))+\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\Phi}(x)-x\|^2-\hat{\Phi}(\mathbf{prox}_{\lambda\hat{\Phi}}(x))-\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-x\|^2\\
&= F(\mathbf{prox}_{\lambda\Phi}(x),\hat{y}(\mathbf{prox}_{\lambda\Phi}(x)))-\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\Phi}(x))\|^2+\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\Phi}(x)-x\|^2\\
&\qquad -F(\mathbf{prox}_{\lambda\hat{\Phi}}(x),\hat{y}(\mathbf{prox}_{\lambda\hat{\Phi}}(x)))+\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\hat{\Phi}}(x))\|^2-\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-x\|^2\\
&\leq F(\mathbf{prox}_{\lambda\Phi}(x),y^*(\mathbf{prox}_{\lambda\Phi}(x)))+\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\Phi}(x)-x\|^2-\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\Phi}(x))\|^2\\
&\qquad -F(\mathbf{prox}_{\lambda\hat{\Phi}}(x),\hat{y}(\mathbf{prox}_{\lambda\hat{\Phi}}(x)))-\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-x\|^2+\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\hat{\Phi}}(x))\|^2\\
&\leq F(\mathbf{prox}_{\lambda\Phi}(x),y^*(\mathbf{prox}_{\lambda\Phi}(x)))+\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\Phi}(x)-x\|^2-\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\Phi}(x))\|^2\\
&\qquad -F(\mathbf{prox}_{\lambda\hat{\Phi}}(x),y^*(\mathbf{prox}_{\lambda\hat{\Phi}}(x)))-\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-x\|^2+\frac{\nu}{2}\|y^*(\mathbf{prox}_{\lambda\hat{\Phi}}(x))\|^2\\
&= \Phi(\mathbf{prox}_{\lambda\Phi}(x))+\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\Phi}(x)-x\|^2-\Phi(\mathbf{prox}_{\lambda\hat{\Phi}}(x))-\frac{1}{2\lambda}\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-x\|^2\\
&\qquad +\frac{\nu}{2}\|y^*(\mathbf{prox}_{\lambda\hat{\Phi}}(x))\|^2-\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\Phi}(x))\|^2\\
&\leq \frac{\nu}{2}\|y^*(\mathbf{prox}_{\lambda\hat{\Phi}}(x))\|^2-\frac{\nu}{2}\|\hat{y}(\mathbf{prox}_{\lambda\Phi}(x))\|^2\\
&\leq \frac{\nu D_{\mathcal{Y}}}{2},
\end{aligned}
\tag{17}
$$

where the first inequality holds by strong convexity of $\hat{\Phi}(z)+\frac{1}{2\lambda}\|z-x\|^2$ and optimality of $\mathbf{prox}_{\lambda\hat{\Phi}}(x)$ for $\min_{z\in\mathcal{X}}\hat{\Phi}(z)+\frac{1}{2\lambda}\|z-x\|^2$, the first equality holds by definition of $\hat{\Phi}$, the second inequality holds by optimality of $y^*(\mathbf{prox}_{\lambda\Phi}(x)))=\operatorname{argmax}_{y\in\mathcal{Y}}F(\mathbf{prox}_{\lambda\Phi}(x),y)$, the third inequality holds by optimality of $\hat{y}(\mathbf{prox}_{\lambda\Phi}(x)))=\operatorname{argmax}_{y\in\mathcal{Y}}F(\mathbf{prox}_{\lambda\Phi}(x),y)-\frac{\nu}{2}\|y\|^2$, the second equality holds by definition of Φ, the fourth inequality holds by optimality of $\mathbf{prox}_{\lambda\Phi}(x)=\operatorname{argmin}_{x\in\mathcal{X}}\{\Phi(z)+\frac{1}{2\lambda}\|z-x\|^2\}$, the last inequality holds by the compactness of domain $\mathcal{Y}$. □

Next, we demonstrate the proof of Theorem 3.2.

Proof. By Lemma 3.3, we have

$$
\begin{aligned}
\|\mathbf{prox}_{\lambda\Phi}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|&\leq\sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}};\\
\|\mathbf{prox}_{\lambda\Phi_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)\|&\leq\sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}}.
\end{aligned}
$$

To derive the desired uniform convergence, similar to the proof of Theorem 3.1, we take an ρ-net $\{x_k\}_{k=1}^Q$ on $\mathcal{X}$ so that there exists a $k\in\{1,\cdots,Q\}$ for any $x\in\mathcal{X}$ such that $\|x-x_k\|\leq\rho$. Note that such ρ-net exists with $Q=\mathcal{O}(\rho^{-d})$ for compact $\mathcal{X}$. We first decompose the error as the approximation error from NC-SC minimax problems to NC-C minimax problems. Then we utilize the ρ-net to address the dependence between S and $\operatorname{argmax}_{x\in\mathcal{X}}\|\nabla\Phi_S^\lambda(x)-\nabla\Phi^\lambda(x)\|$. First, note that

$$
\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S^\lambda(x)-\nabla\Phi^\lambda(x)\|\\
=&\frac{1}{\lambda}\,\mathbb{E}\max_{x\in\mathcal{X}}\|\mathbf{prox}_{\lambda\Phi_S}(x)-\mathbf{prox}_{\lambda\Phi}(x)\|\\
\le&\frac{1}{\lambda}\,\mathbb{E}\max_{x\in\mathcal{X}}\|\mathbf{prox}_{\lambda\Phi_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)\|+\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|\\
&\qquad+\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-\mathbf{prox}_{\lambda\Phi}(x)\|\\
\le&\frac{2}{\lambda}\sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}}+\frac{1}{\lambda}\,\mathbb{E}\max_{x\in\mathcal{X}}\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|\\
\le&\frac{2}{\lambda}\sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}}+\frac{1}{\lambda}\,\mathbb{E}\max_{x\in\mathcal{X}}\big[\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)\|\\
&\qquad+\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\|+\|\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|\big]\\
\le&2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda(L+\nu))}}+\frac{1}{\lambda}\,\mathbb{E}\max_{k\in[Q]}\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\|+\frac{2\rho}{\lambda(1-\lambda(L+\nu))}\\
\le&2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda(L+\nu))}}+\frac{1}{\lambda s}\log\left(\sum_{k\in[Q]}\mathbb{E}\exp\left(s\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\right)\\
&\qquad+\frac{2\rho}{\lambda(1-\lambda(L+\nu))},
\end{aligned}
\tag{18}
$$

where the first and the third inequality use the triangle inequality, the second inequality uses Lemma 3.3 for Φ and Φ_S, x_k is the closest point to x in the ρ-net, the fourth inequality holds by $(1-\lambda(L+\nu))^{-1}$-Lipschitz continuity of proximal operator (Davis and Drusvyatskiy, 2022, Lemma 4.3) since $F(x,y)-\frac{\nu}{2}\|y\|^2$ is a $(L+\nu)$-smooth function, and the last inequality follows a similar argument in (6). All that remains is to bounding $\mathbb{E}\exp\left(s\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\right\|\right)$ for $x\in\mathcal{X}$ that is independent of S. Notice that

$$
\begin{aligned}
&\mathbb{E}\exp\left(s\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\\
=\;&\mathbb{E}\exp\left(s\left[\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|-\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right]\right)\\
&\qquad\cdot\exp\left(s\,\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)
\end{aligned}
$$

Next, we show that $\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|-\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|$ is a zero-mean sub-Guassian random variable and $\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|$ is bounded. Since x_k is independent of S, it is sufficient to show an upper bound of the following term where $x\in\mathcal{X}$ is independent of S.

$$
\mathbb{E}\,\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|.
$$

Recall the definition that

$$
\mathbf{prox}_{\lambda\hat{\Phi}}(x)=\operatorname*{argmin}_{z\in\mathcal{X}}\Big\{\max_{y\in\mathcal{Y}}\mathbb{E}_\xi\, f(z,y;\xi)-\frac{\nu}{2}\|y\|^2+\frac{1}{2\lambda}\|z-x\|^2\Big\},
\tag{19}
$$

$$
\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)=\operatorname*{argmin}_{z\in\mathcal{X}}\left\{\max_{y\in\mathcal{Y}}\frac{1}{n}\sum_{i=1}^n\Big[f(z,y;\xi_i)-\frac{\nu}{2}\|y\|^2+\frac{1}{2\lambda}\|z-x\|^2\Big]\right\}.
\tag{20}
$$

Denote the solution of (19) as $(z^*(x),y^*(x))$ and the solution of (20) as $(z_S(x),y_S(x))$. We need to bound the distance between $z^*(x)$ and $z_S(x)$, note that this $(z^*(x),y^*(x))$ comes from a strongly-convex-strongly-concave stochastic minimax problem, where the modulus is $\frac{1-\lambda L}{\lambda}$ and ν, respectively; while the other comes from the sample average approximation counterpart. By Zhang et al. (2021a, Theorem 1 and Appendix A.1), we have the following results:

$$
\frac{1-\lambda L}{2\lambda}\,\mathbb{E}\,\|z_S(x)-z^*(x)\|^2+\frac{\nu}{2}\,\mathbb{E}\,\|y_S(x)-y^*(x)\|^2\le\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right),
$$

where $\hat{L}_x$ is the Lipschitz continuity parameter of $f(z,y;\xi)+\frac{1}{2\lambda}\|z-x\|^2$ in $z\in\mathcal{X}$ for any given $y\in\mathcal{Y}$ and ξ, and $\hat{L}_y$ is the Lipschitz continuity parameter of $f(z,y;\xi)-\frac{\nu}{2}\|y\|^2$ in $y\in\mathcal{Y}$ for any given $z\in\mathcal{X}$ and ξ. More specifically, since $f(\cdot,\cdot;\xi)$ is G-Lipschitz continuous for any ξ, we have

$$\hat{L}_x \le G+\frac{2\sqrt{D_{\mathcal{X}}}}{\lambda},\quad \hat{L}_y \le G+\nu\sqrt{D_{\mathcal{Y}}}.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\right\| &= \mathbb{E}\left\|z_S(x)-z^*(x)\right\| \\ &\le \sqrt{\mathbb{E}\left\|z_S(x)-z^*(x)\right\|^2} \le \sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}.\end{aligned} \tag{21}$$

Next, we show that $\|z_S(x)-z^*(x)\|-\mathbb{E}\|z_S(x)-z^*(x)\|$ is a zero-mean sub-Gaussian random variable. Replacing one sample ξ_i in S with an i.i.d. sample ξ_i' and denote the new dataset as $S^{(i)}$, by Zhang et al. (2021a, Lemma 2), it holds that

$$\|z_S(x)-z^*(x)\|-\|z_{S^{(i)}}(x)-z^*(x)\| \le \|z_S(x)-z_{S^{(i)}}(x)\| \le \frac{2}{n}\sqrt{\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}},$$

where $z_{S^{(i)}}$ follows a similar definition of z_S but with a different dataset $S^{(i)}$. By McDiarmid's inequality (Lemma A.2) and the definition of sub-Gaussian random variables, it holds that $\|z_S(x)-z^*(x)\|-\mathbb{E}\|z_S(x)-z^*(x)\|$ is a zero-mean sub-Gaussian random variable with variance proxy $\frac{1}{n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)$. By the definition of sub-Gaussian random variable and (21), it holds that

$$\begin{aligned}&\mathbb{E}\exp\left(s\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\\ =\ &\mathbb{E}\exp\left(s\left[\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|-\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right]\right)\\ &\qquad\cdot\exp\left(s\,\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\\ \le\ &\mathbb{E}\exp\left(s\left[\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|-\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right]\right)\\ &\qquad\cdot\exp\left(s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}\right)\\ \le\ &\exp\left(\frac{s^2}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)\right)\exp\left(s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}\right),\end{aligned} \tag{22}$$

where the second inequality uses definition of zero-mean sub-Gaussian random variable. Combining (22) with (18), for

$$\lambda=\frac{1}{2L},\quad \rho=\frac{\epsilon\lambda(1-\lambda L)}{8}=\frac{\epsilon}{32L},\quad s=\sqrt{2n\log(Q)\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)^{-1}}, \tag{23}$$

it holds that

$$
\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S^\lambda(x)-\nabla\Phi^\lambda(x)\| \\
&\le 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda(L+\nu))}}+\frac{2\rho}{\lambda(1-\lambda(L+\nu))} \\
&\quad+\frac{1}{\lambda s}\log\left(Q\exp\left(\frac{s^2}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)\right)\right) \\
&\qquad\qquad+\frac{1}{\lambda s}\log\left(\exp\left(s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}\right)\right) \\
&\le 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda L)}}+\frac{1}{\lambda s}\log(Q)+\frac{1}{\lambda s}\frac{s^2}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right) \\
&\qquad\qquad+\frac{1}{\lambda s}s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\Big(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\Big)}+\frac{2\rho}{\lambda(1-\lambda L)} \\
&= 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda L)}}+\frac{\log(Q)}{\lambda s}+\frac{1}{\lambda}\frac{s}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right) \\
&\quad+\frac{1}{\lambda}\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}+\frac{\epsilon}{4} \\
&= 2\sqrt{4L\nu D_{\mathcal{Y}}}+4L\sqrt{\frac{\log(Q)}{2n}\left(\frac{\hat{L}_x^2}{L^2}+\frac{\hat{L}_y^2}{\nu L}\right)}+2L\sqrt{\frac{4\sqrt{2}}{Ln}\left(\frac{\hat{L}_x^2}{L}+\frac{\hat{L}_y^2}{\nu}\right)}+\frac{\epsilon}{4} \\
&= 2\sqrt{4L\nu D_{\mathcal{Y}}}+4L\sqrt{\frac{\log(Q)}{2n}\left(\frac{\hat{L}_x^2}{L^2}+\frac{\hat{L}_y^2}{\nu L}\right)} \\
&\qquad\qquad+2L\sqrt{\frac{4\sqrt{2}}{Ln}\left(\frac{(G+4L\sqrt{D_{\mathcal{X}}})^2}{L}+\frac{(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{\nu}\right)}+\frac{\epsilon}{4}.
\end{aligned}
\tag{24}
$$

Here the first equality holds by the selection of ρ, the second equality holds by the selection of λ and s, and the last equality holds by plugging in $\hat{L}_x$ and $\hat{L}_y$. Note that ρ, s, and ν are only used for analysis purposes, and λ is only used in the definition of gradient mapping. Thus one has free choices on these parameters. Since $Q=\mathcal{O}\left(\left(\frac{D_{\mathcal{X}}}{\rho}\right)^d\right)$, then we choose $\nu=\tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right)$ in the right-hand side above, which verifies the first statement. For the sample complexity result, to make sure that the right-hand side of (24) of order $\mathcal{O}(\epsilon)$, it suffices to have

$$
n=\mathcal{O}\left(\frac{\log(Q)}{\nu}\epsilon^{-2}\right)=\mathcal{O}\left(d\epsilon^{-4}\log\left(\epsilon^{-1}\right)\right), \tag{25}
$$

which concludes the proof. □

D Proof of Theorem 4.2

For simplicity we define the following notations:

$$
\begin{aligned}
&F(x,y)\triangleq\mathbb{E}_\xi\left[f(x,y;\xi)\right],\quad \Phi(x)\triangleq\max_y F(x,y),\quad F_S(x,y)\triangleq\frac{1}{n}\sum_{i=1}^n\left[f(x,y;\xi_i)\right],\quad \Phi_S(x)\triangleq\max_y F_S(x,y),\\
&y^*(x)\triangleq\operatorname*{argmax}_y F(x,y),\quad y_S^*(x)\triangleq\operatorname*{argmax}_y F_S(x,y),\quad \Phi(x;\xi)\triangleq\max_y f(x,y;\xi),
\end{aligned}
\tag{26}
$$

and the Moreau envelope of a function Φ:

$$
\Phi^\lambda(x)\triangleq\min_{z\in\mathcal{X}}\left\{\Phi(x)+\frac{1}{2\lambda}\|z-x\|_2^2\right\},\quad \mathbf{prox}_{\lambda\Phi}(x)\triangleq\operatorname*{argmin}_{z\in\mathcal{X}}\left\{\Phi(x)+\frac{1}{2\lambda}\|z-x\|_2^2\right\}, \tag{27}
$$

similar notations can be defined for Φ_S, which we do not repeat here.

Definition D.1 (Uniform Stability)**.** We say a randomized algorithm $\mathcal{A}$ is δ-uniformly stable in x-gradients if for every two dataset S, S' which differ in only one sample, for every $\xi \in \Xi$ we have

$$\sup_{\xi} \mathbb{E}_{\mathcal{A}} \|\nabla_x f(\mathcal{A}_x(S), \mathcal{A}_y(S); \xi) - \nabla_x f(\mathcal{A}_x(S'), \mathcal{A}_y(S'); \xi)\|^2 \leq \delta^2. \tag{28}$$

Lemma D.2 (Concentration of Optimizers)**.** For y^* and y_S^* defined above, with Assumption 2.1, we have for any $x \in \mathcal{X}$,

$$\|y^*(x) - y_S^*(x)\| \leq \frac{1}{\mu} \|\nabla_y F_S(x, y^*(x)) - \nabla_y F(x, y^*(x))\|. \tag{29}$$

Proof. By the optimality of $y^*(x)$ and $y_S^*(x)$, we have for any $y \in \mathcal{Y}$

$$\begin{aligned} \langle y - y^*(x), \nabla_y F(x, y^*(x)) \rangle &\leq 0 \\ \langle y - y_S^*(x), \nabla_y F_S(x, y_S^*(x)) \rangle &\leq 0. \end{aligned} \tag{30}$$

Setting $y = y_S^*(x)$ and $y = y^*(x)$ in the above inequalities respectively, we have

$$\langle y_S^*(x) - y^*(x), \nabla_y F(x, y^*(x)) - \nabla_y F_S(x, y_S^*(x)) \rangle \leq 0. \tag{31}$$

In addition, by strong concavity of $F_S(x, \cdot)$, we have

$$\langle y_S^*(x) - y^*(x), \nabla_y F_S(x, y_S^*(x)) - \nabla_y F_S(x, y^*(x)) \rangle + \mu \|y_S^*(x) - y^*(x)\|^2 \leq 0. \tag{32}$$

Combining (31) and (32), we have

$$\langle y_S^*(x) - y^*(x), \nabla_y F(x, y^*(x)) - \nabla_y F_S(x, y^*(x)) \rangle + \mu \|y_S^*(x) - y^*(x)\|^2 \leq 0. \tag{33}$$

Rearranging terms, it holds that

$$\begin{aligned} \mu \|y_S^*(x) - y^*(x)\|^2 &\leq \langle y_S^*(x) - y^*(x), \nabla_y F_S(x, y^*(x)) - \nabla_y F(x, y^*(x)) \rangle \\ &\leq \|y_S^*(x) - y^*(x)\| \cdot \|\nabla_y F_S(x, y^*(x)) - \nabla_y F(x, y^*(x))\|, \end{aligned} \tag{34}$$

which implies

$$\|y_S^*(x) - y^*(x)\| \leq \frac{1}{\mu} \|\nabla_y F_S(x, y^*(x)) - \nabla_y F(x, y^*(x))\|. \tag{35}$$

It concludes the proof. □

Lemma D.3 (Stability of Optimizers)**.** For y_S^* and $y_{S'}^*$ defined above where S and S' are two dataset differing in only one sample (ξ_i and ξ_i'), with Assumption 2.1 while $\mu > 0$, we have for any $x \in \mathcal{X}$,

$$\|y_S^*(x) - y_{S'}^*(x)\| \leq \frac{1}{\mu} \|\nabla_y F_S(x, y_{S'}^*(x)) - \nabla_y F_{S'}(x, y_{S'}^*(x))\| \leq \frac{2G}{n\mu}. \tag{36}$$

Proof. The proof is similar to that of Lemma D.2. By the optimality of $y_S^*(x)$ and $y_{S'}^*(x)$, we have for any $y \in \mathcal{Y}$

$$\begin{aligned} \langle y - y_S^*(x), \nabla_y F_S(x, y_S^*(x)) \rangle &\leq 0 \\ \langle y - y_{S'}^*(x), \nabla_y F_{S'}(x, y_{S'}^*(x)) \rangle &\leq 0. \end{aligned} \tag{37}$$

Setting $y = y_{S'}^*(x)$ and $y = y_S^*(x)$ in the above inequalities respectively, we have

$$\langle y_S^*(x) - y_{S'}^*(x), \nabla_y F_{S'}(x, y_{S'}^*(x)) - \nabla_y F_S(x, y_S^*(x)) \rangle \leq 0. \tag{38}$$

In addition, by strong concavity of $F_S(x, \cdot)$, we have

$$\langle y_S^*(x) - y_{S'}^*(x), \nabla_y F_S(x, y_S^*(x)) - \nabla_y F_S(x, y_{S'}^*(x)) \rangle + \mu \|y_S^*(x) - y_{S'}^*(x)\|^2 \leq 0. \tag{39}$$

Combining (38) and (39), we have

$$\langle y_S^*(x) - y_{S'}^*(x), \nabla_y F_S(x, y_{S'}^*(x)) - \nabla_y F_S(x, y_{S'}^*(x))\rangle + \mu\|y_S^*(x) - y_{S'}^*(x)\|^2 \le 0. \tag{40}$$

Rearranging terms, it holds that

$$\begin{aligned}\mu\|y_S^*(x) - y_{S'}^*(x)\|^2 &\le \langle y_S^*(x) - y_{S'}^*(x), \nabla_y F_S(x, y_{S'}^*(x)) - \nabla_y F_{S'}(x, y_{S'}^*(x))\rangle \\ &\le \|y_S^*(x) - y_{S'}^*(x)\| \cdot \|\nabla_y F_S(x, y_{S'}^*(x)) - \nabla_y F_{S'}(x, y_{S'}^*(x))\|,\end{aligned} \tag{41}$$

which implies

$$\begin{aligned}\|y_S^*(x) - y_{S'}^*(x)\| &\le \frac{1}{\mu}\|\nabla_y F_S(x, y_{S'}^*(x)) - \nabla_y F_{S'}(x, y_{S'}^*(x))\| \\ &= \frac{1}{\mu}\left\|\frac{1}{n}(\nabla_y f(x, y_{S'}^*(x); \xi_i) - \nabla_y f(x, y_{S'}^*(x); \xi_i'))\right\| \le \frac{2G}{n\mu},\end{aligned} \tag{42}$$

which concludes the proof. Here the equality above is due to the variables being the same $(x, y_{S'}^*(x))$, while S and S' differ in only one sample. □

Theorem D.4 (Stability and Generalization, NC-SC)**.** Let $\mathcal{A}$ be an δ-uniformly primal stable algorithm, for any function f satisfying Assumption 2.1 with $\mu > 0$, we have

$$\mathbb{E}_{\mathcal{A},S}\|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\| \le (1+\kappa)\left(4\delta + \frac{G}{\sqrt{n}}\right). \tag{43}$$

Proof. Following the definition, we have

$$\begin{aligned}&\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S)) \\ &= \nabla_x F(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y_S^*(\mathcal{A}_x(S))) \\ &= \nabla_x F(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) + \nabla_x F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y_S^*(\mathcal{A}_x(S))),\end{aligned} \tag{44}$$

so we know that

$$\begin{aligned}&\|\nabla\Phi(\mathcal{A}_x(S)) - \nabla\Phi_S(\mathcal{A}_x(S))\| \\ &\le \|\nabla_x F(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S)))\| \\ &\qquad + \|\nabla_x F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y_S^*(\mathcal{A}_x(S)))\|,\end{aligned} \tag{45}$$

for the first term above, by Lei (2022, Theorem 2) (i.e., regarding $(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S)))$ as one single variable to recover their conclusion), we have

$$\mathbb{E}_{\mathcal{A},S}\|\nabla_x F(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S)))\| \le 4\delta + \sqrt{\frac{\mathrm{Var}(\nabla_x f)}{n}} \le 4\delta + \frac{G}{\sqrt{n}}, \tag{46}$$

for the second term above, by Lemma D.2, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{A},S}\,\|\nabla_x F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_x F_S(\mathcal{A}_x(S), y_S^*(\mathcal{A}_x(S)))\| \\ &\le L\,\mathbb{E}_{\mathcal{A},S}\,\|y^*(\mathcal{A}_x(S)) - y_S^*(\mathcal{A}_x(S))\| \\ &\le \kappa\,\mathbb{E}_{\mathcal{A},S}\,\|\nabla_y F_S(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S))) - \nabla_y F(\mathcal{A}_x(S), y^*(\mathcal{A}_x(S)))\| \\ &\le \kappa\left(4\delta + \sqrt{\frac{\mathrm{Var}(\nabla_y f)}{n}}\right) \\ &\le \kappa\left(4\delta + \frac{G}{\sqrt{n}}\right),\end{aligned} \tag{47}$$

where the third inequality applies the same argument as that in (46). We conclude the proof by combining the two bounds above together. □

E Proof of Theorem 4.4

The proof uses the idea from Lei (2022, Theorem 3) and our proof of Theorem 3.2. Unlike Lei (2022) which considers the minimization case, with $\Phi(x) \neq \mathbb{E}[\Phi(x;\xi)]$, we need some modification in the proof. To address the non-uniqueness of $y^*(x)$ in the NC-C case, similar to the uniform convergence analysis in the NC-C case (Theorem 3.2), we resort to the regularized objective in the proof to characterize corresponding distances.

For convenience, we recall the definition of regularized objective functions here.

$$\begin{aligned} \widehat{\Phi}(x) &= \max_{y\in\mathcal{Y}} F(x,y) - \frac{\nu}{2}\|y\|^2, \quad \widehat{\Phi}_S(x) = \max_{y\in\mathcal{Y}} F_S(x,y) - \frac{\nu}{2}\|y\|^2, \\ \widehat{y}^*(x) &= \operatorname*{argmax}_{y\in\mathcal{Y}} F(x,y) - \frac{\nu}{2}\|y\|^2, \quad \widehat{y}^*_S(x) = \operatorname*{argmax}_{y\in\mathcal{Y}} F_S(x,y) - \frac{\nu}{2}\|y\|^2. \end{aligned} \tag{48}$$

In addition, following the notation in Lei (2022), we define

$$\begin{aligned} \widetilde{w}_S &= \mathbf{prox}_{\frac{\widehat{\Phi}}{2L}}(\mathcal{A}_x(S)) = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\widehat{\Phi}(x) + L\|x - \mathcal{A}_x(S)\|^2\right\}, \\ w_S &= \mathbf{prox}_{\frac{\widehat{\Phi}_S}{2L}}(\mathcal{A}_x(S)) = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\widehat{\Phi}_S(x) + L\|x - \mathcal{A}_x(S)\|^2\right\}. \end{aligned} \tag{49}$$

As discussed in Appendix C, the function $F(x,y) - \frac{\nu}{2}\|y\|^2$ is $(L+\nu)$-smooth, and the function $\widehat{\Phi}(x)$ is $(L+\nu)$-weakly-convex (the same hold for $F_S(x,y) - \frac{\nu}{2}\|y\|^2$ and $\widehat{\Phi}_S(x)$).

First, we build up a connection between algorithm stability and proximal operators to facilitate the analysis.

Lemma E.1 (Algorithm Stability and Proximal Operators)**.** Let $\mathcal{A}$ be an algorithm. For any function f satisfying Assumption 2.1 with $\mu = 0$, we have the following inequalities for any two neighboring dataset S and S', we have

$$\begin{aligned} \|\widetilde{w}_S - \widetilde{w}_{S'}\| &\leq \frac{2L}{L-\nu}\|\mathcal{A}(S) - \mathcal{A}(S')\| \\ \|w_S - w_{S'}\| &\leq \frac{2L}{L-\nu}\|\mathcal{A}(S) - \mathcal{A}(S')\| + \frac{2G}{n(L-\nu)} + \frac{2L(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu(L-\nu)}, \end{aligned} \tag{50}$$

where $\widehat{\Phi}$ and $\widehat{\Phi}_S$ follows the definitions in (48) and (49).

The proof basically follows the proof of Lei (2022, Lemma 15 and 16) with some differences in detailed parameters.

Proof. For the first result, note that $\widehat{\Phi}(x)$ is $(L+\nu)$-weakly-convex and differentiable, so we have

$$\left\langle \widetilde{w}_S - \widetilde{w}_{S'}, \nabla\widehat{\Phi}(\widetilde{w}_S) - \nabla\widehat{\Phi}(\widetilde{w}_{S'}) \right\rangle \geq -(L+\nu)\|\widetilde{w}_S - \widetilde{w}_{S'}\|^2. \tag{51}$$

On the other hand, by the optimality of $\widetilde{w}_S$, we have

$$-2L(\widetilde{w}_S - \mathcal{A}_x(S)) - \nabla\widehat{\Phi}\widetilde{w}_S) \in \partial\mathcal{I}_{\mathcal{X}}(\widetilde{w}_S), \quad -2L(\widetilde{w}_{S'} - \mathcal{A}_x(S')) - \nabla\widehat{\Phi}(\widetilde{w}_{S'}) \in \partial\mathcal{I}_{\mathcal{X}}(\widetilde{w}_{S'}), \tag{52}$$

where $\mathcal{I}_{\mathcal{X}}(x)$ is the indicator function of the set $\mathcal{X}$, i.e., $\mathcal{I}_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $\mathcal{I}_{\mathcal{X}}(x) = \infty$ otherwise. Since $\mathcal{X}$ is convex, the subgradient $\partial\mathcal{I}_{\mathcal{X}}$ is monotone, and thus

$$\begin{aligned} \left\langle \widetilde{w}_S - \widetilde{w}_{S'}, 2L(\widetilde{w}_{S'} - \mathcal{A}_x(S')) - 2L(\widetilde{w}_S - \mathcal{A}_x(S)) + \nabla\widehat{\Phi}(\widetilde{w}_{S'}) - \nabla\widehat{\Phi}(\widetilde{w}_S) \right\rangle &= \langle \widetilde{w}_S - \widetilde{w}_{S'}, \partial\mathcal{I}_{\mathcal{X}}(\widetilde{w}_S) - \partial\mathcal{I}_{\mathcal{X}}(\widetilde{w}_{S'})\rangle \\ &\geq 0. \end{aligned} \tag{53}$$

Combining (51) and (53), it follows that

$$\langle \widetilde{w}_S - \widetilde{w}_{S'}, 2L(\widetilde{w}_{S'} - \mathcal{A}_x(S')) - 2L(\widetilde{w}_S - \mathcal{A}_x(S))\rangle \geq -(L+\nu)\|\widetilde{w}_S - \widetilde{w}_{S'}\|^2. \tag{54}$$

Rearranging the terms, we have

$$(L-\nu)\|\widetilde{w}_S-\widetilde{w}_{S'}\|^2\le 2L\langle \widetilde{w}_S-\widetilde{w}_{S'},\mathcal{A}_x(S)-\mathcal{A}_x(S')\rangle\le 2L\|\widetilde{w}_S-\widetilde{w}_{S'}\|\|\mathcal{A}_x(S)-\mathcal{A}_x(S')\|. \tag{55}$$

We obtain the first result by dividing both sides by $(L-\nu)\|\widetilde{w}_S-\widetilde{w}_{S'}\|$.

For the second statement, applying the fact that $\widehat{\Phi}_S$ is weakly-convex and differentiable,

$$\left\langle w_S-w_{S'},\nabla\widehat{\Phi}_S(w_S)-\nabla\widehat{\Phi}_S(w_{S'})\right\rangle\ge-(L+\nu)\|w_S-w_{S'}\|^2. \tag{56}$$

Similar as (53), by the optimality condition of w_S and $w_{S'}$,

$$\left\langle w_S-w_{S'},2L(w_{S'}-\mathcal{A}_x(S'))-2L(w_S-\mathcal{A}_x(S))+\nabla\widehat{\Phi}_{S'}(w_{S'})-\nabla\widehat{\Phi}_S(w_S)\right\rangle\ge 0. \tag{57}$$

Therefore, by the above two equations, we obtain that

$$-(L+\nu)\|w_S-w_{S'}\|^2\le\left\langle w_S-w_{S'},2L(w_{S'}-\mathcal{A}_x(S'))-2L(w_S-\mathcal{A}_x(S))+\nabla\widehat{\Phi}_{S'}(w_{S'})-\nabla\widehat{\Phi}_S(w_{S'})\right\rangle. \tag{58}$$

By the definition of $\widehat{\Phi}_S$ and w_S, we rewrite the additional term $\nabla\widehat{\Phi}_{S'}(w_{S'})-\nabla\widehat{\Phi}_S(w_{S'})$ as

$$\begin{aligned}
&\nabla\widehat{\Phi}_{S'}(w_{S'})-\nabla\widehat{\Phi}_S(w_{S'})\\
=&\nabla_x\Big(F_{S'}(w_{S'};\widehat{y}^*_{S'}(w_{S'}))-\frac{\nu}{2}\|\widehat{y}^*_{S'}(w_{S'})\|^2\Big)-\nabla\widehat{\Phi}_S(w_{S'})\\
=&\nabla_x F_{S'}(w_{S'};\widehat{y}^*_{S'}(w_{S'}))-\nabla\widehat{\Phi}_S(w_{S'})\\
=&\nabla_x F_{S'}(w_{S'};\widehat{y}^*_{S'}(w_{S'}))-\nabla_x F_S(w_{S'};\widehat{y}^*_{S'}(w_{S'}))+\nabla_x F_S(w_{S'};\widehat{y}^*_{S'}(w_{S'}))-\nabla_x F_S(w_{S'};\widehat{y}^*_S(w_{S'}))\\
=&\underbrace{\frac{1}{n}\nabla_x f(w_{S'},\widehat{y}^*_{S'}(w_{S'});\xi'_i)-\frac{1}{n}\nabla_x f(w_{S'},\widehat{y}^*_{S'}(w_{S'});\xi_i)}_{E_1}+\underbrace{\nabla_x F_S(w_{S'};\widehat{y}^*_{S'}(w_{S'}))-\nabla_x F_S(w_{S'};\widehat{y}^*_S(w_{S'}))}_{E_2},
\end{aligned} \tag{59}$$

where the third equation holds since $\nabla\widehat{\Phi}_S(w_{S'})=\nabla_x F_S(w_{S'};\widehat{y}^*_S(w_{S'}))$. Thus it holds that

$$-(L+\nu)\|w_S-w_{S'}\|^2\le\langle w_S-w_{S'},-2L(w_S-\mathcal{A}_x(S))+2L(w_{S'}-\mathcal{A}_x(S'))+E_1+E_2\rangle. \tag{60}$$

Rearranging terms, we have

$$\begin{aligned}
&(L-\nu)\|w_S-w_{S'}\|^2\\
\le&\langle w_S-w_{S'},2L(\mathcal{A}_x(S)-\mathcal{A}_x(S'))+E_1+E_2\rangle\\
\le&\|w_S-w_{S'}\|\|2L(\mathcal{A}_x(S)-\mathcal{A}_x(S'))+E_1+E_2\|\\
\le&\|w_S-w_{S'}\|(2L\|(\mathcal{A}_x(S)-\mathcal{A}_x(S'))\|+\|E_1\|+\|E_2\|)\\
\le&\|w_S-w_{S'}\|\left(2L\|(\mathcal{A}_x(S)-\mathcal{A}_x(S'))\|+\frac{2G}{n}+\frac{2L(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}\right),
\end{aligned} \tag{61}$$

where the last inequality uses the fact that that $\|E_1\|\le 2G/n$ via Lipschitz continuity, and

$$\|E_2\|\le L\|\widehat{y}^*_{S'}(w_{S'})-\widehat{y}^*_S(w_{S'})\|\overset{\text{Lemma D.3}}{\le}\frac{2L(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}.$$

It concludes the proof by diving $(L-\nu)\|w_S-w_{S'}\|$ on both sides of (61). □

Lemma E.2. Let $\mathcal{A}$ be an δ-uniformly primal argument stable algorithm. For any function f satisfying Assumption 2.1 with $\mu=0$, we have

$$\mathbb{E}\left[\widehat{\Phi}_S(\widetilde{w}_S)-\widehat{\Phi}(\widetilde{w}_S)\right]\le\frac{2GL(L+2\nu)}{\nu(L-\nu)}\delta+\frac{G}{\nu}\left(4\sqrt{\frac{8L^4(L+2\nu)^2}{\nu^2(L-\nu)^2}\delta}+\frac{G}{\sqrt{n}}\right)+\frac{\nu}{2}D_{\mathcal{Y}}. \tag{62}$$

Proof. Note that

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\Phi}_S(\widetilde{w}_S) - \widehat{\Phi}(\widetilde{w}_S)\right] \\
=\ &\mathbb{E}\left[F_S(\widetilde{w}_S, \widehat{y}_S^*(\widetilde{w}_S)) - F(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) - \frac{\nu}{2}\|\widehat{y}_S^*(\widetilde{w}_S)\|^2 + \frac{\nu}{2}\|\widehat{y}^*(\widetilde{w}_S)\|^2\right] \\
\le\ &\mathbb{E}\left[\underbrace{F_S(\widetilde{w}_S, \widehat{y}_S^*(\widetilde{w}_S)) - F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))}_{H_1} + \underbrace{F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) - F(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))}_{H_2}\right] + \frac{\nu}{2}D_{\mathcal{Y}}.
\end{aligned} \tag{63}$$

We bound H_2 via the stability argument of $f(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S); \xi)$, i.e., regarding $(\widetilde{w}_S, \widehat{y}_S^*(\widetilde{w}_S))$ as one single variable.

$$\begin{aligned}
&\mathbb{E}\left[f(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S); \xi)\right] - \mathbb{E}\left[f(\widetilde{w}_{S'}, \widehat{y}^*(\widetilde{w}_{S'}); \xi)\right] \\
\le\ &G\,\mathbb{E}\left[\|\widetilde{w}_S - \widetilde{w}_{S'}\| + \|\widehat{y}^*(\widetilde{w}_S) - \widehat{y}^*(\widetilde{w}_{S'})\|\right] \\
\le\ &G\,\mathbb{E}\left[\|\widetilde{w}_S - \widetilde{w}_{S'}\| + \frac{L+\nu}{\nu}\|\widetilde{w}_S - \widetilde{w}_{S'}\|\right] \\
\le\ &G\,\mathbb{E}\left[\left(1 + \frac{L+\nu}{\nu}\right)\cdot\frac{2L}{L-\nu}\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\|\right] \\
\le\ &\frac{L+2\nu}{\nu}\cdot\frac{2GL}{L-\nu}\,\mathbb{E}\left[\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\|\right] \\
\le\ &\frac{2GL(L+2\nu)}{\nu(L-\nu)}\delta,
\end{aligned} \tag{64}$$

where the second inequality uses Lin et al. (2020a, Lemma 4.3), and the fact that $\hat{y}^*$ is the optimal solution of a $(L+\nu)$-smooth and ν-strongly concave maximization problem defined in (48); the third inequality is due to Lemma E.1, and the last inequality follows the definition of δ-uniform primal argument stability. So we have the "composed algorithm" $\widetilde{w}_S$ is stable[3] in function values, which implies (Hardt et al., 2016)

$$\mathbb{E}\left[F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) - F(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))\right] \le \frac{2GL(L+2\nu)}{\nu(L-\nu)}\delta. \tag{65}$$

For the term H_1 above, we have

$$\begin{aligned}
&\mathbb{E}\left[F_S(\widetilde{w}_S, \widehat{y}_S^*(\widetilde{w}_S)) - F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))\right] \\
\le\ &G\,\mathbb{E}\,\|\widehat{y}_S^*(\widetilde{w}_S) - \widehat{y}^*(\widetilde{w}_S)\| \\
\le\ &\frac{G}{\nu}\,\mathbb{E}\,\|\nabla_y F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) - \nu\widehat{y}^*(\widetilde{w}_S) - \nabla_y F(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) + \nu\widehat{y}^*(\widetilde{w}_S)\| \\
=\ &\frac{G}{\nu}\,\mathbb{E}\,\|\nabla_y F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) - \nabla_y F(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))\|,
\end{aligned} \tag{66}$$

where the second inequality applies Lemma D.2. We further upper bound the RHS above using the stability argument. For $\nabla_y f(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S); \xi)$, similar to the same argument as in (64), we have

$$\begin{aligned}
&\mathbb{E}\,\|\nabla_y f(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S); \xi) - \nabla_y f(\widetilde{w}_{S'}, \widehat{y}^*(\widetilde{w}_{S'}); \xi)\|^2 \\
\le\ &2L^2\,\mathbb{E}\left[\|\widetilde{w}_S - \widetilde{w}_{S'}\|^2 + \|\widehat{y}^*(\widetilde{w}_S) - \widehat{y}^*(\widetilde{w}_{S'})\|^2\right] \\
\le\ &2L^2\,\mathbb{E}\left[\left(1 + \left(\frac{L+\nu}{\nu}\right)^2\right)\|\widetilde{w}_S - \widetilde{w}_{S'}\|^2\right] \\
\le\ &2L^2\left(1 + \left(\frac{L+\nu}{\nu}\right)^2\right)\cdot\left(\frac{2L}{L-\nu}\right)^2\mathbb{E}\left[\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\|^2\right] \\
\le\ &\frac{8L^4(L+2\nu)^2}{\nu^2(L-\nu)^2}\delta^2,
\end{aligned} \tag{67}$$

[3]Here we call the iteration $\widetilde{w}_S = \mathbf{prox}_{\frac{\widehat{\Phi}}{2L}}(\mathcal{A}_x(S)) = \operatorname{argmin}_{x\in\mathcal{X}}\left\{\widehat{\Phi}(x) + L\|x - \mathcal{A}_x(S)\|^2\right\}$ as an algorithm regarding that it is a composition of the algorithm $\mathcal{A}$ and the proximal operator.

where the second inequality comes from Lin et al. (2020a, Lemma 4.3). It concludes that algorithm $\mathcal{A}$ is δ-uniformly primal stable. Applying Lei (2022, Theorem 2) to (66), we have

$$\begin{aligned}\mathbb{E}\left[F_S(\widetilde{w}_S, \widehat{y}_S^*(\widetilde{w}_S)) - F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))\right] &\le \frac{G}{\nu}\,\mathbb{E}\,\|\nabla_y F_S(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S)) - \nabla_y F(\widetilde{w}_S, \widehat{y}^*(\widetilde{w}_S))\| \\ &\le \frac{G}{\nu}\left(4\sqrt{\frac{8L^4(L+2\nu)^2}{\nu^2(L-\nu)^2}}\delta + \sqrt{\frac{\mathrm{Var}(\nabla_y f)}{n}}\right) \\ &\le \frac{G}{\nu}\left(4\sqrt{\frac{8L^4(L+2\nu)^2}{\nu^2(L-\nu)^2}}\delta + \frac{G}{\sqrt{n}}\right),\end{aligned} \tag{68}$$

which concludes the proof. □

Lemma E.3. Let $\mathcal{A}$ be an δ-uniformly primal argument stable algorithm. For any function f satisfying Assumption 2.1 with $\mu = 0$, we have

$$\begin{aligned}\mathbb{E}\left[\widehat{\Phi}(w_S) - \widehat{\Phi}_S(w_S)\right] \le \frac{G}{\nu}\left(4\sqrt{\frac{8L^2(L+2\nu)^2}{\nu^2}\left(\frac{4L^2}{(L-\nu)^2}\delta^2 + \frac{4G^2}{n^2(L-\nu)^2} + \frac{2L^2(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2(L-\nu)^2}\right)} + \frac{G}{\sqrt{n}}\right) \\ + \frac{G(L+2\nu)}{\nu}\left(\frac{2L}{L-\nu}\delta + \frac{2G}{n(L-\nu)} + \frac{2L(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu(L-\nu)}\right) + \frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu} + \frac{\nu}{2}D_{\mathcal{Y}}.\end{aligned} \tag{69}$$

Proof. Note that

$$\begin{aligned}&\mathbb{E}\left[\widehat{\Phi}(w_S) - \widehat{\Phi}_S(w_S)\right] \\ =\ &\mathbb{E}\left[F(w_S, \widehat{y}^*(w_S)) - F_S(w_S, \widehat{y}_S^*(w_S)) - \frac{\nu}{2}\|\widehat{y}^*(w_S)\|^2 + \frac{\nu}{2}\|\widehat{y}_S^*(w_S)\|^2\right] \\ \le\ &\mathbb{E}\left[\underbrace{F(w_S, \widehat{y}^*(w_S)) - F(w_S, \widehat{y}_S^*(w_S))}_{J_1} + \underbrace{F(w_S, \widehat{y}_S^*(w_S)) - F_S(w_S, \widehat{y}_S^*(w_S))}_{J_2}\right] + \frac{\nu}{2}D_{\mathcal{Y}}.\end{aligned} \tag{70}$$

For J_2, by Lemma E.1, similar to the analysis of H_2 in the proof of Lemma E.2, we have

$$\begin{aligned}&\mathbb{E}\left[f(w_S, \widehat{y}_S^*(w_S); \xi) - \mathbb{E}\left[f(w_{S'}, \widehat{y}_{S'}^*(w_{S'}); \xi)\right]\right] \\ \le\ &G\,\mathbb{E}\left[\|w_S - w_{S'}\| + \|\widehat{y}_S^*(w_S) - \widehat{y}_{S'}^*(w_S)\| + \|\widehat{y}_{S'}^*(w_S) - \widehat{y}_{S'}^*(w_{S'})\|\right] \\ \le\ &G\,\mathbb{E}\left[\|w_S - w_{S'}\| + \frac{L+\nu}{\nu}\|w_S - w_{S'}\|\right] + \frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu} \\ \le\ &G\,\mathbb{E}\left[\left(1 + \frac{L+\nu}{\nu}\right)\cdot\left(\frac{2L}{L-\nu}\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\| + \frac{2G}{n(L-\nu)} + \frac{2L(G+\nu D_{\mathcal{Y}})}{n\nu(L-\nu)}\right)\right] + \frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu} \\ \le\ &\frac{G(L+2\nu)}{\nu}\cdot\left(\frac{2L}{L-\nu}\,\mathbb{E}\left[\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\|\right] + \frac{2G}{n(L-\nu)} + \frac{2L(G+\nu D_{\mathcal{Y}})}{n\nu(L-\nu)}\right) + \frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu} \\ \le\ &\frac{G(L+2\nu)}{\nu}\left(\frac{2L}{L-\nu}\delta + \frac{2G}{n(L-\nu)} + \frac{2L(G+\nu D_{\mathcal{Y}})}{n\nu(L-\nu)}\right) + \frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}.\end{aligned} \tag{71}$$

It further holds that

$$\mathbb{E}\left[F(w_S, \widehat{y}_S^*(w_S)) - F_S(w_S, \widehat{y}_S^*(w_S))\right] \le \frac{G(L+2\nu)}{\nu}\left(\frac{2L}{L-\nu}\delta + \frac{2G}{n(L-\nu)} + \frac{2L(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu(L-\nu)}\right) + \frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}. \tag{72}$$

For J_1, similar to the analysis of H_1 in the proof of Lemma E.2, we have

$$\begin{aligned}
&\mathbb{E}\left\|\nabla_y f(w_S, \widehat{y}^*(w_S); \xi) - \nabla_y f(w_{S'}, \widehat{y}^*(w_{S'}); \xi)\right\|^2 \\
&\leq 2L^2\,\mathbb{E}\left[\left\|w_S - w_{S'}\right\|^2 + \left\|\widehat{y}^*(w_S) - \widehat{y}^*(w_{S'})\right\|^2\right] \\
&\leq 2L^2\,\mathbb{E}\left[\left(1 + \left(\frac{L+\nu}{\nu}\right)^2\right)\left\|w_S - w_{S'}\right\|^2\right] \\
&\leq 2L^2\left(1 + \left(\frac{L+\nu}{\nu}\right)^2\right)\cdot\mathbb{E}\left[4\left(\frac{2L}{L-\nu}\right)^2\left\|\mathcal{A}_x(S) - \mathcal{A}_x(S')\right\|^2 + 4\frac{4G^2}{n^2(L-\nu)^2} + 2\frac{4L^2(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2(L-\nu)^2}\right] \\
&\leq \frac{8L^2(L+2\nu)^2}{\nu^2}\left(\frac{4L^2}{(L-\nu)^2}\delta^2 + \frac{4G^2}{n^2(L-\nu)^2} + \frac{2L^2(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2(L-\nu)^2}\right).
\end{aligned} \tag{73}$$

Combined with Lei (2022, Theorem 2), we have

$$\begin{aligned}
&\mathbb{E}\left[F(w_S, \widehat{y}_S^*(w_S)) - F(w_S, \widehat{y}^*(w_S))\right] \\
&\leq \frac{G}{\nu}\,\mathbb{E}\left\|\nabla_y F_S(w_S, \widehat{y}^*(w_S)) - \nabla_y F(w_S, \widehat{y}^*(w_S))\right\| \\
&\leq \frac{G}{\nu}\left(4\sqrt{\frac{8L^2(L+2\nu)^2}{\nu^2}\left(\frac{4L^2}{(L-\nu)^2}\delta^2 + \frac{4G^2}{n^2(L-\nu)^2} + \frac{2L^2(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2(L-\nu)^2}\right)} + \sqrt{\frac{\mathrm{Var}(\nabla_y f)}{n}}\right) \\
&\leq \frac{G}{\nu}\left(4\sqrt{\frac{8L^2(L+2\nu)^2}{\nu^2}\left(\frac{4L^2}{(L-\nu)^2}\delta^2 + \frac{4G^2}{n^2(L-\nu)^2} + \frac{2L^2(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2(L-\nu)^2}\right)} + \frac{G}{\sqrt{n}}\right),
\end{aligned} \tag{74}$$

which concludes the proof. □

Next, we formally demonstrate the proof for the generalization bounds in the NC-C setting.

Theorem E.4 (Stability and Generalization, NC-C, repeat Theorem 4.4)**.** Let $\mathcal{A}$ be an δ-uniformly primal argument stable algorithm, for any function f satisfying Assumption 2.1 with $\mu = 0$, we have

$$\mathbb{E}_{\mathcal{A},S}\left\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\right\| \leq \mathcal{O}\left(\delta^{\frac{1}{6}} + \left(\frac{1}{n}\right)^{\frac{1}{12}}\right). \tag{75}$$

Proof. Recall that

$$\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) = 2L\Big(\mathcal{A}_x(S) - \mathbf{prox}_{\frac{\Phi}{2L}}(\mathcal{A}(S))\Big), \quad \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S)) = 2L\Big(\mathcal{A}_x(S) - \mathbf{prox}_{\frac{\Phi_S}{2L}}(\mathcal{A}(S))\Big). \tag{76}$$

Since Φ is L-weakly-convex and G-Lipschitz (Lin et al., 2020a, Lemma 4.7), it holds that

$$\left\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\right\| = 2L\left\|\mathbf{prox}_{\frac{\Phi}{2L}}(\mathcal{A}(S)) - \mathbf{prox}_{\frac{\Phi_S}{2L}}(\mathcal{A}(S))\right\|. \tag{77}$$

Utilizing the regularized objective function, we have

$$\begin{aligned}
&\left\|\mathbf{prox}_{\frac{\Phi}{2L}}(\mathcal{A}(S)) - \mathbf{prox}_{\frac{\Phi_S}{2L}}(\mathcal{A}(S))\right\| \\
&\leq \left\|\mathbf{prox}_{\frac{\Phi}{2L}}(\mathcal{A}(S)) - \mathbf{prox}_{\frac{\widetilde{\Phi}}{2L}}(\mathcal{A}(S))\right\| + \left\|\mathbf{prox}_{\frac{\widetilde{\Phi}}{2L}}(\mathcal{A}(S)) - \mathbf{prox}_{\frac{\widetilde{\Phi}_S}{2L}}(\mathcal{A}(S))\right\| + \left\|\mathbf{prox}_{\frac{\widetilde{\Phi}_S}{2L}}(\mathcal{A}(S)) - \mathbf{prox}_{\frac{\Phi_S}{2L}}(\mathcal{A}(S))\right\| \\
&\leq 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{L-\nu}} + \|\widetilde{w}_S - w_S\|,
\end{aligned} \tag{78}$$

where the second inequality comes from Lemma 3.3 with $\lambda = \frac{1}{2L}$. So now the problem is transformed to characterizing the distance between $\widetilde{w}_S$ and w_S coming from the regularized surrogate objective which is NC-SC.

Since the function $\widehat{\Phi}(x)+L\|x-\mathcal{A}(S)\|^2$ is $(L-\nu)$-strongly convex, and by the definition of $\widetilde{w}_S$, we have

$$\begin{aligned}
&\frac{L-\nu}{2}\,\mathbb{E}\,\|w_S-\widetilde{w}_S\|^2\\
&\le \mathbb{E}\,\widehat{\Phi}(w_S)+L\|w_S-\mathcal{A}(S)\|^2-\Big(\widehat{\Phi}(\widetilde{w}_S)+L\|\widetilde{w}_S-\mathcal{A}(S)\|^2\Big)\\
&= \mathbb{E}\,\widehat{\Phi}_S(w_S)+L\|w_S-\mathcal{A}(S)\|^2-\Big(\widehat{\Phi}_S(\widetilde{w}_S)+L\|\widetilde{w}_S-\mathcal{A}(S)\|^2\Big)+\Big(\widehat{\Phi}(w_S)-\widehat{\Phi}_S(w_S)\Big)+\Big(\widehat{\Phi}_S(\widetilde{w}_S)-\widehat{\Phi}(\widetilde{w}_S)\Big)\\
&\le \mathbb{E}\,\Big(\widehat{\Phi}(w_S)-\widehat{\Phi}_S(w_S)\Big)+\Big(\widehat{\Phi}_S(\widetilde{w}_S)-\widehat{\Phi}(\widetilde{w}_S)\Big)\\
&\le \frac{G}{\nu}\Bigg(4\sqrt{\frac{8L^2(L+2\nu)^2}{\nu^2}\Big(\frac{4L^2}{(L-\nu)^2}\delta^2+\frac{4G^2}{n^2(L-\nu)^2}+\frac{2L^2(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2(L-\nu)^2}\Big)}+\frac{G}{\sqrt{n}}\Bigg)\\
&\qquad+\frac{G(L+2\nu)}{\nu}\Big(\frac{2L}{L-\nu}\delta+\frac{2G}{n(L-\nu)}+\frac{2L(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu(L-\nu)}\Big)+\frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}\\
&\qquad+\frac{2GL(L+2\nu)}{\nu(L-\nu)}\delta+\frac{G}{\nu}\Bigg(4\sqrt{\frac{8L^4(L+2\nu)^2}{\nu^2(L-\nu)^2}}\delta+\frac{G}{\sqrt{n}}\Bigg)+\nu D_{\mathcal{Y}},
\end{aligned}\tag{79}$$

where the second inequality uses the optimality of w_S and $\widehat{w}_S$, the last inequality is due to Lemma E.2 and E.3. Now we choose ν to simplify the RHS above. For simplicity, first we set $\nu\le\frac{L}{2}$, so $L-\nu\ge\frac{L}{2}$, $L+2\nu\le 2L$. The RHS above simplifies to

$$\begin{aligned}
&\frac{L-\nu}{2}\,\mathbb{E}\,\|w_S-\widetilde{w}_S\|^2\\
&\le \frac{G}{\nu}\Bigg(4\sqrt{\frac{32L^4}{\nu^2}\Big(16\delta^2+\frac{16G^2}{n^2L^2}+\frac{16(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2}\Big)}+\frac{G}{\sqrt{n}}\Bigg)+\frac{2GL}{\nu}\Big(4\delta+\frac{4G}{nL}+\frac{4(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}\Big)\\
&\qquad+\frac{2G(G+\nu\sqrt{D_{\mathcal{Y}}})}{n\nu}+\frac{8GL}{\nu}\delta+\frac{G}{\nu}\Bigg(4\sqrt{\frac{128L^4}{\nu^2}}\delta+\frac{G}{\sqrt{n}}\Bigg)+\nu D_{\mathcal{Y}}\\
&\le \frac{G}{\nu}\Bigg(\frac{128L^2}{\nu}\sqrt{\delta^2+\frac{G^2}{n^2L^2}+\frac{(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{n^2\nu^2}}+\frac{G}{\sqrt{n}}\Bigg)+\frac{8GL}{\nu}\Big(2\delta+\frac{G}{nL}+\frac{G+\nu\sqrt{D_{\mathcal{Y}}}}{n\nu}\Big)\\
&\qquad+\frac{G}{\nu}\Big(\frac{64L^2}{\nu}\delta+\frac{2(G+\nu\sqrt{D_{\mathcal{Y}}})}{n}+\frac{G}{\sqrt{n}}\Big)+\nu D_{\mathcal{Y}}\\
&\le \frac{G}{\nu}\Bigg(\frac{128L^2}{\nu}\Big(\delta+\frac{G}{nL}+\frac{G+\nu\sqrt{D_{\mathcal{Y}}}}{n\nu}\Big)+\frac{G}{\sqrt{n}}\Bigg)+\frac{8GL}{\nu}\Big(2\delta+\frac{G}{nL}+\frac{G+\nu\sqrt{D_{\mathcal{Y}}}}{n\nu}\Big)\\
&\qquad+\frac{G}{\nu}\Big(\frac{64L^2}{\nu}\delta+\frac{2(G+\nu\sqrt{D_{\mathcal{Y}}})}{n}+\frac{G}{\sqrt{n}}\Big)+\nu D_{\mathcal{Y}}\\
&= 64G\Big(3\delta+\frac{2G}{nL}+\frac{2G+2\nu\sqrt{D_{\mathcal{Y}}}}{n\nu}\Big)\frac{L^2}{\nu^2}+2G\Big(8\delta+\frac{G}{\sqrt{n}L}+\frac{4G}{nL}+\frac{G+\nu\sqrt{D_{\mathcal{Y}}}}{n}\Big(\frac{4}{\nu}+\frac{1}{L}\Big)\Big)\frac{L}{\nu}+\nu D_{\mathcal{Y}}\\
&= \mathcal{O}\Big(\frac{1}{n}\Big)\cdot\mathcal{O}\Big(\frac{1}{\nu^3}+\frac{1}{\nu^2}+\frac{1}{\nu}+1\Big)+\mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big)\cdot\mathcal{O}\Big(\frac{1}{\nu}\Big)+\mathcal{O}(\delta)\cdot\mathcal{O}\Big(\frac{1}{\nu^2}+\frac{1}{\nu}\Big)+\mathcal{O}(1)\nu,
\end{aligned}\tag{80}$$

where the last step hides all other dependence on parameters except δ and n. Let

$$\frac{1}{\nu}=\mathcal{O}\Big(\min\Big(\delta^{-\frac{1}{3}},n^{\frac{1}{4}}\Big)\Big),\tag{81}$$

with δ and $1/n$ small enough such that $\nu\le L/2$ holds. The setting of ν implies that

$$\begin{aligned}
\mathbb{E}\,\|w_S-\widetilde{w}_S\|^2&\le\mathcal{O}\Big(\frac{1}{n}\Big)\cdot\mathcal{O}\Big(\frac{1}{\nu^3}\Big)+\mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big)\cdot\mathcal{O}\Big(\frac{1}{\nu}\Big)+\mathcal{O}(\delta)\cdot\mathcal{O}\Big(\frac{1}{\nu^2}\Big)+\mathcal{O}(1)\nu\\
&\le\mathcal{O}\Big(\frac{1}{n}\Big)\cdot\mathcal{O}\Big(n^{\frac{3}{4}}\Big)+\mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big)\cdot\mathcal{O}\Big(n^{\frac{1}{4}}\Big)+\mathcal{O}(\delta)\cdot\mathcal{O}\Big(\delta^{-\frac{2}{3}}\Big)+\mathcal{O}\Big(\delta^{\frac{1}{3}}+n^{-\frac{1}{4}}\Big)\\
&=\mathcal{O}\Big(\delta^{\frac{1}{3}}+n^{-\frac{1}{4}}\Big).
\end{aligned}\tag{82}$$

As a result, we have

$$\mathbb{E}\left\|w_S-\widetilde{w}_S\right\|\leq\mathcal{O}\Big(\delta^{\frac{1}{6}}+n^{-\frac{1}{8}}\Big). \tag{83}$$

Further incorporating (78), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{A},S}\Big\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S))-\nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\Big\|\\&\leq 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{L-\nu}}+\mathbb{E}\left\|\widetilde{w}_S-w_S\right\|\\&\leq\mathcal{O}\big(\sqrt{\nu}\big)+\mathbb{E}\left\|\widetilde{w}_S-w_S\right\|\\&=\mathcal{O}\Big(\delta^{\frac{1}{6}}+n^{-\frac{1}{8}}\Big),\end{aligned} \tag{84}$$

which concludes the proof. □

F Proof of Corollary 4.5 and 4.6

Corollary F.1. Assume the function f is NC-SC as defined in Assumption 2.1, then if we run SGDA for T iterations with stepsize $(\alpha_x,\alpha_y)=\left(\frac{c}{t},\frac{cr^2}{t}\right)$ for some constant $c>0$ and $1\leq r<\kappa$, we have

$$\mathbb{E}_{S,\mathcal{A}}\left\|\nabla\Phi(\mathcal{A}_x(S))-\nabla\Phi_S(\mathcal{A}_x(S))\right\|\leq(1+\kappa)\left(\frac{8G\Big(1+\frac{1}{cL(r+1)}\Big)}{n}(24\kappa cL(r+1))^{\frac{1}{cL(r+1)+1}}T^{\frac{cL(r+1)}{cL(r+1)+1}}+\frac{G}{\sqrt{n}}\right). \tag{85}$$

Proof. Denote $\Delta_t\triangleq\sqrt{\left\|x_t-x_t'\right\|^2+\left\|y_t-y_t'\right\|^2}$, and the event $E_{t_0}=\mathbf{1}(\Delta_{t_0}=0)$, we have for the full gradient $\nabla f=(\nabla_x f,\nabla_y f)^\top$,

$$\begin{aligned}&\mathbb{E}\|\nabla f(x_t,y^*(x_t);\xi)-\nabla f(x_t',y^*(x_t');\xi)\|\\&=\mathbb{P}(E_{t_0})\mathbb{E}[\|\nabla f(x_t,y^*(x_t);\xi)-\nabla f(x_t',y^*(x_t');\xi)\||E_{t_0}]\\&\qquad+\mathbb{P}(E_{t_0}^C)\mathbb{E}\big[\|\nabla f(x_t,y^*(x_t);\xi)-\nabla f(x_t',y^*(x_t');\xi)\||E_{t_0}^C\big]\\&\leq\mathbb{E}[\|\nabla f(x_t,y^*(x_t);\xi)-\nabla f(x_t',y^*(x_t');\xi)\||E_{t_0}]+2G\mathbb{P}(E_{t_0}^C)\\&\leq\mathbb{E}[\|\nabla_x f(x_t,y^*(x_t);\xi)-\nabla_x f(x_t',y^*(x_t');\xi)\|+\|\nabla_y f(x_t,y^*(x_t);\xi)-\nabla_y f(x_t',y^*(x_t');\xi)\||E_{t_0}]+2G\mathbb{P}(E_{t_0}^C)\\&\leq 2L\mathbb{E}[\|x_t-x_t'\|+\|y^*(x_t)-y^*(x_t')\||E_{t_0}]+2G\mathbb{P}(E_{t_0}^C)\\&\leq 2(1+\kappa)L\mathbb{E}[\|x_t-x_t'\||E_{t_0}]+2G\frac{t_0}{n}\\&\leq 4\kappa L\mathbb{E}[\Delta_t|\Delta_{t_0}=0]+2G\frac{t_0}{n},\end{aligned} \tag{86}$$

the remaining steps aims to bound $\mathbb{E}[\Delta_t|\Delta_{t_0}=0]$, which are the same as those in (Farnia and Ozdaglar, 2021, Appendix B.8), with that we will get

$$\mathbb{E}\|\nabla f(x_T,y^*(x_T);\xi)-\nabla f(x_T',y^*(x_T');\xi)\|\leq\frac{4\kappa L\cdot 12G}{nL}\left(\frac{T}{t_0}\right)^{cL(r+1)}+\frac{2G}{n}t_0, \tag{87}$$

to minimize the RHS above over t_0, we set

$$t_0=\left(\frac{\frac{4\kappa L\cdot 12G}{nL}\cdot cL(r+1)}{\frac{2G}{n}}\right)^{\frac{1}{cL(r+1)+1}}\cdot T^{\frac{cL(r+1)}{cL(r+1)+1}}=(24\kappa cL(r+1))^{\frac{1}{cL(r+1)+1}}T^{\frac{cL(r+1)}{cL(r+1)+1}} \tag{88}$$

and we get

$$\mathbb{E}\|\nabla f(x_T,y^*(x_T);\xi)-\nabla f(x_T',y^*(x_T');\xi)\|\leq\frac{2G\Big(1+\frac{1}{cL(r+1)}\Big)}{n}(24\kappa cL(r+1))^{\frac{1}{cL(r+1)+1}}T^{\frac{cL(r+1)}{cL(r+1)+1}}. \tag{89}$$

We conclude the proof by incorporating the above bound with Theorem 4.2. □

Corollary F.2. Assume the function f is NC-C as defined in Assumption 2.1 with $\mu = 0$, then if we run SGDA for T iterations with stepsize $\max\{\alpha_x, \alpha_y\} \leq \frac{c}{t}$ for some constant $c > 0$, we have

$$\mathbb{E}_{S,\mathcal{A}} \left\| \nabla \Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla \Phi_S^{1/(2L)}(\mathcal{A}_x(S)) \right\| \leq \mathcal{O}\left(\left(\frac{T^{\frac{cL}{cL+1}}}{n} \right)^{1/6} + \left(\frac{1}{n} \right)^{1/8} \right). \tag{90}$$

Proof. Denote $\Delta_t \triangleq \sqrt{\|x_t - x_t'\|^2 + \|y_t - y_t'\|^2}$, and the event $E_{t_0} = \mathbf{1}(\Delta_{t_0} = 0)$, we have that

$$\begin{aligned} \mathbb{E}\|x_t - x_t'\| &\leq \mathbb{E}\, \Delta_t \\ &= \mathbb{P}(E_{t_0})\mathbb{E}[\Delta_t | E_{t_0}] + \mathbb{P}(E_{t_0}^C)\mathbb{E}\big[\Delta_t | E_{t_0}^C\big] \\ &\leq \mathbb{E}[\Delta_t | E_{t_0}] + 2\sqrt{D_{\mathcal{X}} + D_{\mathcal{Y}}}\mathbb{P}(E_{t_0}^C) \\ &\leq \mathbb{E}[\Delta_t | \Delta_{t_0} = 0] + 2\sqrt{D_{\mathcal{X}} + D_{\mathcal{Y}}}\frac{t_0}{n}, \end{aligned} \tag{91}$$

the remaining steps aims to bound $\mathbb{E}[\Delta_t | \Delta_{t_0} = 0]$, with the results in (Farnia and Ozdaglar, 2021, Appendix B.9), we get

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{2G}{nL}\left(\frac{T}{t_0} \right)^{cL} + 2\sqrt{D}\frac{t_0}{n}, \tag{92}$$

where we let $D = D_{\mathcal{X}} + D_{\mathcal{Y}}$. To minimize the RHS above over t_0, we set

$$t_0 = \left(\frac{cG}{\sqrt{D}} \right)^{1/(cL+1)} T^{\frac{cL}{cL+1}}, \tag{93}$$

and we get

$$\mathbb{E}\|x_T - x_T'\| \leq 2\left(\frac{G}{L}\left(\frac{1}{cG} \right)^{\frac{cL}{cL+1}} + (cG)^{\frac{1}{cL+1}} \right) D^{\frac{cL}{2(cL+1)}} \frac{T^{\frac{cL}{cL+1}}}{n}. \tag{94}$$

The proof is complete by incorporating the above bound with Theorem 4.4. □

G Proof of Corollary 4.8 and 4.9

Proof. For the NC-SC case, by Lei (2022, Corollary 6), we know the algorithm is δ uniformly primal stable in gradients with $\delta = 2G\sqrt{T/n}$, the proof is complete by Theorem 4.2.

For the NC-C case, we want to derive the uniform primal argument stability, the flow here is almost the same as the proof of Lei (2022, Corollary 6), let $\Omega_t = \|x_t - x_t'\|^2$, define the event E_Ω as that the only different data point ξ_i is selected by the algorithm $\mathcal{A}$, so we have

$$\mathbb{E}\left[\Omega_t\right] \leq \mathbb{E}\left[\Omega_t \mid E_\Omega\right]P(E_\Omega) + \mathbb{E}\left[\Omega_t \mid E_\Omega^C\right]\mathbb{P}(E_\Omega^C) \leq \mathbb{E}\left[\Omega_t \mid E_\Omega^C\right]\frac{T}{n} \leq \frac{4D_{\mathcal{X}}T}{n}, \tag{95}$$

so the algorithm is $\sqrt{4D_{\mathcal{X}}T/n}$-uniformly primal argument stable. Then we conclude the proof by substituting the above stability results into Theorem 4.4, i.e.,

$$\begin{aligned} &\mathbb{E}_{S,\mathcal{A}} \left\| \nabla \Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla \Phi_S^{1/(2L)}(\mathcal{A}_x(S)) \right\| \\ &= \mathcal{O}\Big(\delta^{\frac{1}{6}} + n^{-\frac{1}{8}}\Big) = \mathcal{O}\left(\left(\frac{T}{n} \right)^{\frac{1}{12}} + \left(\frac{1}{n} \right)^{\frac{1}{8}} \right) = \mathcal{O}\left(\left(\frac{T}{n} \right)^{\frac{1}{12}} + \left(\frac{1}{n} \right)^{\frac{1}{8}} \right). \end{aligned} \tag{96}$$

□