OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Uniform Convergence and Generalization for Nonconvex Stochastic Minimax Problems

Siqi Zhang[*] SZHAN207@JHU.EDU
Department of Applied Mathematics and Statistics, Johns Hopkins University, USA

Yifan Hu[*] YIFAN.HU@EPFL.CH
Risk Analytics and Optimization Chair, EPFL, Switzerland

Liang Zhang LIANG.ZHANG@INF.ETHZ.CH
Niao He NIAO.HE@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Switzerland

Abstract

This paper studies the uniform convergence and generalization bounds for nonconvex-(strongly)-concave (NC-SC / NC-C) stochastic minimax optimization. We first establish the uniform convergence between the empirical minimax problem and the population minimax problem and show the $\tilde{\mathcal{O}}\big(d\kappa^2\epsilon^{-2}\big)$ and $\tilde{\mathcal{O}}\big(d\epsilon^{-4}\big)$ sample complexities respectively for the NC-SC and NC-C settings, where d is the dimension number and κ is the condition number. To the best of our knowledge, this is the first uniform convergence result measured by the first-order stationarity in stochastic minimax optimization literature.

1. Introduction

In this paper, we consider nonconvex 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 problems widely appear in practical applications like adversarial training [27, 42], generative adversarial networks (GANs) [12, 19, 33], reinforcement learning [5, 6, 16] and robust training [37]. The distribution $\mathcal{D}$ is often unknown and one generally only has access to a dataset $S=\{\xi_1,\cdots,\xi_n\}$ consisting of n i.i.d. samples from $\mathcal{D}$ and instead solves the following *empirical minimax problem*:

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

Since functions F and F_S are nonconvex in x and pursuing their global optimal solutions is intractable in general, instead one aims to design an algorithm $\mathcal{A}$ that finds an ϵ-*stationary point*, i.e., $\|\nabla\Phi(\mathcal{A}_x(S))\|\le\epsilon$ or $\mathbf{dist}\left(0,\partial\Phi(\mathcal{A}_x(S))\right)\le\epsilon$, where $\Phi(x)\triangleq\max_{y\in\mathcal{Y}}F(x,y)$ and $\Phi_S(x)\triangleq$

$\max_{y\in\mathcal{Y}} F_S(x,y)$ are primal functions, $\mathcal{A}_x(S)$ 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 $\mathbf{dist}\,(0,\partial\Phi(\cdot))$ [7].

Take the NC-SC setting as an example. The optimization error for solving the population minimax problem (1) consists of two terms[1]:

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

where the first term on the right-hand-side corresponds to the optimization error of solving the empirical minimax problem (2) and the second term corresponds to the generalization error. Such decomposition on the gradient norm has been studied recently in nonconvex minimization, e.g., [8, 11, 28]. Recently, there is a line of work that develops efficient algorithms for solving the empirical minimax problems, which gives a hint on the optimization error; see e.g., [26, 46], just to list a few. However, a full characterization of the generalization error is still lacking.

Characterizing the generalization error is not easy as both Φ_S and $\mathcal{A}_x(S)$ depend on the dataset S, which induces some correlation. One way to address such dependence issue in generalization bounds is to establish the *stability argument* of specific algorithms in stochastic optimization [4, 15, 35] and stochastic minimax optimization [2, 9, 20, 48]. However, these stability-based generalization bounds have several drawbacks:

1. Generally, they require case-by-case analysis for different algorithms, i.e., these bounds are *algorithm-dependent*.
2. Existing stability analysis only applies to simple gradient-based algorithms for minimization and minimax problems (note that for minimax optimization, simple algorithms such as stochastic gradient descent ascent often turn out to be suboptimal), yet such analysis can be difficult to generalize to more sophisticated state-of-art algorithms.
3. Existing stability analysis generally requires specific parameters (e.g., stepsizes), which may misalign with those required for convergence analysis, thus making the generalization bounds less informative.
4. Existing stability-based generalization bounds generally use function value-based gap as the measurement of the algorithm, which may not be suitable concerning the nonconvex landscape. To the best of our knowledge, there are no generalization bound results measured by the first-order stationarity in nonconvex minimax optimization.

To overcome these difficulties, we aim to derive generalization bounds via establishing the *uniform convergence* from the empirical minimax optimization to the population minimax problem, i.e., $\mathbb{E}\sup_{x\in\mathcal{X}}\|\nabla\Phi(x)-\nabla\Phi_S(x)\|$. Note that uniform convergence is invariant to the choice of algorithms and provides an upper bound on the generalization error for any $\mathcal{A}_x(S)\in\mathcal{X}$, thus the derived generalization bound is *algorithm-agnostic*. Although uniform convergence has been extensively studied in the literature of stochastic optimization [8, 17, 28], a key difference in uniform

[1] Here for simplicity of illustration, we assume there is no constraint and primal functions are differentiable, detailed setting will be formally introduced in Section 2.

convergence for minimax optimization is that the primal function cannot be written as the average over n i.i.d. random functions and one needs to additionally characterize the differences between $\operatorname{argmax}_{y\in\mathcal{Y}} F_S(x,y)$ and $\operatorname{argmax}_{y\in\mathcal{Y}} F(x,y)$. Thus techniques in uniform convergence for classical stochastic optimization are not directly applicable.

Related Works In the NC-SC setting, many algorithms have been proposed in the literature, e.g., [3, 13, 23–26, 29, 36, 43–45], also recent years witnessed a surge of algorithms for NC-C problems in deterministic, finite-sum, and stochastic settings, e.g., [3, 23, 29, 30, 38, 46, 49, 52], to name a few. Uniform convergence is a very important topic and has been extensively studied in statistical learning theory [10, 39, 40] and stochastic optimization [8, 17, 28, 40], but to the best of our knowledge, we have not found any work that investigates the uniform convergence measured by the primal stationarity in nonconvex minimax optimization.

Contributions In this work, 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 the gradients of primal functions (or its Moreau envelope). Our results provide an algorithm-agnostic generalization bound for any algorithms that solve the empirical nonconvex minimax problem. Specifically, the sample complexities to achieve an ϵ-uniform convergence and 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.

2. Problem Setting

First, we introduce the main assumptions used throughout the paper. For more notations and standard definitions, we refer readers to check Appendix A.

Assumption 1 (Main Settings) *We assume the following:*

- *The function $f(x,y;\xi)$ is L-smooth jointly in $(x,y)\in\mathcal{X}\times\mathcal{Y}$ for any ξ.*
- *The function $f(x,y;\xi)$ is μ-strongly concave in $y\in\mathcal{Y}$ for any $x\in\mathcal{X}$ and any ξ where $\mu\geq 0$.*
- *The gradient norms of $f(\cdot,\cdot;\xi)$ and $\Phi(\cdot)$ are bounded by $G, G_\Phi > 0$ respectively for any ξ.*
- *The domains $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets, i.e., there exists constants $D_\mathcal{X}, D_\mathcal{Y} > 0$ such that for any $x\in\mathcal{X}$, $\|x\|^2\leq D_\mathcal{X}$ and for any $y\in\mathcal{Y}$, $\|y\|^2\leq D_\mathcal{Y}$, respectively.*

Note that compact domain assumption is widely used in uniform convergence literature [8, 17]. Under Assumption 1, the objective function F is L-smooth in (x,y) and μ-strongly concave for any ξ. When $\mu>0$, we call the population minimax problem (1) a *nonconvex-strongly-concave* (NC-SC) minimax problem; when $\mu=0$, we call it a *nonconvex-concave* (NC-C) minimax problem.

Definition 1 (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:*

$$\Phi^\lambda(x)\triangleq\min_{z\in\mathcal{X}}\left\{\Phi(z)+\frac{1}{2\lambda}\|z-x\|^2\right\},\quad \mathbf{prox}_{\lambda\Phi}(x)\triangleq\operatorname*{argmin}_{z\in\mathcal{X}}\left\{\Phi(z)+\frac{1}{2\lambda}\|z-x\|^2\right\}. \tag{4}$$

Performance Measurement In the NC-SC setting, the primal functions Φ and Φ_S are both $\tilde{L}$-smooth. Regarding the constraint, we measure the difference between the population and empirical minimax problems using *the generalized gradient of the population and the empirical primal functions*, i.e., $\mathbb{E}\left\|\mathcal{G}_{\Phi}(\mathcal{A}_x(S)) - \mathcal{G}_{\Phi_S}(\mathcal{A}_x(S))\right\|$, where $\mathcal{G}_{\Phi}(x) \triangleq \tilde{L}(x - \mathbf{proj}_{\mathcal{X}}(x - (1/\tilde{L})\nabla\Phi(x)))$. The following inequality summarized the relationship of measurements in terms of generalized gradient and in terms of gradient used in Section 1.

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

where the first inequality holds as projection is a non-expansive operator. The term in the left-hand side (LHS) above is the generalization error of an algorithm $\mathcal{A}$ we desire in the NC-SC case.

For the NC-C case, as the primal function $\Phi(x)$ is L-weakly convex, we use the gradient of its Moreau Envelope to characterize the (near)-stationarity [7]. We measure the difference between the population and empirical problems using the difference between *the gradients of their respective Moreau envelopes*. The generalization error and the uniform convergence in the NC-C case are as follows:

$$\underbrace{\mathbb{E}\left\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\right\|}_{\text{generalization error of Algorithm } \mathcal{A}} \leq \underbrace{\mathbb{E}\left[\max_{x\in\mathcal{X}}\left\|\nabla\Phi^{1/(2L)}(x) - \nabla\Phi_S^{1/(2L)}(x)\right\|\right]}_{\text{algorithm-agnostic uniform convergence}}. \tag{5}$$

The term in the LHS above is the generalization error of an algorithm $\mathcal{A}$ we desire in the NC-C case.

3. Uniform Convergence and Generalization Bounds

In this section, we discuss the sample complexity for achieving ϵ-uniform convergence and ϵ-generalization error for NC-SC and NC-C stochastic minimax optimization.

3.1. NC-SC Stochastic Minimax Optimization

Under the NC-SC setting, we demonstrate in the following theorem the uniform convergence between gradients of primal functions of the population and empirical minimax problems, which provides an upper bound on the generalization error for any algorithm $\mathcal{A}$. We defer the proof to Appendix C.

Theorem 2 (Uniform Convergence and Generalization Error, NC-SC) *Under Assumption 1 with $\mu > 0$, we have*

$$\mathbb{E}\left[\max_{x\in\mathcal{X}}\left\|\nabla\Phi(x) - \nabla\Phi_S(x)\right\|\right] = \tilde{\mathcal{O}}\Big(d^{1/2}\kappa n^{-1/2}\Big). \tag{6}$$

Furthermore, to achieve ϵ-uniform convergence and ϵ-generalization error for any algorithm $\mathcal{A}$ such that $\mathbb{E}\left\|\mathcal{G}_{\Phi}(\mathcal{A}_x(S)) - \mathcal{G}_{\Phi_S}(\mathcal{A}_x(S))\right\| \leq \epsilon$, it suffices to have

$$n = n^*_{\text{NCSC}} \triangleq \tilde{\mathcal{O}}\big(d\kappa^2\epsilon^{-2}\big). \tag{7}$$

To the best of our knowledge, it is the first uniform convergence and algorithm-agnostic generalization error bound result for NC-SC stochastic minimax problem. In comparison, existing works in the generalization error analysis [9, 20] utilize stability arguments for certain algorithms and thus are

algorithm-specific. Zhang et al. [50] establish algorithm-agnostic stability and generalization in the strongly-convex-strongly-concave regime, yet their analysis does not extend to the nonconvex regime. Our generalization results apply to any algorithms for solving finite-sum problems, especially the SOTA algorithms like Catalyst-SVRG [51] and finite-sum version SREDA [26]. These algorithms are generally very complicated, and they lack stability-based generalization bounds.

The achieved sample complexity further implies that for any algorithm that achieves an ϵ-stationarity point of the empirical minimax problem, its sample complexity for finding an ϵ-stationary point of the population minimax problem is $\tilde{\mathcal{O}}\big(d\kappa^2\epsilon^{-2}\big)$. In terms of the dependence on the accuracy ϵ and the condition number κ, such sample complexity is better than the SOTA sample complexity results achieved via directly applying gradient-based methods on the population minimax optimization, i.e., $\mathcal{O}\big(\kappa^2\epsilon^{-4}\big)$ by Stochastic Smoothed-AGDA [47] and $\mathcal{O}\big(\kappa^3\epsilon^{-3}\big)$ by SREDA [26].

3.2. NC-C Stochastic Minimax Optimization

In this subsection, we derive the uniform convergence and algorithm-agnostic generalization bounds for NC-C stochastic minimax problems. Recall that the primal function Φ is L-weakly convex [38] and $\nabla\Phi$ is not well-defined. We use the gradient of the Moreau envelope of the primal function as the measurement [7].

Theorem 3 (Uniform Convergence and Generalization Error, NC-C) *Under Assumption 1 with $\mu = 0$, we have*

$$\mathbb{E}\left[\max_{x\in\mathcal{X}}\left\|\nabla\Phi_S^{1/(2L)}(x) - \nabla\Phi^{1/(2L)}(x)\right\|\right] = \tilde{\mathcal{O}}\Big(d^{1/4}n^{-1/4}\Big). \tag{8}$$

Furthermore, to achieve ϵ-uniform convergence and ϵ-generalization error for any algorithm $\mathcal{A}$ such that $\mathbb{E}\left[\left\|\nabla\Phi^{1/(2L)}(\mathcal{A}_x(S)) - \nabla\Phi_S^{1/(2L)}(\mathcal{A}_x(S))\right\|\right] \le \epsilon$, it suffices to have

$$n = n^*_{\mathrm{NCC}} \triangleq \tilde{\mathcal{O}}\big(d\epsilon^{-4}\big). \tag{9}$$

The proof of Theorem 3 is deferred to Appendix D. To the best of our knowledge, this is the first algorithm-agnostic generalization error result in NC-C stochastic minimax optimization. Similar to the NC-SC setting, Theorem 3 indicates that the sample complexity to guarantee an ϵ-generalization error in the NC-C case for any algorithm is $\tilde{\mathcal{O}}\big(d\epsilon^{-4}\big)$. In comparison, it is much better than the $\tilde{\mathcal{O}}\big(\epsilon^{-6}\big)$ sample complexity achieved by the SOTA stochastic approximation-based algorithms [32] for NC-C stochastic minimax optimization for small accuracy ϵ and moderate dimension d.

4. Conclusion

In this paper, we take an initial step towards understanding the uniform convergence and corresponding generalization performances of NC-SC and NC-C minimax problems measured by the first-order stationarity. We hope that this work will shed light on the design of algorithms with improved complexities for solving stochastic nonconvex minimax optimization.

Several future directions are worthy of further investigation. It remains interesting to see whether we can improve the uniform convergence results under the NC-C setting, particularly the dependence on accuracy ϵ. In terms of generalization bounds, it remains open to derive algorithm-specific stability-based generalization bounds under the stationarity measurement.

Acknowledgement

Siqi Zhang and Yifan Hu contributed equally to this paper. This work was done while Siqi Zhang and Yifan Hu were at University of Illinois Urbana-Champaign (UIUC). Liang Zhang gratefully acknowledges funding by the Max Planck ETH Center for Learning Systems (CLS). Niao He is supported by ETH research grant funded through ETH Zurich Foundations and NCCR Automation funded through Swiss National Science Foundations.

References

[1] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[2] Digvijay Boob and Cristóbal Guzmán. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *arXiv preprint arXiv:2104.02988*, 2021.

[3] Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

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

[5] Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pages 1458–1467, 2017.

[6] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134, 2018.

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

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

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

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

[11] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family and beyond. *arXiv preprint arXiv:2104.14840*, 2021.

[14] Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.

[15] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[16] Jiawei Huang and Nan Jiang. On the convergence rate of density-ratio basedoff-policy policy gradient methods. In *Neural Information Processing Systems Offline Reinforcement Learning Workshop*, 2020.

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

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

[19] Qi Lei, Jason Lee, Alex Dimakis, and Constantinos Daskalakis. Sgd learns one-layer networks in wgans. In *International Conference on Machine Learning*, pages 5799–5808. PMLR, 2020.

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

[21] Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3384–3392, 2015.

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

[24] Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

[25] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

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

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

[29] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32:14934–14942, 2019.

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

[31] Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst for gradient-based nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 613–622. PMLR, 2018.

[32] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35, 2021.

[33] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. 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, 2018.

[34] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.

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

[36] Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod K Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. *arXiv preprint arXiv:2203.04850*, 2022.

[37] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[38] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32:12680–12691, 2019.

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

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

[41] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

[42] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595. PMLR, 2019.

[43] Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2020.

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

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

[46] Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. In *Advances in Neural Information Processing Systems*, 2020.

[47] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022.

[48] Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Varshney, Siwei Lyu, and Yiming Ying. Differentially private sgda for minimax problems. *arXiv preprint arXiv:2201.09046*, 2022.

[49] Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.

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

[51] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

[52] Renbo Zhao. A primal dual smoothing framework for max-structured nonconvex optimization. *arXiv preprint arXiv:2003.04375*, 2020.

Appendix 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.

[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) [23, 43]. We refer readers to [23, 47] 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.

- 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 $g(px_1 + (1-p)x_2) \geq pg(x_1) + (1-p)g(x_2)$.
- A function $h : \mathcal{X} \to \mathbb{R}$ is L-smooth[3] if h is continuously differentiable in $\mathcal{X}$ and there exists a constant $L > 0$ such that $\|\nabla h(x_1) - \nabla h(x_2)\| \leq L\|x_1 - x_2\|$ holds for any x_1, x_2.
- Given $\mu \geq 0$, we say a function $g : \mathcal{X} \to \mathbb{R}$ μ-strongly convex if $g(x) - \frac{\mu}{2}\|x\|^2$ is convex, and it is μ-strongly concave if $-g$ is μ-strongly convex. Also we say a function g is μ-weakly convex if $g(x) + \frac{\mu}{2}\|x\|^2$ is convex

Definition 4 (Smooth Function) *We say a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is L-smooth jointly in (x, y) if the function is continuous differentiable, and there exists a constant $L > 0$ such that for any $(x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$, we have $\|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| \leq L(\|x_1 - x_2\| + \|y_1 - y_2\|)$ and $\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \leq L(\|x_1 - x_2\| + \|y_1 - y_2\|)$.*

By definition, it is easy to find that an L-smooth function is also L-weakly convex. For completeness, we introduce the definition of a sub-Gaussian random variable and related lemma, which are important tools in the analysis.

Definition 5 (Sub-Gaussian Random Variable) *A random variable η is a zero-mean sub-Gaussian random variable with variance proxy σ_η^2 if $\mathbb{E}\,\eta = 0$ and either of the following two conditions hold:*

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

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

Lemma 6 (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 $j \neq i$), we have*

$$\left|h(\eta_1, \ldots, \eta_n) - h(\eta_1', \ldots, \eta_n')\right| \leq c_i.$$

For any $t > 0$, it holds that

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

Below we recall some important properties on the primal function Φ and its Moreau envelope $\Phi^\lambda(x)$ presented in the literature [7, 23, 38].

Lemma 7 (Properties of Φ and Φ^λ) *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*

[3] Here the smoothness definition for single-variable functions is subtly different from that of two-variable functions in Definition 4, so we list it here for completeness.

($\mu = 0$), the primal function Φ is L-weakly convex, and its its Moreau envelope $\Phi^\lambda(x)$ is differentiable, Lipschitz smooth, also

$$\nabla\Phi^\lambda(x) = \lambda^{-1}(x - \hat{x}), \quad \left\|\nabla\Phi^\lambda(x)\right\| \geq \mathbf{dist}\,(0, \partial\Phi(\hat{x})), \tag{10}$$

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 8 (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 $\|\mathcal{G}_\Phi(x)\| \leq \epsilon$, 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 $\left\|\nabla\Phi^{1/(2L)}(x)\right\| \leq \epsilon$.*

Appendix B. Additional Related Literature

Nonconvex Minimax Optimization Among the rich library of algorithms for NC-SC problems, [51] achieved the optimal complexity $\mathcal{O}(\sqrt{\kappa}\epsilon^{-2})$ in the deterministic case by introducing the Catalyst acceleration scheme [22, 31] into minimax problems, and Luo et al. [26], Zhang et al. [51] achieved the best complexity in the finite-sum case for now, which are $\mathcal{O}(\sqrt{n}\kappa^2\epsilon^{-2})$ and $\mathcal{O}(n^{3/4}\sqrt{\kappa}\epsilon^{-2})$, respectively. For the purely stochastic NC-SC minimax problems, Yang et al. [47] introduced a stochastic smoothed-AGDA algorithm, which achieves the best $\mathcal{O}(\kappa^2\epsilon^{-4})$ complexity, while Luo et al. [26] achieves the best $\mathcal{O}(\kappa^3\epsilon^{-3})$ complexity if further assuming average smoothness. The lower bounds of NC-SC problems in deterministic, finite-sum, and stochastic settings have been extensively studied recently in [14, 21, 51]

In general, NC-C problems are harder than NC-SC problems since besides nonconvexity, their primal functions can also be nonsmooth [23, 38]. To the best of our knowledge, [24, 38, 46] achieved the best $\tilde{\mathcal{O}}(\epsilon^{-3})$ complexity in the deterministic case, while [46] achieved the best $\tilde{\mathcal{O}}(n^{3/4}\epsilon^{-3})$ complexity in the finite-sum case, and [32] provided the best $\tilde{\mathcal{O}}(\epsilon^{-6})$ complexity in the purely stochastic case.

Uniform Convergence A series of works from stochastic optimization and statistical learning theory studied uniform convergence on the worst-case differences between the population objective $L(x)$ and its empirical objective $L_S(x)$ constructed via sample average approximation (SAA, also known as empirical risk minimization). Interested readers may refer to prominent results in statistical learning [10, 39, 40]. For finite-dimensional problems, Kleywegt et al. [17] showed that the sample complexity is $\mathcal{O}(d\epsilon^{-2})$ to achieve an ϵ-uniform convergence in high probability, i.e., $\mathbb{P}(\sup_{x\in\mathcal{X}} |L(x) - L_S(x)| \geq \epsilon)$. For nonconvex empirical objectives, Mei et al. [28] and Davis and Drusvyatskiy [8] established $\tilde{\mathcal{O}}(d\epsilon^{-2})$ sample complexity of uniform convergence measured by the stationarity for nonconvex smooth and weakly convex functions, respectively. For infinite-dimensional functional stochastic optimization with a finite VC-dimension, uniform convergence still holds [40]. In addition, Wang et al. [41] use uniform convergence to demonstrate the generalization and the gradient complexity of differential private algorithms for stochastic optimization.

Stability-Based Generalization Bounds Another line of research focuses on analyzing generalization bounds of stochastic optimization via the uniform stability property of specific algorithms, including SAA [4, 34], stochastic gradient descent [1, 15], and uniformly stable algorithms [18]. Recently, a series of works further extended the analysis to understand the generalization performances of various algorithms in minimax problems. Farnia and Ozdaglar [9] gave the generalization bound for the outputs of gradient-descent-ascent (GDA) and proximal-point algorithm (PPA) in both (strongly)-convex-(strongly)-concave and nonconvex-nonconcave smooth minimax problems. Lei et al. [20] focused on GDA and provided a comprehensive study for different settings of minimax problems with various generalization measures on function value gaps. Boob and Guzmán [2] provided stability and generalization results of extragradient algorithm (EG) in the smooth convex-concave setting. On the other hand, Zhang et al. [50] studied stability and generalization of the empirical minimax problem under the (strongly)-convex-(strongly)-concave setting, assuming that one can find the optimal solution to the empirical minimax problem. But as we discussed before, there are several key restrictions for the stability argument approach. To accommodate more involved algorithms, here we will study the generalization performances via the lens of uniform convergence.

Appendix C. Proof of Theorem 2

We first briefly sketch the main flow of the proof of Theorem 2, which consists of two main steps:
Step 1: First, we use a υ-net $\{x_k\}_{k=1}^Q$ [40] to decompose the error and handle the dependence issue between $\operatorname{argmax}_{x\in\mathcal{X}} \|\nabla\Phi_S(x) - \nabla\Phi(x)\|$ and $\Phi_S(x)$.
Step 2: For any x_k within the υ-net, we have the following decomposition

$$\begin{aligned}&\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\|\\ \leq&(\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| - \mathbb{E}\,\|\nabla\Phi_S(x) - \nabla\Phi(x_k)\|) + (\mathbb{E}\,\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\|).\end{aligned}$$

We first upper bound the second term $\mathbb{E}\,\|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\|$ in the right-hand side (RHS) using the stability argument as the maximization problem over y is strongly concave. Then we utilize the established stability argument to show that the first term in the RHS is sub-Gaussian and apply the concentration inequality.

Now we proceed to the proof of Theorem 2.

Proof To derive the desired generalization bounds, we take an υ-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 $\|x - x_k\| \leq \upsilon$. Note that such υ-net exists with $Q = \mathcal{O}(\upsilon^{-d})$ for compact $\mathcal{X}$ [17]. Utilizing the definition of the υ-net, we have

$$\begin{aligned}&\mathbb{E}\max_{x\in\mathcal{X}} \|\nabla\Phi_S(x) - \nabla\Phi(x)\|\\ \leq\ &\mathbb{E}\max_{x\in\mathcal{X}} [\|\nabla\Phi_S(x) - \nabla\Phi_S(x_k)\| + \|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| + \|\nabla\Phi(x_k) - \nabla\Phi(x)\|] \qquad (11)\\ \leq\ &\mathbb{E}\max_{k\in[Q]} \|\nabla\Phi_S(x_k) - \nabla\Phi(x_k)\| + 2L(1+\kappa)\upsilon,\end{aligned}$$

where the last inequality holds as Φ and Φ_S are $L(1+\kappa)$-smooth following Lemma 7. For any $s>0$, we have

$$\begin{aligned}
&\exp\left(s\,\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S(x)-\nabla\Phi(x)\|\right)\\
\leq\ &\exp\left(s\Big[\mathbb{E}\max_{k\in[Q]}\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\upsilon\Big]\right)\\
\leq\ &\mathbb{E}\max_{k\in[Q]}\exp\left(s\Big[\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\upsilon\Big]\right)\\
\leq\ &\mathbb{E}\sum_{k\in[Q]}\exp\left(s\Big[\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\upsilon\Big]\right)\\
=\ &\sum_{k\in[Q]}\mathbb{E}\exp\left(s\Big[\|\nabla\Phi_S(x_k)-\nabla\Phi(x_k)\|+2L(1+\kappa)\upsilon\Big]\right),
\end{aligned}\tag{12}$$

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 $\mathcal{O}(\upsilon^{-d})$. Without loss of generality, selecting υ such that $2L(1+\kappa)\upsilon=\frac{\epsilon}{2}$, we have

$$\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S(x)-\nabla\Phi(x)\|\\
\leq\ &\frac{1}{s}\log\Bigg(\sum_{k\in[Q]}\mathbb{E}\exp\left(s[\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|-\mathbb{E}\,\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|]\right)\\
&\qquad\cdot\exp\left(s\,\mathbb{E}\,\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|\right)\exp\left(\frac{s\epsilon}{2}\right)\Bigg).
\end{aligned}\tag{13}$$

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 $S=\{\xi_i\}_{i=1}^n$, $S^{(i)}=\{\xi_1,\ldots,\xi_{i-1},\xi_i',\xi_{i+1},\ldots,\xi_n\}$ 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

$$\begin{aligned}
&\mathbb{E}\|\nabla\Phi(x)-\nabla\Phi_S(x)\| = \mathbb{E}\left\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi) - \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y_S^*(x);\xi_i)\right\| \\
=\;& \mathbb{E}\Bigg\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi) - \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y^*(x);\xi_i) \\
&\qquad + \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y^*(x);\xi_i) - \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y_S^*(x);\xi_i)\Bigg\| \\
\le\;& \mathbb{E}\left\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi) - \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y^*(x);\xi_i)\right\| \\
&\qquad + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y^*(x);\xi_i) - \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y_S^*(x);\xi_i)\right\| \\
\le\;& \mathbb{E}\left\|\mathbb{E}_\xi\,\nabla_x f(x,y^*(x);\xi) - \frac{1}{n}\sum_{i=1}^n \nabla_x f(x,y^*(x);\xi_i)\right\| + L\|y^*(x)-y_S^*(x)\| \\
\le\;& \sqrt{\frac{\mathrm{Var}(\nabla_x f)}{n}} + L\|y^*(x)-y_S^*(x)\|,
\end{aligned} \tag{14}$$

where $\mathrm{Var}(\nabla_x f)$ is the variance of $\nabla_x f(\cdot,\cdot;\xi)$ and the second inequality holds by smoothness of f. Since the variance is upper bounded by the second moment:

$$\mathrm{Var}(\nabla_x f) \le \mathbb{E}\|\nabla_x f(x,y^*(x);\xi)\|^2 \le G^2, \tag{15}$$

it further holds that

$$\mathbb{E}\|\nabla\Phi(x)-\nabla\Phi_S(x)\| \le \frac{G}{\sqrt{n}} + L\|y^*(x)-y_S^*(x)\|. \tag{16}$$

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

$$\big(-F_S\big(x,y_{S^{(i)}}^*(x)\big)\big) - \big(-F_S(x,y_S^*(x))\big) \ge \frac{\mu}{2}\left\|y_{S^{(i)}}^*(x)-y_S^*(x)\right\|^2, \tag{17}$$

On the other hand, we have

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

where the last inequality holds by Lipschitz continuity and the optimality of $y^*_{S^{(i)}}(x)$. Combined with (17), it holds that

$$\left\|y^*_{S^{(i)}}(x) - y^*_S(x)\right\| \leq \frac{4G}{\mu n}.$$

In addition, we have

$$\begin{aligned}
&\mathbb{E}[F(x, y^*(x)) - F(x, y^*_S(x))] \\
=& \mathbb{E}\left[F(x, y^*(x)) - F_S(x, y^*(x))\right] + \mathbb{E}\left[F_S(x, y^*(x)) - F_S(x, y^*_S(x))\right] \\
&\qquad + \mathbb{E}\left[F_S(x, y^*_S(x)) - F(x, y^*_S(x))\right] \\
\leq& \mathbb{E}\left[F_S(x, y^*_S(x)) - F(x, y^*_S(x))\right] \\
=& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x, y^*_S(x); \xi_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\xi f(x, y^*_S(x); \xi)\right] \\
=& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x, y^*_S(x); \xi_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi_i} f(x, y^*_{S^{(i)}}(x); \xi_i)\right] \\
=& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(x, y^*_S(x); \xi_i) - \frac{1}{n}\sum_{i=1}^n f(x, y^*_{S^{(i)}}(x); \xi_i)\right] \\
\leq& G\,\mathbb{E}\left\|y^*_S(x) - y^*_{S^{(i)}}(x)\right\| \\
\leq& \frac{4G^2}{\mu n}
\end{aligned} \tag{18}$$

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

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

Plugging into (16), it holds that

$$\mathbb{E}\left\|\nabla\Phi(x_k) - \nabla\Phi_S(x_k)\right\| \leq L\sqrt{\frac{8G^2}{\mu^2 n}} + \frac{G}{\sqrt{n}}. \tag{19}$$

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

$$\begin{aligned}
&\|\nabla\Phi(x)-\nabla\Phi_S(x)\|-\|\nabla\Phi(x)-\nabla\Phi_{S^{(i)}}(x)\| \\
&\le \|\nabla\Phi_S(x)-\nabla\Phi_{S^{(i)}}(x)\| \\
&= \left\|\frac{1}{n}\sum_{j=1}^{n}\nabla_x f(x,y_S^*(x),\xi_j)-\frac{1}{n}\sum_{j\neq i}^{n}\nabla_x f\big(x,y_{S^{(i)}}^*(x),\xi_j\big)-\frac{1}{n}\nabla_x f\big(x,y_{S^{(i)}}^*(x),\xi_i'\big)\right\| \\
&\le L\big\|y_{S^{(i)}}^*(x)-y_S^*(x)\big\|+\frac{1}{n}\big\|\nabla_x f(x,y_{S^{(i)}}^*(x);\xi_i')-\nabla_x f(x,y_{S^{(i)}}^*(x);\xi_i)\big\| \\
&\le \frac{4LG/\mu+2G}{n},
\end{aligned} \tag{20}$$

where the first inequality uses triangle inequality, the first equality uses 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 6) and the definition of sub-Gaussian random variable, 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 variable, it holds that

$$\mathbb{E}\exp(s[\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|-\mathbb{E}\|\nabla\Phi(x_k)-\nabla\Phi_S(x_k)\|])\le\exp\left(\frac{s^2\sigma^2}{2}\right). \tag{21}$$

Plugging (19) and (21) into (13), we have

$$\mathbb{E}\|\nabla\Phi_S(x)-\nabla\Phi(x)\|\le\frac{log(Q)}{s}+\frac{s\sigma^2}{2}+L\sqrt{\frac{8G^2}{\mu^2 n}}+\frac{G}{\sqrt{n}}+\frac{\epsilon}{2} \tag{22}$$

Minimizing the right hand side over s, we have

$$\begin{aligned}
\mathbb{E}\|\nabla\Phi_S(x)-\nabla\Phi(x)\| &\le 2\sqrt{\frac{log(Q)\sigma^2}{2}}+L\sqrt{\frac{8G^2}{\mu^2 n}}+\frac{G}{\sqrt{n}}+\frac{\epsilon}{2} \\
&= \sqrt{\frac{2log(Q)(2LG/\mu+G)^2}{n}}+L\sqrt{\frac{8G^2}{\mu^2 n}}+\frac{G}{\sqrt{n}}+\frac{\epsilon}{2}.
\end{aligned} \tag{23}$$

Recall that $Q=\mathcal{O}(\upsilon^{-d})$ with $\upsilon=\epsilon/(4L(1+\kappa))$, 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

$$n=\mathcal{O}\left(2d\epsilon^{-2}(2LG/\mu+G)^2\log(4L(1+\kappa)\epsilon^{-1})\right)=\tilde{\mathcal{O}}(d\kappa^2\epsilon^{-2}) \tag{24}$$

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

Appendix D. Proof of Theorem 3

D.1. Important Lemma

The following lemma characterizes the distance between the proximal points of the primal function of the original NC-C problem Φ and the regularized NC-SC problem $\hat{\Phi}$. Note that the lemma may be of independent interest for the design and the analysis of gradient-based methods for NC-C problem.

Lemma 9 *For $\nu > 0$, 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*

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

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

We first provide the proof of Lemma 9.

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 [38, 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{25}$$

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 $y^*(\mathbf{prox}_{\lambda\Phi}(x))) = \text{argmax}_{y\in\mathcal{Y}} F(\mathbf{prox}_{\lambda\Phi}(x), y)$, the third inequality holds by optimality of $\hat{y}(\mathbf{prox}_{\lambda\Phi}(x))) = \text{argmax}_{y\in\mathcal{Y}} F(\mathbf{prox}_{\lambda\Phi}(x), y) - \frac{\nu}{2}\|y\|^2$, the second equality holds by definition of Φ, the fourth inequality holds by optimality of $\mathbf{prox}_{\lambda\Phi}(x) = \text{argmin}_{x\in\mathcal{X}}\{\Phi(z) + \frac{1}{2\lambda}\|z - x\|^2\}$, the last inequality holds by compact domain $\mathcal{Y}$. ■

D.2. Proof of Theorem 3

The analysis of Theorem 3 is closely related to the analysis of NC-SC setting and consists of three parts.

Step 1: By the expression of the gradient of the Moreau envelope, it holds that $\|\nabla\Phi_S^\lambda(x) - \nabla\Phi^\lambda(x)\| \leq \frac{1}{\lambda}\|\mathbf{prox}_{\lambda\Phi}(x) - \mathbf{prox}_{\lambda\Phi_S}(x)\|$. We first use a υ-net $\{x_k\}_{k=1}^Q$ [40] to handle the dependence issue between $\tilde{x}^* \triangleq \text{argmax}_{x\in\mathcal{X}} \|\mathbf{prox}_{\lambda\Phi}(x) - \mathbf{prox}_{\lambda\Phi_S}(x)\|$ and Φ_S.

Step 2: Then we build up a connection between NC-C stochastic minimax optimization problems and NC-SC stochastic minimax optimization problems via adding an ℓ_2-regularization and carefully choosing a regularization parameter.

Step 3: It remains to characterize the distance between $\mathbf{prox}_{\lambda\hat{\Phi}}(x)$ and $\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)$ and show that $\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x) - \mathbf{prox}_{\lambda\hat{\Phi}}(x)\| - \mathbb{E}\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x) - \mathbf{prox}_{\lambda\hat{\Phi}}(x)\|$ is a sub-Gaussian random variable. For the distance between $\mathbf{prox}_{\lambda\hat{\Phi}}(x)$ and $\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)$, by definition, it is equivalent to the difference between the optimal solutions on x of strongly-convex strongly-concave (SC-SC) population minimax problem and its empirical minimax problem. We utilize the existing stability-based results for SC-SC minimax optimization [50] to upper bound such distance and show the variable is sub-Gaussian.

Next, we demonstrate the proof of Theorem 3.

Proof By Lemma 9, we have

$$\|\mathbf{prox}_{\lambda\Phi}(x) - \mathbf{prox}_{\lambda\hat{\Phi}}(x)\| \leq \sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}};$$

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

To derive the desired uniform convergence, similar to the proof of Theorem 2, we take an υ-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 $\|x - x_k\| \leq \upsilon$. Note that such υ-net exists with $Q = \mathcal{O}(\upsilon^{-d})$ for compact $\mathcal{X}$. We first decompose the error as approximation error from NC-SC minimax problems to NC-C minimax problems. Then we utilize the υ-net to address the dependence between S and $\text{argmax}_{x\in\mathcal{X}} \|\nabla\Phi_S^\lambda(x) - \nabla\Phi^\lambda(x)\|$. First note that

$$
\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S^\lambda(x)-\nabla\Phi^\lambda(x)\| \\
=&\frac{1}{\lambda}\mathbb{E}\max_{x\in\mathcal{X}}\|\mathbf{prox}_{\lambda\Phi_S}(x)-\mathbf{prox}_{\lambda\Phi}(x)\| \\
\le&\frac{1}{\lambda}\mathbb{E}\max_{x\in\mathcal{X}}\|\mathbf{prox}_{\lambda\Phi_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)\|+\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\| \\
&\qquad\qquad+\|\mathbf{prox}_{\lambda\hat{\Phi}}(x)-\mathbf{prox}_{\lambda\Phi}(x)\| \\
\le&\frac{2}{\lambda}\sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}}+\frac{1}{\lambda}\mathbb{E}\max_{x\in\mathcal{X}}\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\| \\
\le&\frac{2}{\lambda}\sqrt{\frac{\lambda\nu D_{\mathcal{Y}}}{1-\lambda(L+\nu)}}+\frac{1}{\lambda}\mathbb{E}\max_{x\in\mathcal{X}}\big[\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x)-\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)\| \\
&\qquad\qquad+\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\|+\|\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x)\|\big] \\
\le&2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda(L+\nu))}}+\frac{1}{\lambda}\mathbb{E}\max_{k\in[Q]}\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\|+\frac{2\upsilon}{\lambda(1-\lambda(L+\nu))} \\
\le&2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda(L+\nu))}}+\frac{1}{\lambda s}\log\left(\sum_{k\in[Q]}\mathbb{E}\exp\left(s\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\right) \\
&\qquad\qquad+\frac{2\upsilon}{\lambda(1-\lambda(L+\nu))},
\end{aligned}
\tag{26}
$$

where the first and the third inequality use the triangle inequality, the second inequality uses Lemma 9 for Φ and Φ_S, x_k is the closest point to x in the υ-net, the fourth inequality holds by $(1-\lambda(L+\nu))^{-1}$-Lipschitz continuity of proximal operator [8, 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 (12). 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

$$\mathbf{prox}_{\lambda\hat{\Phi}}(x) = \operatorname*{argmin}_{z\in\mathcal{X}}\left\{\max_{y\in\mathcal{Y}} \mathbb{E}_\xi f(z,y;\xi) - \frac{\nu}{2}\|y\|^2 + \frac{1}{2\lambda}\|z-x\|^2\right\}, \tag{27}$$

$$\mathbf{prox}_{\lambda\hat{\Phi}_S}(x) = \operatorname*{argmin}_{z\in\mathcal{X}}\left\{\max_{y\in\mathcal{Y}} \frac{1}{n}\sum_{i=1}^n\left[f(z,y;\xi_i) - \frac{\nu}{2}\|y\|^2 + \frac{1}{2\lambda}\|z-x\|^2\right]\right\}. \tag{28}$$

Denote the solution of (27) as $(z^*(x), y^*(x))$ and the solution of (28) as $(z_S(x), y_S(x))$. We need to bound the distance between $z^*(x)$ and $z_S(x)$, note that these $(z^*(x), y^*(x))$ comes from a strongly-convex-strongly-concave stochastic minimax problem, where the modulus are $\frac{1-\lambda L}{\lambda}$ and ν, respectively; while the other comes from the sample average approximation counterpart. By [50, 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

$$\begin{aligned}\mathbb{E}\,\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x) - \mathbf{prox}_{\lambda\hat{\Phi}}(x)\| &= \mathbb{E}\,\|z_S(x)-z^*(x)\| \\ &\le \sqrt{\mathbb{E}\,\|z_S(x)-z^*(x)\|^2} \le \sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L} + \frac{\hat{L}_y^2}{\nu}\right)}.\end{aligned} \tag{29}$$

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 [50, 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 6) and the definition of sub-Gaussian random variable, 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 (29), it holds that

$$\begin{aligned}
&\mathbb{E}\exp\left(s\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\\
=\;&\mathbb{E}\exp\left(s\left[\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|-\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right]\right)\\
&\qquad\cdot\exp\left(s\,\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right)\\
\leq\;&\mathbb{E}\exp\left(s\left[\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|-\mathbb{E}\left\|\mathbf{prox}_{\lambda\hat{\Phi}_S}(x_k)-\mathbf{prox}_{\lambda\hat{\Phi}}(x_k)\right\|\right]\right)\\
&\qquad\cdot\exp\left(s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}\right)\\
\leq\;&\exp\left(\frac{s^2}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)\right)\exp\left(s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}\right),
\end{aligned}\tag{30}$$

where the second inequality uses definition of zero-mean sub-Gaussian random variable. Combining (30) with (26), for

$$\lambda=\frac{1}{2L},\quad \upsilon=\frac{\epsilon\lambda(1-\lambda L)}{8}=\frac{\epsilon}{32L},\quad s=\sqrt{2n\log(Q)\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)^{-1}},\tag{31}$$

it holds that

$$
\begin{aligned}
&\mathbb{E}\max_{x\in\mathcal{X}}\|\nabla\Phi_S^\lambda(x)-\nabla\Phi^\lambda(x)\| \\
&\le 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda(L+\nu))}}+\frac{2\upsilon}{\lambda(1-\lambda(L+\nu))} \\
&\quad+\frac{1}{\lambda s}\log\left(Q\exp\left(\frac{s^2}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right)\right)\right) \\
&\qquad\qquad+\frac{1}{\lambda s}\log\left(\exp\left(s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}\right)\right) \\
&\le 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda L)}}+\frac{1}{\lambda s}\log(Q)+\frac{1}{\lambda s}\frac{s^2}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right) \\
&\qquad\qquad+\frac{1}{\lambda s}s\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\Big(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\Big)}+\frac{2\upsilon}{\lambda(1-\lambda L)} \\
&= 2\sqrt{\frac{\nu D_{\mathcal{Y}}}{\lambda(1-\lambda L)}}+\frac{\log(Q)}{\lambda s}+\frac{1}{\lambda}\frac{s}{2n}\left(\frac{\hat{L}_x^2\lambda^2}{(1-\lambda L)^2}+\frac{\hat{L}_y^2\lambda}{\nu(1-\lambda L)}\right) \\
&\quad+\frac{1}{\lambda}\sqrt{\frac{2\lambda}{1-\lambda L}\frac{2\sqrt{2}}{n}\left(\frac{\hat{L}_x^2\lambda}{1-\lambda L}+\frac{\hat{L}_y^2}{\nu}\right)}+\frac{\epsilon}{4} \\
&= 2\sqrt{4L\nu D_{\mathcal{Y}}}+4L\sqrt{\frac{\log(Q)}{2n}\left(\frac{\hat{L}_x^2}{L^2}+\frac{\hat{L}_y^2}{\nu L}\right)}+2L\sqrt{\frac{4\sqrt{2}}{Ln}\left(\frac{\hat{L}_x^2}{L}+\frac{\hat{L}_y^2}{\nu}\right)}+\frac{\epsilon}{4} \\
&= 2\sqrt{4L\nu D_{\mathcal{Y}}}+4L\sqrt{\frac{\log(Q)}{2n}\left(\frac{\hat{L}_x^2}{L^2}+\frac{\hat{L}_y^2}{\nu L}\right)} \\
&\qquad\qquad+2L\sqrt{\frac{4\sqrt{2}}{Ln}\left(\frac{(G+4L\sqrt{D_{\mathcal{X}}})^2}{L}+\frac{(G+\nu\sqrt{D_{\mathcal{Y}}})^2}{\nu}\right)}+\frac{\epsilon}{4}.
\end{aligned}
\tag{32}
$$

Here the first equality holds by the selection of υ, 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 υ, 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 $Q=\mathcal{O}\left(\left(\frac{D_{\mathcal{X}}}{\upsilon}\right)^d\right)$, then we choose $\upsilon=\tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right)$ 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 (32) of order $\mathcal{O}(\epsilon)$, it suffices to have

$$
\nu=\mathcal{O}(\epsilon^2),\quad n=\mathcal{O}\left(\frac{\log(Q)}{\nu}\epsilon^{-2}\right)=\mathcal{O}\left(d\epsilon^{-4}\log\left(\epsilon^{-1}\right)\right),
\tag{33}
$$

which concludes the proof. ■

Comparison Among Minimization, NC-SC, and NC-C Settings For general stochastic nonconvex optimization $\min_{x\in\mathcal{X}} \mathbb{E}[f(x;\xi)]$, the sample complexity of achieving ϵ-uniform convergence,

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

between the gradient of the population problem and the empricial problem is $\tilde{\mathcal{O}}(d\epsilon^{-2})$ [8, 28]. For nonconvex minimax optimization, if we care about the uniform convergence in terms of the gradient of F, i.e.,

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

existing analysis in Mei et al. [28] directly gives a $\tilde{\mathcal{O}}(d\epsilon^{-2})$ sample complexity. However, since we care about the gradient of the primal function, the analysis becomes more complicated.

1. In the NC-SC setting, to establish the uniform convergence, we bound

$$\mathbb{E}\max_{x\in\mathcal{X}} \|\nabla\Phi_S(x) - \nabla\Phi(x)\| = \mathbb{E}\max_{x\in\mathcal{X}} \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f(x,y_S^*;\xi_i) - \mathbb{E}\nabla f(x,y^*;\xi_i) \right\|,$$

   where $y_S^*(x) \triangleq \operatorname{argmax}_{y\in\mathcal{Y}} F_S(x,y)$ and $y^*(x) \triangleq \operatorname{argmax}_{y\in\mathcal{Y}} F(x,y)$. The primal function Φ_S is not in the form of averaging over n samples and thus existing analysis for minimization problem is not directly applicable. In addition, as the optimal point $y_S^*(x)$ differs from $y^*(x)$, such difference brings in an additional error term. In the NC-SC case, such error is upper bounded by $\mathcal{O}(n^{-1/2})$, which is of the same scale of the error from establishing uniform convergence on x. Thus the eventual uniform convergence bound established in Theorem 2 is of the same order as that for minimization problem [8, 28] except for an additional dependence on the condition number κ.

2. In the NC-C case, since there may exists multiple maximizers, we have

$$y^* \in \mathcal{Y}^* = \operatorname{argmax}_{y\in\mathcal{Y}} \mathbb{E}f(x,y;\xi), \quad y_S^* \in \mathcal{Y}_S^* = \operatorname{argmax}_{y\in\mathcal{Y}} \frac{1}{n}\sum_{i=1}^{n} f(x,y;\xi_i).$$

   Thus, the distance between y^* and y_S^* may not be well-defined. Instead, we bound the distance between $\hat{y}_S^*(x) \triangleq \operatorname{argmax}_{y\in\mathcal{Y}} F_S(x,y) - \frac{\nu}{2}\|y\|^2$ and $\hat{y}^*(x) \triangleq \operatorname{argmax}_{y\in\mathcal{Y}} F(x,y) - \frac{\nu}{2}\|y\|^2$ for a small regularization parameter $\nu = \mathcal{O}(n^{-1/2})$. Such distance is controlled by $\mathcal{O}(n^{-1/4})$. Thus the sample complexity for achieving ϵ-uniform convergence for the NC-C case is large than that of the NC-SC case. We leave it for future investigation to see if one could achieve smaller sample complexity in the NC-C case via a better characterization of the extra error brought in by y in the NC-C setting.