# Stability and Generalization of Stochastic Gradient Methods for Minimax Problems

**Yunwen Lei** [*1] **Zhenhuan Yang** [*2] **Tianbao Yang** [3] **Yiming Ying** [2]

## Abstract

Many machine learning problems can be formulated as minimax problems such as Generative Adversarial Networks (GANs), AUC maximization and robust estimation, to mention but a few. A substantial amount of studies are devoted to studying the convergence behavior of their stochastic gradient-type algorithms. In contrast, there is relatively little work on understanding their generalization, i.e., how the learning models built from training examples would behave on test examples. In this paper, we provide a comprehensive generalization analysis of stochastic gradient methods for minimax problems under both convex-concave and nonconvex-nonconcave cases through the lens of algorithmic stability. We establish a quantitative connection between stability and several generalization measures both in expectation and with high probability. For the convex-concave setting, our stability analysis shows that stochastic gradient descent ascent attains optimal generalization bounds for both smooth and nonsmooth minimax problems. We also establish generalization bounds for both weakly-convex-weakly-concave and gradient-dominated problems. We report preliminary experimental results to verify our theory.

## 1. Introduction

In machine learning we often encounter minimax optimization problems, where the decision variables are partitioned into two groups: one for minimization and one for maximization. This framework covers many important problems as specific instantiations, including adversarial learning (Goodfellow et al., 2014), robust optimization (Chen et al., 2017; Namkoong & Duchi, 2017), reinforcement learning (Dai et al., 2018; Du et al., 2017) and AUC maximization (Gao et al., 2013; Lei & Ying, 2021b; Liu et al., 2018; Ying et al., 2016; Zhao et al., 2011). To solve these problems, researchers have proposed various efficient optimization algorithms, for which a representative algorithm is the stochastic gradient descent ascent (SGDA) due to its simplicity and widespread use in real-world applications.

There is a large amount of work on the convergence analysis of minimax optimization algorithms in different settings such as convex-concave (Nemirovski et al., 2009), strongly-convex-strongly-concave (SC-SC) (Balamurugan & Bach, 2016), nonconvex-concave (Rafique et al., 2018) and nonconvex-nonconcave (Liu et al., 2020; Yang et al., 2020) cases. However, there is relatively little work on studying the generalization, i.e., how the model trained based on the training examples would generalize to test examples. Indeed, a model with good performance on training data may not generalize well if the models are too complex. It is imperative to study the generalization error of the trained models to foresee their prediction behavior. This often entails the investigation of the tradeoff between optimization and estimation for an implicit regularization.

To our best knowledge, there is only two recent work on the generalization analysis for minimax optimization algorithms (Farnia & Ozdaglar, 2020; Zhang et al., 2020). The argument stability for the specific empirical saddle point (ESP) was studied (Zhang et al., 2020), which implies weak generalization and strong generalization bounds. However, the discussion there ignored optimization errors and nonconvex-nonconcave cases, which can be restrictive in practice. For SC-SC, convex-concave, nonconvex-nonconcave objective functions, the uniform stability of several gradient-based minimax learners was developed in a smooth setting (Farnia & Ozdaglar, 2020), including gradient descent ascent (GDA), proximal point method (PPM) and GDmax. While they developed optimal generalization bounds for PPM, their discussions did not yield vanishing risk bounds for GDA in the general convex-concave case since their generalization bounds grow exponentially in terms of the iteration number. Furthermore, the above mentioned papers only study generalization bounds in ex-

[*]Equal contribution [1]School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK [2]Department of Mathematics and Statistics, State University of New York at Albany, USA [3]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA. Correspondence to: Yiming Ying <yying@albany.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

pectation, and there is a lack of high-probability analysis.

In this paper, we leverage the lens of algorithmic stability to study the generalization behavior of minimax learners for both convex-concave and nonconvex-nonconcave problems. Our discussion shows how the optimization and generalization should be balanced for good prediction performance. Our main results are listed in Table 1. In particular, our contributions can be summarized as follows.

1. We establish a quantitative connection between stability and generalization for minimax learners in different forms including weak/strong primal-dual generalization, primal generalization and generalization with high probability. For the technical contributions, we introduce novel decompositions to handle the correlation between the primal model and dual model for connecting stability and generalization.

2. We establish stability bounds of SGDA for convex-concave problems, from which we derive its optimal population risk bounds under an appropriate early-stopping strategy. We consider several measures of generalization and show that the optimal population risk bounds can be derived even in the nonsmooth case. To the best of our knowledge, our results are the first-ever known population risk bounds for minimax problems in the nonsmooth setting and the high-probability format.

3. We further extend our analysis to the nonconvex-nonconcave setting and give the first generalization bounds for nonsmooth objective functions. Our analysis relaxes the range of step size for a controllable stability and implies meaningful primal population risk bounds under some regularity assumptions of objective functions, e.g., a decay of weak-convexity-weak-concavity parameter along the optimization process or a two-sided PL condition.

The paper is organized as follows. The related work is discussed in Section 1.1 and the minimax problem formulation is given in Section 2. The connection between stability and generalization is studied in Section 3. We develop population risk bounds in the convex-concave case in Section 4 and extend our discussions to the nonconvex-nonconcave case in Section 5. We report preliminary experiments in Section 6 and conclude the paper in Section 7.

### 1.1. Related Work

We first review related work of stochastic optimization for minimax problems. Convergence rates of order $O(1/\sqrt{T})$ were established for SGDA with $T$ iterations in the convex-concave case (Nedić & Ozdaglar, 2009; Nemirovski et al., 2009), which can be further improved for SC-SC problems (Balamurugan & Bach, 2016; Hsieh et al., 2019). These discussions were extended to nonconvex-strongly-concave (Lin et al., 2020; Luo et al., 2020; Rafique et al., 2018; Yan et al., 2020), nonconvex-concave (Lin et al., 2020; Thekumparampil et al., 2019) and nonconvex-nonconcave (Liu et al., 2020; Loizou et al., 2020; Yang et al., 2020) minimax optimization problems. All the above mentioned work consider the convergence rate of optimization errors, while the generalization analysis was much less studied (Farnia & Ozdaglar, 2020; Zhang et al., 2020).

We now survey related work on stability and generalization. The framework of stability analysis was established in a seminal paper (Bousquet & Elisseeff, 2002), where the celebrated concept of uniform stability was introduced. This stability was extended to study randomized algorithms (Elisseeff et al., 2005). It was shown that stability is closely related to the fundamental problem of learnability (Rakhlin et al., 2005; Shalev-Shwartz et al., 2010). Hardt et al. (2016) pioneered the generalization analysis of SGD via stability, which inspired several upcoming work to understand stochastic optimization algorithms based on different algorithmic stability measures, e.g., uniform stability (Chen et al., 2018; Lin et al., 2016; Madden et al., 2020; Mou et al., 2018; Richards et al., 2020), argument stability (Bassily et al., 2020; Lei & Ying, 2020; Liu et al., 2017), on-average stability (Kuzborskij & Lampert, 2018; Lei & Ying, 2021a), hypothesis stability (Charles & Papailiopoulos, 2018; Foster et al., 2019; London, 2017), Bayes stability (Li et al., 2020) and locally elastic stability (Deng et al., 2020).

## 2. Problem Formulation

Let $\mathcal{W}$ and $\mathcal{V}$ be two parameter spaces in $\mathbb{R}^d$. Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z}$ and $f : \mathcal{W} \times \mathcal{V} \times \mathcal{Z} \mapsto \mathbb{R}$. We consider the following minimax optimization problem:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w}, \mathbf{v}) := \mathbb{E}_{z\sim\mathbb{P}}[f(\mathbf{w}, \mathbf{v}; z)]. \tag{2.1}$$

In practice, we do not know $\mathbb{P}$ but instead have access to a dataset $S = \{z_1, \ldots, z_n\}$ independently drawn from $\mathbb{P}$. Then, we approximate $F$ by an empirical risk

$$F_S(\mathbf{w}, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \mathbf{v}; z_i).$$

We apply a (randomized) algorithm $A$ to the dataset $S$ and get a model $A(S) := (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \in \mathcal{W} \times \mathcal{V}$ as an approximate solution of the problem (2.1). Since the model $A(S)$ is trained based on the training dataset $S$, its empirical behavior as measured by $F_S$ may not generalize well to a test example (Bousquet & Elisseeff, 2002). We are interested in studying the test error (population risk) of $A(S)$. Unlike the standard statistical learning theory (SLT) setting where there is only a minimization of $\mathbf{w}$, we have different measures of population risk due to the minimax structure (Zhang et al., 2020). We collect the notations of these performance measures in Table A.1. Let $\mathbb{E}[\cdot]$ denote the expectation w.r.t. the randomness of both the algorithm $A$ and the dataset $S$.

| Algorithm | Reference | Assumption | Measure | Rate |
|---|---|---|---|---|
| ESP | Zhang et al. (2020) | $\rho$-SC-SC, Lip | Weak PD Risk | $O(1/(n\rho))$ |
| | | $\rho$-SC-SC, Lip, S | Strong PD Risk | $O(1/(n\rho^2))$ |
| R-ESP | | C-C, Lip | Weak PD Risk | $O(1/\sqrt{n})$ |
| SGDA, SGDmax | Farnia & Ozdaglar (2020) | $\rho$-SC-SC, Lip, S | Weak PD Generalization[1] | $O(\log(n)/(n\rho))$ |
| PPM | | C-C, Lip, S | Weak PD Risk | $O(1/\sqrt{n})$ |
| SGDA | This work | C-C, Lip (S) | Weak PD Risk | $O(1/\sqrt{n})$ |
| | | C-$\rho$-SC, Lip, S | (H.P.) Primal Risk | $O(1/(\sqrt{n}\rho))$ |
| | | C-C, Lip | H.P. Plain Risk | $O(\log(n)/\sqrt{n})$ |
| | | $\rho$-SC-SC, Lip | Weak PD Risk | $O(\sqrt{\log n}/(n\rho))$ |
| SGDA | Farnia & Ozdaglar (2020) | Lip, S | Weak PD Generalization | $O\big(T^{\frac{Lc}{Lc+1}}/n\big)$ |
| SGDA | This work | $\rho$-WC-WC, Lip | Weak PD Generalization | $O\big(T^{\frac{2c\rho}{2c\rho+3}}/n^{\frac{2c\rho+1}{2c\rho+3}}\big)$ |
| | | D, Lip, S | Weak PD Generalization | $O\big(1/\sqrt{n}+\sqrt{T}/n\big)$ |
| AGDA | | $\rho$-SC, PL, Lip, S | Primal Risk | $O\big(n^{-\frac{cL+1}{2cL+1}}\big)$ |

*Table 1.* Summary of Results. Bounds are stated in expectation or with high probability (H.P.). For risk bounds, the optimal $T$ (number of iterations) is chosen to trade-off generalization and optimization. Here, C-C means convex-concave, C-$\rho$-SC means convex-$\rho$-strongly-concave, $\rho$-SC means nonconvex-$\rho$-strongly-concave, Lip means Lipschitz continuity, S means the smoothness, D means a decay of weak-convexity-weak-concavity parameter along the optimization process as Eq. (5.1) and PL means the two-sided condition as Assumption 3 . AGDA means Alternating Gradient Descent Ascent and (R)-ESP means the (regularized)-empirical risk saddle point. $c$ is a parameter in the step size and $L$ is given in Assumption 2.

**Definition 1** (Weak Primal-Dual Risk)**.** The weak Primal-Dual (PD) population risk of a (randomized) model $(\mathbf{w},\mathbf{v})$ is defined as (Zhang et al., 2020)

$$\triangle^w(\mathbf{w},\mathbf{v})=\sup_{\mathbf{v}'\in\mathcal{V}}\mathbb{E}\big[F(\mathbf{w},\mathbf{v}')\big]-\inf_{\mathbf{w}'\in\mathcal{W}}\mathbb{E}\big[F(\mathbf{w}',\mathbf{v})\big].$$

The weak PD empirical risk of $(\mathbf{w},\mathbf{v})$ is defined as

$$\triangle^w_S(\mathbf{w},\mathbf{v})=\sup_{\mathbf{v}'\in\mathcal{V}}\mathbb{E}\big[F_S(\mathbf{w},\mathbf{v}')\big]-\inf_{\mathbf{w}'\in\mathcal{W}}\mathbb{E}\big[F_S(\mathbf{w}',\mathbf{v})\big].$$

We refer to $\triangle^w(\mathbf{w},\mathbf{v})-\triangle^w_S(\mathbf{w},\mathbf{v})$ as the weak PD generalization error of the model $(\mathbf{w},\mathbf{v})$.

**Definition 2** (Strong Primal-Dual Risk)**.** The strong PD population risk of a model $(\mathbf{w},\mathbf{v})$ is defined as

$$\triangle^s(\mathbf{w},\mathbf{v})=\sup_{\mathbf{v}'\in\mathcal{V}}F(\mathbf{w},\mathbf{v}')-\inf_{\mathbf{w}'\in\mathcal{W}}F(\mathbf{w}',\mathbf{v}).$$

The strong PD empirical risk of $(\mathbf{w},\mathbf{v})$ is defined as

$$\triangle^s_S(\mathbf{w},\mathbf{v})=\sup_{\mathbf{v}'\in\mathcal{V}}F_S(\mathbf{w},\mathbf{v}')-\inf_{\mathbf{w}'\in\mathcal{W}}F_S(\mathbf{w}',\mathbf{v}).$$

We refer to $\triangle^s(\mathbf{w},\mathbf{v})-\triangle^s_S(\mathbf{w},\mathbf{v})$ as the strong PD generalization error of the model $(\mathbf{w},\mathbf{v})$.

**Definition 3** (Primal Risk)**.** The primal population risk of a model $\mathbf{w}$ is defined as $R(\mathbf{w})=\sup_{\mathbf{v}\in\mathcal{V}}F(\mathbf{w},\mathbf{v})$. The primal empirical risk of $\mathbf{w}$ is defined as $R_S(\mathbf{w})=\sup_{\mathbf{v}\in\mathcal{V}}F_S(\mathbf{w},\mathbf{v})$. We refer to $R(\mathbf{w})-R_S(\mathbf{w})$ as the primal generalization error of the model $\mathbf{w}$, and $R(\mathbf{w})-\inf_{\mathbf{w}'}R(\mathbf{w}')$ as the excess primal population risk.

According to the above definitions, we know $\triangle^w(\mathbf{w},\mathbf{v})\leq\mathbb{E}\big[\triangle^s(\mathbf{w},\mathbf{v})\big]$ and $R(\mathbf{w})-R_S(\mathbf{w})$ is closely related to $\triangle^s(\mathbf{w},\mathbf{v})-\triangle^s_S(\mathbf{w},\mathbf{v})$. The key difference between the weak PD risk and the strong PD risk is that the expectation is inside of the supremum/infimum for weak PD risk, while outside of the supremum/infimum for strong PD risk. In this way, one does not need to consider the coupling between primal and dual models for studying weak PD risks, and has to consider this coupling for strong PD risks. Furthermore, we refer to $F(\mathbf{w},\mathbf{v})-F_S(\mathbf{w},\mathbf{v})$ as the plain generalization error as it is standard in SLT. An approach to handle a population risk is to decompose it into a generalization error (estimation error) and an empirical risk (optimization error) (Bousquet & Bottou, 2008). For example, the weak PD population risk can be decomposed as

$$\triangle^w(\mathbf{w},\mathbf{v})=\big(\triangle^w(\mathbf{w},\mathbf{v})-\triangle^w_S(\mathbf{w},\mathbf{v})\big)+\triangle^w_S(\mathbf{w},\mathbf{v}). \quad (2.2)$$

The generalization error comes from the approximation of $\mathbb{P}$ with $S$, while the empirical risk comes since the algorithm may not find the saddle point of $F_S$. Our basic idea is to use algorithmic stability to study the generalization error and use optimization theory to study the empirical risk.

We now introduce necessary definitions and assumptions. Denote $\|\cdot\|_2$ as the Euclidean norm and $\langle\cdot,\cdot\rangle$ as the inner product. A function $g:\mathcal{W}\mapsto\mathbb{R}$ is said to be $\rho$-strongly-convex ($\rho\geq 0$) iff for all $\mathbf{w},\mathbf{w}'\in\mathcal{W}$ there holds $g(\mathbf{w})\geq g(\mathbf{w}')+\langle\mathbf{w}-\mathbf{w}',\nabla g(\mathbf{w}')\rangle+\frac{\rho}{2}\|\mathbf{w}-\mathbf{w}'\|_2^2$, where $\nabla$ is the gradient operator. We say $g$ is convex if $g$ is 0-strongly-convex. We say $g$ is $\rho$-strongly concave if $-g$ is $\rho$-strongly convex and concave if $-g$ is convex.

**Definition 4.** Let $\rho\geq 0$ and $g:\mathcal{W}\times\mathcal{V}\mapsto\mathbb{R}$. We say

(a) $g$ is $\rho$-strongly-convex-strongly-concave ($\rho$-SC-SC) if for any $\mathbf{v}\in\mathcal{V}$, the function $\mathbf{w}\mapsto g(\mathbf{w},\mathbf{v})$ is $\rho$-

strongly-convex and for any $\mathbf{w} \in \mathcal{W}$, the function $\mathbf{v} \mapsto g(\mathbf{w}, \mathbf{v})$ is $\rho$-strongly-concave.

(b) $g$ is convex-concave if $g$ is 0-SC-SC.

(c) $g$ is $\rho$-weakly-convex-weakly-concave ($\rho$-WC-WC) if $g + \frac{\rho}{2}\big(\|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2\big)$ is convex-concave.

The following two assumptions are standard (Farnia & Ozdaglar, 2020; Zhang et al., 2020). Assumption 1 amounts to saying $f$ is Lipschitz continuous with respect to (w.r.t.) both $\mathbf{w}$ and $\mathbf{v}$. Let $\nabla_{\mathbf{w}} f$ denote the gradient w.r.t. $\mathbf{w}$.

**Assumption 1.** Let $G > 0$. Assume for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$, there holds $\big\|\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z)\big\|_2 \leq G$ and $\big\|\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z)\big\|_2 \leq G$.

**Assumption 2.** Let $L > 0$. For any $z$, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is said to be $L$-smooth, if the following inequality holds for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$

$$\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}'; z) \\ \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}'; z) \end{pmatrix} \right\|_2 \leq L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2.$$

### 2.1. Motivating Examples

The minimax formulation (2.1) has broad applications in machine learning. Here we give some examples.

**AUC Maximization.** Area Under ROC Curve (AUC) is a popular measure for binary classification. Let $h(\mathbf{w}; x)$ denote a scoring function parameterized by $\mathbf{w}$ at $x$. It was shown that AUC maximization for learning $h$ under the square loss reduces to the problem (Ying et al., 2016)

$$\min_{(\mathbf{w},a,b)\in\mathbb{R}^{d+2}} \max_{\alpha\in\mathbb{R}} \mathbb{E}[f(\mathbf{w}, a, b, \alpha; z)], \tag{2.3}$$

where $p = \mathbb{P}[y = 1]$ and $f(\mathbf{w}, a, b, \alpha; z) = (h(\mathbf{w}; x) - a)^2 \mathbb{I}_{[y=1]}/p + (h(\mathbf{w}; x) - b)^2 \mathbb{I}_{[y=-1]}/(1-p) + 2(1+\alpha) h(\mathbf{w}; x)(\mathbb{I}_{[y=-1]}/(1-p) - \mathbb{I}_{[y=1]}/p) - \alpha^2$ ($\mathbb{I}_{[\cdot]}$ is the indicator function). It is clear that $\alpha \mapsto f(\mathbf{w}, a, b, \alpha; z)$ is a (strongly) concave function. Depending on $h$, the function $f$ can be convex, nonconvex, smooth or nonsmooth.

**Generative Adversarial Networks.** GAN (Goodfellow et al., 2014) refers to a popular class of generative models that consider generative modeling as a game between a generator network $G_{\mathbf{v}}$ and a discriminator network $D_{\mathbf{w}}$. The generator network produces synthetic data from random noise $\xi \sim \mathbb{P}_\xi$, while the discriminator network discriminates between the true data and the synthetic data. In particular, a popular variant of GAN named as WGAN (Arjovsky et al., 2017) can be written as a minimax problem

$$\min_{\mathbf{w}} \max_{\mathbf{v}} \mathbb{E}[f(\mathbf{w}, \mathbf{v}; z, \xi)] = \mathbb{E}_z[D_{\mathbf{w}}(z)] - \mathbb{E}_\xi[D_{\mathbf{w}}(G_{\mathbf{v}}(\xi))].$$

[1]Primal generalization bounds were presented in Farnia & Ozdaglar (2020). However, the stability analysis there actually only implies bounds on weak PD risk.

While this problem is generally nonconvex-nonconcave, it is weakly-convex-weakly-concave under smoothness assumptions on $D$ and $G$ (Liu et al., 2020).

**Robust Estimation with minimax estimator**. Audibert & Catoni (2011) formulated robust estimation as a minimax problem as follows

$$\min_{\mathbf{w}} \max_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v}) = \frac{1}{n} \sum_{i=1}^{n} \psi(\ell_1(\mathbf{w}; z_i) - \ell_2(\mathbf{v}; z_i)),$$

where $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a truncated loss, and $\ell_1, \ell_2$ are Lipschitz continuous and convex loss functions. A typical truncated loss is $\psi(x) = \log(1 + |x| + x^2/2)\text{sign}(x)$ to compute a mean estimator under heavy-tailed distribution of data (Brownlees et al., 2015; Xu et al., 2020), where $\text{sign}(x)$ is the sign of $x$. The composition function $F_S$ can be nonconvex and nonsmooth since $\psi$ is nonconvex and $\ell_1, \ell_2$ can be nonsmooth. Following Xu et al. (2020), it can be shown that $F_S(\mathbf{w}, \mathbf{v})$ is weakly-convex-weakly-concave.

## 3. Connecting Stability and Generalization

A fundamental concept in our analysis is the algorithmic stability, which measures the sensitivity of an algorithm w.r.t. the perturbation of training datasets (Bousquet & Elisseeff, 2002). We say $S, S' \subset \mathcal{Z}$ are neighboring datasets if they differ by at most a single example. We introduce three stability measures to the minimax learning setting. The weak-stability and uniform-stability quantify the sensitivity measured by function values, while the argument-stability quantifies the sensitivity measured by arguments. We collect these notations of stabilities in Table A.2 in Appendix A.

**Definition 5** (Algorithmic Stability)**.** Let $A$ be a randomized algorithm, $\epsilon > 0$ and $\delta \in (0, 1)$. Then we say

(a) $A$ is $\epsilon$-weakly-stable if for all neighboring $S$ and $S'$, there holds

$$\sup_z \Big( \sup_{\mathbf{v}' \in \mathcal{V}} \mathbb{E}_A\big[f(A_{\mathbf{w}}(S), \mathbf{v}'; z) - f(A_{\mathbf{w}}(S'), \mathbf{v}'; z)\big] + \sup_{\mathbf{w}' \in \mathcal{W}} \mathbb{E}_A\big[f(\mathbf{w}', A_{\mathbf{v}}(S); z) - f(\mathbf{w}', A_{\mathbf{v}}(S'); z)\big] \Big) \leq \epsilon.$$

(b) $A$ is $\epsilon$-argument-stable in expectation if for all neighboring $S$ and $S'$, there holds

$$\mathbb{E}_A\left[ \left\| \begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S') \end{pmatrix} \right\|_2 \right] \leq \epsilon.$$

$A$ is $\epsilon$-argument-stable with probability at least $1 - \delta$ if with probability at least $1 - \delta$

$$\left\| \begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S') \end{pmatrix} \right\|_2 \leq \epsilon.$$

(c) $A$ is $\epsilon$-uniformly-stable with probability at least $1-\delta$ if with probability at least $1-\delta$

$$\sup_z \big[f(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S); z) - f(A_{\mathbf{w}}(S'), A_{\mathbf{v}}(S'); z)\big] \leq \epsilon.$$

Under Assumption 1, argument stability implies weak and uniform stability. As we will see, argument stability plays an important role in getting primal population risk bounds.

As our first main result, we establish a quantitative connection between algorithmic stability and generalization in the following theorem to be proved in Appendix B. Part (a) establishes the connection between weak-stability and weak PD generalization error. Part (b) and Part (c) establish the connection between argument stability and strong/primal generalization error under a further assumption on the strong convexity/concavity. Part (d) and Part (e) establish high-probability bounds based on the uniform stability, which are much more challenging to derive than bounds in expectation and are important to understand the variation of an algorithm in several independent runs (Bousquet et al., 2020; Feldman & Vondrak, 2019). Regarding the technical contributions, we introduce novel decompositions in handling the correlation between $A_{\mathbf{w}}(S)$ and $\mathbf{v}_S^* = \arg\sup_{\mathbf{v}} F(A_{\mathbf{w}}(S), \mathbf{v})$, especially for high-probability analysis.

**Theorem 1.** *Let $A$ be a randomized algorithm and $\epsilon > 0$.*

*(a) If $A$ is $\epsilon$-weakly-stable, then the weak PD generalization error of $(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ satisfies*

$$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle_S^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \epsilon.$$

*(b) If $A$ is $\epsilon$-argument-stable in expectation, the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is $\rho$-strongly-concave and Assumptions 1, 2 hold, then the primal generalization error satisfies*

$$\mathbb{E}_{S,A}\Big[R(A_{\mathbf{w}}(S)) - R_S(A_{\mathbf{w}}(S))\Big] \leq \big(1 + L/\rho\big)G\epsilon.$$

*(c) If $A$ is $\epsilon$-argument-stable in expectation, $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is $\rho$-SC-SC and Assumptions 1, 2 hold, then the strong PD generalization error satisfies*

$$\mathbb{E}_{S,A}\Big[\triangle^s(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle_S^s(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))\Big] \leq \big(1 + L/\rho\big)G\sqrt{2}\epsilon.$$

*(d) Assume $|f(\mathbf{w}, \mathbf{v}; z)| \leq R$ for some $R > 0$ and $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}, z \in \mathcal{Z}$. Assume for all $\mathbf{w}$, the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is $\rho$-strongly-concave and Assumptions 1, 2 hold. Let $\delta \in (0,1)$. If $A$ is $\epsilon$-uniformly stable almost surely (a.s.), then with probability at least $1-\delta$*

$$R(A_{\mathbf{w}}(S)) - R_S(A_{\mathbf{w}}(S)) = O\Big(GL\rho^{-1}\epsilon \log n \log(1/\delta) + Rn^{-\frac{1}{2}}\sqrt{\log(1/\delta)}\Big).$$

*(e) Assume $|f(\mathbf{w}, \mathbf{v}; z)| \leq R$ for some $R > 0$ and $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}, z \in \mathcal{Z}$. Let $\delta \in (0,1)$. If $A$ is $\epsilon$-uniformly-stable a.s., then with probability $1-\delta$ there holds*

$$\big|F(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - F_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))\big| = O\Big(\epsilon \log n \log(1/\delta) + Rn^{-\frac{1}{2}}\sqrt{\log(1/\delta)}\Big).$$

**Remark 1.** We compare Theorem 1 with related work. Weak and strong PD generalization error bounds were established for (R)-ESP (Zhang et al., 2020). However, the discussion there does not consider the connection between stability and generalization. Primal generalization bounds were studied for stable algorithms (Farnia & Ozdaglar, 2020). However, the discussion there is not rigorous since they used an identity $nR_S(A_{\mathbf{w}}(S)) = \sum_{i=1}^n \max_{\mathbf{v}} f(A_{\mathbf{w}}(S), \mathbf{v}; z_i)$, which does not hold. To our best knowledge, Theorem 1 gives the first systematic connection between stability and generalization for minimax problems.

**Remark 2.** We provide some intuitive understanding of Theorem 1 here. Part (a) shows that weak-stability is sufficient for weak PD generalization. This is as expected since both the supremum over $\mathbf{w}'$ and $\mathbf{v}'$ are outside of the expectation operator in the definition of weak stability/generalization. We do not need to consider the correlation between $A_{\mathbf{w}}(S)$ and $\mathbf{v}'$. As a comparison, the primal generalization needs the much stronger argument-stability. The reason is that the supremum over $\mathbf{w}'$ is inside the expectation and $\mathbf{v}^{(i)} := \arg\sup_{\mathbf{v}'} F(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}')$ is different for different $i$ ($\mathbf{v}^{(i)}$ correlates to $A_{\mathbf{w}}(S^{(i)})$ and $S^{(i)}$ is derived by replacing the $i$-th example in $S$ with $z_i'$). We need to estimate how $\mathbf{v}^{(i)}$ differs from each other due to the difference among $A_{\mathbf{w}}(S^{(i)})$. This explains why we need argument-stability and a strong-concavity in Parts (b), (d) for primal generalization. Similarly, the strong PD generalization assumes SC-SC functions.

## 4. SGDA: Convex-Concave Case

In this section, we are interested in SGDA for solving minimax optimization problems in the convex-concave case. Let $\mathbf{w}_1 = 0 \in \mathcal{W}$ and $\mathbf{v}_1 = 0 \in \mathcal{V}$ be the initial point. Let $\text{Proj}_{\mathcal{W}}(\cdot)$ and $\text{Proj}_{\mathcal{V}}(\cdot)$ denote the projections onto $\mathcal{W}$ and $\mathcal{V}$, respectively. Let $\{\eta_t\}_t$ be a sequence of positive stepsizes. At each iteration, we randomly draw $i_t$ from the uniform distribution over $[n] := \{1, 2, \ldots, n\}$ and do the update

$$\begin{cases} \mathbf{w}_{t+1} = \text{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big), \\ \mathbf{v}_{t+1} = \text{Proj}_{\mathcal{V}}\big(\mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big). \end{cases} \tag{4.1}$$

### 4.1. Stability Bounds

In this section, we present the stability bounds for SGDA in the convex-concave case. We consider both the nonsmooth

setting and smooth setting. Part (a) and Part (b) establish stability bounds in expectation, while Part (c) and Part (d) give stability bounds with high probability. Part (e) consider the SC-SC case. The proofs are given in Appendix C.

**Theorem 2.** *Assume for all $z$, the function $(\mathbf{w},\mathbf{v}) \mapsto f(\mathbf{w},\mathbf{v};z)$ is convex-concave. Let the algorithm $A$ be SGDA (4.1) with $t$ iterations. Let $\delta \in (0,1)$.*

*(a) Assume $\eta_t = \eta$. If Assumption 1 holds, then $A$ is $4\eta G\big(\sqrt{t}+t/n\big)$-argument-stable in expectation.*

*(b) If Assumptions 1, 2 hold, then $A$ is $\epsilon$-argument-stable in expectation, where*

$$\epsilon \leq \frac{\sqrt{8e(1+t/n)}G}{\sqrt{n}} \exp\Big(2^{-1}L^2\sum_{j=1}^{t}\eta_j^2\Big)\Big(\sum_{k=1}^{t}\eta_k^2\Big)^{\frac{1}{2}}.$$

*(c) Let $\eta_t = \eta$. If Assumption 1 holds, then $A$ is $\epsilon$-argument-stable with probability at least $1-\delta$, where*

$$\epsilon \leq \sqrt{8e}G\eta\Big(\sqrt{t}+t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}\Big).$$

*(d) Let $\eta_t = \eta$. If Assumptions 1, 2 hold, then $A$ is $\epsilon$-argument-stable with probability at least $1-\delta$, where*

$$\epsilon \leq \sqrt{8e}G\eta\exp\big(2^{-1}L^2t\eta^2\big)\times \Big(1+t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}\Big).$$

*(e) If $(\mathbf{w},\mathbf{v}) \mapsto f(\mathbf{w},\mathbf{v};z)$ is $\rho$-SC-SC, Assumption 1 holds and $\eta_t = 1/(\rho t)$, then $A$ is $\epsilon$-argument-stable in expectation, where $\epsilon \leq \frac{2\sqrt{2}G}{\rho}\big(\frac{\log(et)}{t}+\frac{1}{n(n-2)}\big)^{\frac{1}{2}}$.*

**Remark 3.** If $t = O(n^2)$, then the stability bounds in a nonsmooth case (Part a) become $O(\eta\sqrt{t})$ and we can still get non-vacuous bounds by taking small step size $\eta = o(t^{-1/2})$. If we choose $\eta_j = 1/\sqrt{t}$ for $j \in [t]$, then the stability bound in Part (b) under a further smoothness assumption becomes $O(\sqrt{t}/n+n^{-\frac{1}{2}})$, which matches the existing result for SGD in a convex setting (Hardt et al., 2016). The high-probability bounds in Part (c) and Part (d) enjoy the same behavior.

**Remark 4.** The stability bounds of SGDA and GDA were discussed in Farnia & Ozdaglar (2020) for SC-SC, Lipschitz continuous and smooth problems. We remove the smoothness assumption in Part (e) in the SC-SC case. The stability of GDA was also studied there for convex-concave $f$, which, however, implies non-vanishing generalization bounds growing exponentially with the iteration count (Farnia & Ozdaglar, 2020). We extend their discussions to SGDA in this convex-concave case, and, as we will show in Theorem 3, our stability bounds imply optimal bounds on PD population risks. Furthermore, the existing discussions (Farnia & Ozdaglar, 2020) require the function $f$ to be smooth, while we show that meaningful stability bounds can be achieved in a nonsmooth setting (Parts (a), (c), (e)).

**Remark 5.** We consider stability bounds under various assumptions on loss functions. We now sketch the technical difference in our analysis. Let $\delta_t := \|\mathbf{w}_t-\mathbf{w}_t'\|_2^2+\|\mathbf{v}_t-\mathbf{v}_t'\|_2^2$, where $(\mathbf{w}_t,\mathbf{v}_t),(\mathbf{w}_t',\mathbf{v}_t')$ are SGDA iterates for $S$ and $S'$ differing only by the last element. For convex-concave and nonsmooth problems, we show $\delta_{t+1} = \delta_t + O(\eta_t^2)$ if $i_t \neq n$. For convex-concave and smooth problems, we show $\delta_{t+1} = (1+O(\eta_t^2))\delta_t$ if $i_t \neq n$. For $\rho$-SC-SC and nonsmooth problems, we show $\delta_{t+1} = (1-2\rho\eta_t)\delta_t + O(\eta_t^2)$ if $i_t \neq n$. For the above cases, we first control $\delta_{t+1}$ and then take expectation w.r.t. $i_t$. A key point to tackle *nonsmooth* problems is to consider the evolution $\delta_t$ instead of $\|\mathbf{w}_t-\mathbf{w}_t'\|_2+\|\mathbf{v}_t-\mathbf{v}_t'\|_2$, which is able to yield nontrivial bounds by making $\sum_t \eta_t^2 = o(1)$ with sufficiently small $\eta_t$.

### 4.2. Population Risks

We now use stability bounds in Theorem 2 to develop error bounds of SGDA which outputs an average of iterates

$$\bar{\mathbf{w}}_T = \frac{\sum_{t=1}^T \eta_t\mathbf{w}_t}{\sum_{t=1}^T\eta_t} \text{ and } \bar{\mathbf{v}}_T = \frac{\sum_{t=1}^T \eta_t\mathbf{v}_t}{\sum_{t=1}^T\eta_t}. \tag{4.2}$$

The underlying reason to introduce the average operator is to simplify the optimization error analysis (Nemirovski et al., 2009). Indeed, our stability and generalization analysis applies to any individual iterates. As a comparison, the optimization error analysis for the last iterate is much more difficult than that for the averaged iterate. We use the notation $B \asymp \widetilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1\widetilde{B} < B \leq c_2\widetilde{B}$. The following theorem to be proved in Appendix E gives weak PD population risk bounds.

**Theorem 3** (Weak PD risk)**.** *Let $\{\mathbf{w}_t,\mathbf{v}_t\}$ be produced by (4.1). Assume for all $z$, the function $(\mathbf{w},\mathbf{v}) \mapsto f(\mathbf{w},\mathbf{v};z)$ is convex-concave. Let $A$ be defined by $A_{\mathbf{w}}(S) = \bar{\mathbf{w}}_T$ and $A_{\mathbf{v}}(S) = \bar{\mathbf{v}}_T$ for $(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)$ in (4.2). Assume $\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{w}\|_2 \leq B_W$ and $\sup_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|_2 \leq B_V$.*

*(a) If $\eta_t = \eta$ and Assumption 1 holds, then*

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) \leq 4\sqrt{2}\eta G^2\Big(\sqrt{T}+\frac{T}{n}\Big)+\eta G^2 + \frac{B_W^2+B_V^2}{2\eta T}+\frac{G(B_W+B_V)}{\sqrt{T}}. \tag{4.3}$$

*If we choose $T \asymp n^2$ and $\eta \asymp T^{-3/4}$, then we get $\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) = O(n^{-1/2})$.*

*(b) If $\eta_t = \eta$ and Assumptions 1, 2 hold, then*

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) \leq \frac{4\sqrt{e(T+T^2/n)}G^2\eta\exp\big(LT\eta^2/2\big)}{\sqrt{n}} + \eta G^2 + \frac{B_W^2+B_V^2}{2\eta T}+\frac{G(B_W+B_V)}{\sqrt{T}}. \tag{4.4}$$

*We can choose* $T \asymp n$ *and* $\eta \asymp T^{-1/2}$ *to derive* $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(n^{-1/2})$.

*(c) If* $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ *is* $\rho$*-SC-SC* ($\rho > 0$)*, Assumption 1 holds,* $\eta_t = 1/(\rho t)$ *and* $T \asymp n^2$*, then*

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(\sqrt{\log n}/(n\rho)).$$

*(d) If* $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ *is* $\rho$*-SC-SC* ($\rho > 0$)*, Assumptions 1, 2 hold,* $\eta_t = 1/(\rho(t + t_0))$ *with* $t_0 \geq L^2/\rho^2$ *and* $T \asymp n$*, then* $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(\log(n)/(n\rho))$.

**Remark 6.** We first compare our bounds with the related work in a convex-concave setting. Weak PD population risk bounds were established for PPM under Assumptions 1, 2 (Farnia & Ozdaglar, 2020), which updates $(\mathbf{w}_{t+1}^{\text{PPM}}, \mathbf{v}_{t+1}^{\text{PPM}})$ as the saddle point of the following minimax problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}, \mathbf{v}) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t^{\text{PPM}}\|_2^2 + \frac{1}{2\eta_t}\|\mathbf{v} - \mathbf{v}_t^{\text{PPM}}\|_2^2.$$

In particular, they developed population risk bounds $O(1/\sqrt{n})$ by taking $T \asymp \sqrt{n}$ for PPM. However, the implementation of PPM requires to find the exact saddle point at each iteration, which is often computationally expensive. As a comparison, Part (b) shows the minimax optimal population risk bounds $O(1/\sqrt{n})$ for SGDA with $O(n)$ iterations. Weak PD population risk bounds $O(1/\sqrt{n})$ were established for R-ESP (Zhang et al., 2020) without a smoothness assumption, which, however, ignore the interplay between generalization and optimization. In this setting, we show SGDA achieves the same population risk bounds $O(1/\sqrt{n})$ by taking $\eta \asymp T^{-3/4}$ and $T \asymp n^2$ in Part (a). We now consider the SC-SC setting. Weak PD risk bounds $O(1/(n\rho))$ were established for ESP (Zhang et al., 2020). Since Farnia & Ozdaglar (2020) did not present an explicit risk bound, we use their stability analysis to give an explicit risk bound $O(\log(n)/(n\rho))$ in the smooth case (Part (d)). As a comparison, we establish the same population risk bounds for SGDA within a logarithmic factor by taking $\eta_t = 1/(\rho t)$ and $T \asymp n^2$ without the smoothness assumption (Part (c)).

We further develop bounds on primal population risks under a strong concavity assumption on $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$. Primal risk bounds measure the performance of primal variables, which are of real interest in some learning problems, e.g., AUC maximization and robust optimization. We consider both bounds in expectation and bounds with high probability. Let $(\mathbf{w}^*, \mathbf{v}^*)$ be a saddle point of $F$, i.e., for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{v} \in \mathcal{V}$, there holds $F(\mathbf{w}^*, \mathbf{v}) \leq F(\mathbf{w}, \mathbf{v}) \leq F(\mathbf{w}, \mathbf{v}^*)$.

**Theorem 4** (Excess primal risk)**.** *Let* $\{\mathbf{w}_t, \mathbf{v}_t\}$ *be produced by (4.1) with* $\eta_t = \eta$*. Assume for all* $z$*, the function* $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ *is convex-concave and the function* $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ *is* $\rho$*-strongly-concave. Assume* $\sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_2 \leq B_W$ *and* $\sup_{\mathbf{v}\in\mathcal{V}} \|\mathbf{v}\|_2 \leq B_V$*. Let the algorithm* $A$ *be defined by* $A_{\mathbf{w}}(S) = \bar{\mathbf{w}}_T$ *and* $A_{\mathbf{v}}(S) = \bar{\mathbf{v}}_T$ *for* $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ *in (4.2). If Assumptions 1, 2 hold, then*

$$\mathbb{E}\big[R(\bar{\mathbf{w}}_T)\big] - \inf_{\mathbf{w}} R(\mathbf{w}) \leq \eta G^2 + \frac{B_W^2 + B_V^2}{2\eta T} + \frac{G(B_W + B_V)}{\sqrt{T}} + \frac{(1 + L/\rho)\sqrt{32e(T + T^2/n)}G^2\eta \exp\left(L^2 T\eta^2/2\right)}{\sqrt{n}}.$$

*In particular, if we choose* $T \asymp n$ *and* $\eta \asymp T^{-1/2}$ *then*

$$\mathbb{E}\big[R(\bar{\mathbf{w}}_T)\big] - \inf_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}) = O((L/\rho)n^{-1/2}). \quad (4.5)$$

*Furthermore, for any* $\delta \in (0, 1)$ *we can choose* $T \asymp n$ *and* $\eta \asymp T^{-1/2}$ *to show with probability at least* $1 - \delta$

$$R(\bar{\mathbf{w}}_T) - R(\mathbf{w}^*) = O\Big((L/\rho)n^{-\frac{1}{2}} \log n \log^2(1/\delta)\Big). \quad (4.6)$$

Theorem 4 is proved in Appendix E. In Theorem E.1, we also develop high-probability bounds of order $O(n^{-\frac{1}{2}} \log n)$ on plain generalization errors $|F(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \mathbf{v}^*)|$.

## 5. Nonconvex-Nonconcave Objectives

In this section, we extend our analysis to nonconvex-nonconcave minimax learning problems.

### 5.1. Stability and Generalization of SGDA

We first study the generalization bounds of SGDA for WC-WC problems. The proof is given in Appendix F.1. In Appendix F.2, we further give high-probability bounds.

**Theorem 5** (Weak generalization bound)**.** *Let* $\{\mathbf{w}_t, \mathbf{v}_t\}$ *be produced by (4.1) with* $T$ *iterations. Assume for all* $z$*, the function* $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ *is* $\rho$*-WC-WC and* $|f(\cdot, \cdot; z)| \leq 1$*. If Assumption 1 holds and* $\eta_t = c/t$*, then the weak PD generalization error of SGDA is bounded by*

$$O\Big(\Big(1 + \frac{\sqrt{T}}{n}\Big)T^{c\rho}\Big)^{\frac{2}{2c\rho+3}}\Big(\frac{1}{n}\Big)^{\frac{2c\rho+1}{2c\rho+3}}.$$

**Remark 7.** If $T = O(n^2)$, our weak PD generalization error bound is of the order $O\big(n^{-\frac{2c\rho+1}{2c\rho+3}}T^{\frac{2c\rho}{2c\rho+3}}\big)$. This is the first generalization bound of SGDA for nonsmooth and nonconvex-nonconcave objectives. Farnia & Ozdaglar (2020) also studied generalization under nonconvex-nonconcave setting but required the objectives to be smooth, which is relaxed to a milder WC-WC assumption here. Our analysis readily applies to stochastic gradient descent (SGD) with nonsmooth weakly-convex functions, which has not been studied in the literature.

We further consider a variant of weak-convexity-weak-concavity. The proof is given in Appendix F.3.

**Theorem 6** (Weak generalization bound). *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be produced by (4.1) with $T$ iterations. Let Assumptions 1, 2 hold. Assume there are non-negative numbers $\{\rho_t\}_{t\in\mathbb{N}}$ such that the following inequality holds a.s.*

$$\left\langle \begin{pmatrix} \mathbf{w}_t - \mathbf{w} \\ \mathbf{v}_t - \mathbf{v} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v}) \\ \nabla_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v}) - \nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t) \end{pmatrix} \right\rangle \geq -\rho_t \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w} \\ \mathbf{v}_t - \mathbf{v} \end{pmatrix} \right\|_2^2, \ \forall \mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}. \quad (5.1)$$

*Then the weak PD generalization error of SGDA with $T$ iterations can be bounded by*

$$O\Big(n^{-1} \sum_{t=1}^{T} \Big(\eta_t^2 + \frac{1}{n}\Big) \exp\Big( \sum_{k=t+1}^{T} \big(2\rho_k \eta_k + (L^2+1)\eta_k^2\big)\Big)\Big)^{\frac{1}{2}}.$$

Eq. (5.1) allows the empirical objective $F_S$ to have varying weak-convexity-weak-concavity at different iterates encountered by the algorithm. This is motivated by the observation that the nonconvex-nonconcave function can have approximate convexity-concavity around a saddle point. For these problems, we can expect the weak-convexity-weak-concavity parameter $\rho_t$ to decrease along the optimization process (Sagun et al., 2017; Yuan et al., 2019).

**Remark 8.** If $F_S$ is convex-concave, then $\rho_t = 0$ and we can take $\eta_t \asymp 1/\sqrt{T}$ to show that SGDA with $T$ iterations enjoys the generalization bound $O\big(1/\sqrt{n} + \sqrt{T}/n\big)$. This extends Theorem 3 since we only require the convexity-concavity of $F_S$ here instead of $f(\cdot, \cdot; z)$ for all $z$ in Theorem 3. If $\rho_t = O(t^{-\alpha})$ ($\alpha \in (0,1)$), then we can take $\eta_t \asymp t^{\min\{\alpha-1, -\frac{1}{2}\}}/\log T$ (note $\sum_{t=1}^{T} \eta_t^2 = O(1), \sum_{t=1}^{T} \eta_t \rho_t = O(1)$) to show that SGDA with $T$ iterations enjoys the weak PD generalization bound $O\big(1/\sqrt{n} + \sqrt{T}/n\big)$. As compared to Theorem 5, the assumption (5.1) allows us to use much larger step sizes ($O(t^{-\beta}), \beta \in (0,1)$ vs $O(t^{-1})$). This larger step size allows for a better trade-off between generalization and optimization. We note that a recent work (Richards & Rabbat, 2021) considered gradient descent under an assumption similar to (5.1), and developed interesting generalization bounds for $\eta_t = O(t^{-\beta})$ ($\beta \in (0,1)$). However, their discussions do not apply to the important SGD and require an additional assumption on the Lipschitz continuity of Hessian matrix which may be restrictive. It is direct to extend Theorem 6 to SGD for learning with weakly-convex functions for relaxing the step size under Eq. (5.1). Therefore, our stability analysis even gives novel results in the standard nonconvex learning setting. We introduce a novel technique in achieving this improvement. Specifically, let $\delta_t := \|\mathbf{w}_t - \mathbf{w}_t'\|_2^2 + \|\mathbf{v}_t - \mathbf{v}_t'\|_2^2$, where $(\mathbf{w}_t, \mathbf{v}_t), (\mathbf{w}_t', \mathbf{v}_t')$ are SGDA iterates for neighboring datasets $S$ and $S'$. For the stability bounds in Section 4.1, we first handle $\delta_{t+1}$ according to different realizations of $i_t$ and then consider the expectation w.r.t. $i_t$. While for $\rho$-WC-WC problems, we first take expectation w.r.t. $i_t$ and then show how $\mathbb{E}_{i_t}[\delta_{t+1}]$ would change along the iterations.

### 5.2. Stability and Generalization of AGDA and Beyond

We now study the Alternating Gradient Descent Ascent (AGDA) proposed recently to optimize nonconvex-nonconcave problems (Yang et al., 2020). Let $\{\eta_{\mathbf{w},t}, \eta_{\mathbf{v},t}\}_t$ be a sequence of positive stepsizes for updating $\{\mathbf{w}_t, \mathbf{v}_t\}_t$. At each iteration, we randomly draw $i_t$ and $j_t$ from the uniform distribution over $[n]$ and do the update

$$\begin{cases} \mathbf{w}_{t+1} = \mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_{\mathbf{w},t} \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big), \\ \mathbf{v}_{t+1} = \mathrm{Proj}_{\mathcal{V}}\big(\mathbf{v}_t + \eta_{\mathbf{v},t} \nabla_{\mathbf{v}} f(\mathbf{w}_{t+1}, \mathbf{v}_t; z_{j_t})\big). \end{cases} \quad (5.2)$$

This algorithm differs from SGDA in two aspects. First, it randomly selects two examples to update $\mathbf{w}$ and $\mathbf{v}$ per iteration. Second, it uses the updated $\mathbf{w}_{t+1}$ when updating $\mathbf{v}_{t+1}$. Theorem 7 to be proved in Appendix G provides generalization bounds for AGDA.

**Theorem 7** (Weak generalization bounds). *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be the sequence produced by (5.2). If Assumptions 1, 2 hold and $\eta_{\mathbf{w},t} + \eta_{\mathbf{v},t} \leq \frac{c}{t}$ for some $c > 0$, then the weak PD generalization error can be upper bounded by $O\big(n^{-1} T^{\frac{cL}{cL+1}}\big)$.*

Global convergence of AGDA was studied based on the two-sided PL condition defined below (Yang et al., 2020), which means the suboptimality of function values can be bounded by gradients and were shown for several rich classes of functions (Karimi et al., 2016). We also refer to the two-sided PL condition as the gradient dominance condition.

**Assumption 3.** Assume $F_S$ satisfies the two-sided PL condition, i.e., there exist constants $\beta_1(S), \beta_2(S) > 0$ such that the following inequalities hold for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$

$$2\beta_1(S)\big(F_S(\mathbf{w}, \mathbf{v}) - \inf_{\mathbf{w}' \in \mathcal{W}} F_S(\mathbf{w}', \mathbf{v})\big) \leq \|\nabla_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v})\|_2^2,$$

$$2\beta_2(S)\big(\sup_{\mathbf{v}' \in \mathcal{V}} F_S(\mathbf{w}, \mathbf{v}') - F_S(\mathbf{w}, \mathbf{v})\big) \leq \|\nabla_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v})\|_2^2.$$

As a combination of our generalization bounds and optimization error bounds in Yang et al. (2020), we can derive the following informal corollary on primal population risks by early stopping the algorithm to balance the optimization and generalization. It gives the first primal risk bounds for learning with nonconvex-strongly-concave functions. The precise statement can be found in Corollary G.3.

**Corollary 8** (Informal). *Let $\beta_1, \rho > 0$. Assume for all $\mathbf{w}$, the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is $\rho$-strongly concave. Let Assumptions 1, 2, 3 with $\beta_1(S) \geq \beta_1, \beta_2(S) \geq \rho$ hold. Then AGDA with some appropriate step size and $T \asymp \big(n\beta_1^{-2}\rho^{-3}\big)^{\frac{cL+1}{2cL+1}}$ satisfy ($c \asymp 1/(\beta_1\rho^2)$)*

$$\mathbb{E}\big[R(\mathbf{w}_T)\big] - R(\mathbf{w}^*) = O\Big(n^{-\frac{cL+1}{2cL+1}} \beta_1^{-\frac{2cL}{2cL+1}} \rho^{-\frac{5cL+1}{2cL+1}}\Big).$$

For gradient dominated problems, we further have the following error bounds to be proved in Appendix H. Note we do not need the smoothness assumption here.

**Theorem 9.** *Let $A$ be an algorithm. Let Assumptions 1, 3 hold. Let $\mathbf{u}_S = (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ and $\mathbf{u}_S^{(S)}$ be the projection of $\mathbf{u}_S$ onto the set of stationary points of $F_S$. Then,*

$$\big|\mathbb{E}\big[F(\mathbf{u}_S) - F_S(\mathbf{u}_S)\big]\big| \leq \frac{2G^2}{n} \max\Big\{\mathbb{E}\big[1/\beta_1(S)\big], \mathbb{E}\big[1/\beta_2(S)\big]\Big\} + 2G\mathbb{E}\big[\|\mathbf{u}_S - \mathbf{u}_S^{(S)}\|_2\big].$$

**Remark 9.** Note $\|\mathbf{u}_S - \mathbf{u}_S^{(S)}\|_2$ measures how far the point found by $A$ is from the set of stationary points of $F_S$, and can be interpreted as an optimization error. Therefore, Theorem 9 gives a connection between generalization error and optimization error. For a variant of AGDA with noiseless stochastic gradients, it was shown that $\|(\mathbf{w}_T, \mathbf{v}_T) - (\mathbf{w}_T, \mathbf{v}_T)^{(S)}\|_2$ decays linearly w.r.t. $T$ (Yang et al., 2020). We can plug this linear convergent optimization bound into Theorem 9 to directly get generalization bounds. If $A$ returns a saddle point of $F_S$, then $\|\mathbf{u}_S - \mathbf{u}_S^{(S)}\|_2 = 0$ and therefore $\big|\mathbb{E}\big[F(\mathbf{u}_S) - F_S(\mathbf{u}_S)\big]\big| = O\big(n^{-1}\max\{\mathbb{E}[1/\beta_1(S)], \mathbb{E}[1/\beta_2(S)]\}\big)$. Generalization errors of this particular ESP were studied in Zhang et al. (2020) for SC-SC minimax problems, which were extended to more general gradient-dominated problems in Theorem 9. Furthermore, Theorem 9 applies to any optimization algorithm instead of the specific ESP. It should be mentioned that Zhang et al. (2020) addressed PD population risks, while we consider plain generalization errors.

## 6. Experiments

In this subsection, we report preliminary experimental results to validate our theoretical results[2]. We consider two datasets available at the LIBSVM website: `svmguide3` and `w5a` (Chang & Lin, 2011). We follow the experimental setup in Hardt et al. (2016) to study how the stability of SGDA would behave along the learning process. To this end, we build a neighboring dataset $S'$ by changing the last example of the training set $S$. We apply the same randomized algorithm to $S$ and $S'$ and get two model sequences $\{(\mathbf{w}_t, \mathbf{v}_t)\}$ and $\{(\mathbf{w}_t', \mathbf{v}_t')\}$. Then we evaluate the Euclidean distance $\triangle_t = \big(\|\mathbf{w}_t - \mathbf{w}_t'\|_2^2 + \|\mathbf{v}_t - \mathbf{v}_t'\|_2^2\big)^{\frac{1}{2}}$. We consider the `SOLAM` (Ying et al., 2016) algorithm, which is the SGDA for the solving the problem (2.3) (a minimax reformulation of the AUC maximization problem). We consider step sizes $\eta_t = \eta/\sqrt{T}$ with $\eta \in \{0.1, 0.3, 1, 3\}$. We repeat the experiments 25 times and report the average of the experimental results as well as the standard deviation. In Figure 1, we report $\triangle_t$ as a function of the number of passes (the iteration number divided by $n$). It is clear that the Euclidean distance continues to increase during the learning process. Furthermore, the Euclidean distances increase if we consider larger step sizes. This phenomenon is consistent with our stability bounds in Theorem 2. More experiments on the stability of SGDA in GAN training can be found in Appendix I.

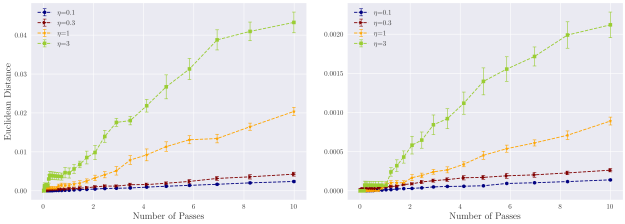


*Figure 1.* $\triangle_t$ versus the number of passes on `svmguide3` (left) and `w5a` (right).

## 7. Conclusion

We presented a comprehensive stability and generalization analysis of stochastic algorithms for minimax objective functions. We introduced various generalization measures and stability measures, and present a systematic study on their quantitative relationship. In particular, we obtained the first minimax optimal risk bounds for SGDA in a general convex-concave case, covering both smooth and nonsmooth setting. We also derived the first non-trivial risk bounds for nonconvex-nonconcave problems. Our bounds show how to early-stop the algorithm in practice to train a model with better generalization. Our theoretical results have potential applications in developing differentially private algorithms to handle sensitive data.

There are some interesting problems for further investigation. Our primal generalization bounds require a strong concavity assumption. It is interesting to remove this assumption. On the other front, it remains an open question to us on understanding how the concavity of dual variables can help generalization in a nonconvex setting.

## Acknowledgments

We are grateful to the anonymous reviewers for their constructive comments and suggestions. The work of Yiming Ying is partially supported by NSF grants IIS-1816227 and IIS-2008532. The work of Tianbao Yang is partially supported by NSF #1933212 and NSF CAREER Award #1844403.

## References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

[2]The source codes are available at `https://github.com/zhenhuan-yang/minimax-stability`.

Audibert, J.-Y. and Catoni, O. Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794, 2011.

Balamurugan, P. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *Advance In Neural Information Processing Systems*, pp. 1416–1424, 2016.

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

Bousquet, O. and Bottou, L. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pp. 161–168, 2008.

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

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626, 2020.

Brownlees, C., Joly, E., Lugosi, G., et al. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

Charles, Z. and Papailiopoulos, D. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pp. 744–753, 2018.

Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pp. 4705–4714, 2017.

Chen, Y., Jin, C., and Yu, B. Stability and convergence tradeoff of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

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

Deng, Z., He, H., and Su, W. J. Toward better generalization bounds with locally elastic stability. *arXiv preprint arXiv:2010.13988*, 2020.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058, 2017.

Elisseeff, A., Evgeniou, T., and Pontil, M. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

Farnia, F. and Ozdaglar, A. Train simultaneously, generalize better: Stability of gradient-based minimax learners. *arXiv preprint arXiv:2010.12561*, 2020.

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

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, pp. 6726–6736. 2019.

Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. One-pass auc optimization. In *International Conference on Machine Learning*, pp. 906–914, 2013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

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

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pp. 6938–6948, 2019.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning*, pp. 795–811, 2016.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2820–2829, 2018.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

Lei, Y. and Ying, Y. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021a.

Lei, Y. and Ying, Y. Stochastic proximal auc maximization. *Journal of Machine Learning Research*, 22(61): 1–45, 2021b.

Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

Lin, J., Camoriano, R., and Rosasco, L. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pp. 2340–2348, 2016.

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

Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic AUC maximization with O(1/n)-convergence rate. In *International Conference on Machine Learning*, pp. 3195–3203, 2018.

Liu, M., Rafique, H., Lin, Q., and Yang, T. First-order convergence theory for weakly-convex-weakly-concave min-max problems, 2020.

Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pp. 2159–2167, 2017.

Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pp. 6370–6381, 2020.

London, B. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2931–2940, 2017.

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

Madden, L., Dall'Anese, E., and Becker, S. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pp. 605–638, 2018.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pp. 2971–2980, 2017.

Nedić, A. and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Raghavan, P. Probabilistic construction of deterministic algorithms: approximating packing integer programs. *Journal of Computer and System Sciences*, 37(2):130–143, 1988.

Rakhlin, A., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

Richards, D. and Rabbat, M. Learning with gradient descent and weakly convex losses. *arXiv preprint arXiv:2101.04968*, 2021.

Richards, D. et al. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(2020), 2020.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

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

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

Vershynin, R. *High-dimensional probability: An introduc-*

*tion with applications in data science*. Cambridge university press, 2018.

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R., and Yang, T. Learning with non-convex truncated losses by sgd. In *Uncertainty in Artificial Intelligence*, pp. 701–711, 2020.

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

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

Ying, Y., Wen, L., and Lyu, S. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.

Yuan, Z., Yan, Y., Jin, R., and Yang, T. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, pp. 2604–2614, 2019.

Zhang, J., Hong, M., Wang, M., and Zhang, S. Generalization bounds for stochastic saddle point problems. *arXiv preprint arXiv:2006.02067*, 2020.

Zhao, P., Hoi, S. C., Jin, R., and Yang, T. Online AUC maximization. In *International Conference on Machine Learning*, pp. 233–240, 2011.