Train simultaneously, generalize better: Stability of gradient-based minimax learners

Farzan Farnia [1] **Asuman Ozdaglar** [1]

Abstract

The success of minimax learning problems of generative adversarial networks (GANs) has been observed to depend on the minimax optimization algorithm used for their training. This dependence is commonly attributed to the convergence speed and robustness properties of the underlying optimization algorithm. In this paper, we show that the optimization algorithm also plays a key role in the *generalization performance* of the trained minimax model. To this end, we analyze the generalization properties of standard gradient descent ascent (GDA) and proximal point method (PPM) algorithms through the lens of algorithmic stability as defined by Bousquet & Elisseeff, 2002 under both convex concave and non-convex non-concave minimax settings. While the GDA algorithm is not guaranteed to have a vanishing excess risk in convex concave problems, we show the PPM algorithm enjoys a bounded excess risk in the same setup. For non-convex non-concave problems, we compare the generalization performance of stochastic GDA and GDmax algorithms where the latter fully solves the maximization subproblem at every iteration. Our generalization analysis suggests the superiority of GDA provided that the minimization and maximization subproblems are solved simultaneously with similar learning rates. We discuss several numerical results indicating the role of optimization algorithms in the generalization of learned minimax models.

1. Introduction

In recent years, minimax learning frameworks including generative adversarial networks (GANs) (Goodfellow et al., 2014) and adversarial training (Madry et al., 2017) have achieved great success over a wide array of learning tasks. In these approaches, the learning problem is modeled as a zero-sum game between two "min" and "max" players that is commonly solved by a minimax optimization algorithm. The minimax optimization problem of these learning frameworks is typically formulated using deep neural networks, which greatly complicates the theoretical and numerical analysis of the optimization problem. Current studies in the machine learning literature focus on fundamental understanding of general minimax problems with emphasis on convergence speed and optimality.

The primary focus of optimization-related studies of minimax learning problems has been on the convergence speed and robustness properties of minimax optimization algorithms. Several recently proposed algorithms have been shown to achieve faster convergence rates and more robust behavior around local solutions. However, training speed and robustness are not the only factors required for the success of a minimax optimization algorithm in a learning task. In this work, we aim to show that the *generalization performance* of the learned minimax model is another key property that is influenced by the underlying optimization algorithm. To this end, we present theoretical and numerical results demonstrating that:

Different minimax optimization algorithms can learn models with different generalization properties.

In order to analyze the generalization behavior of minimax optimization algorithms, we use the algorithmic stability framework as defined by (Bousquet & Elisseeff, 2002) for general learning problems and applied by (Hardt et al., 2016) for analyzing stochastic gradient descent. Our extension of (Bousquet & Elisseeff, 2002)'s stability approach to minimax settings allows us to analyze and compare the generalization properties of standard gradient descent ascent (GDA) and proximal point method (PPM) algorithms. Furthermore, we compare the generalization performance between the following two types of minimax optimization algorithms: 1) simultaneous optimization algorithms such as GDA where the minimization and maximization subproblems are solved simultaneously, and 2) non-simultaneous optimization algo-

[1]Laboratory for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence to: Farzan Farnia <farnia@mit.edu>, Asuman Ozdaglar <asuman@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

rithms such as GDmax where the maximization variable is fully optimized at every iteration.

In our generalization analysis, we consider both the traditional convex concave and general non-convex non-concave classes of minimax optimization problems. For convex concave minimax problems, our bounds indicate a similar generalization performance for simultaneous and non-simultaneous optimization methods. Specifically, we show for strongly-convex strongly-concave minimax problems all the discussed algorithms have a bounded generalization error on the order of $O(1/n)$ with n denoting the number of training samples. However, in general convex concave problems we show that the GDA algorithm with a constant learning rate is not guaranteed to have a properly bounded generalization risk. On the other hand, we prove that proximal point methods still achieve a controlled generalization error, resulting in a vanishing $O(\sqrt{1/n})$ excess risk with respect to the best minimax learner with the optimal performance on the underlying distribution.

For more general minimax problems, our results indicate that models trained by simultaneous and non-simultaneous algorithms can reach different generalization performances. Specifically, we consider the class of non-convex strongly-concave problems where we establish stability-based generalization bounds for both stochastic GDA and GDmax algorithms. Our generalization bounds indicate that the stochastic GDA learner is expected to generalize better provided that the min and max players are trained simultaneously with similar learning rates. In addition, we show a generalization bound for the stochastic GDA algorithm in general non-convex non-concave problems, which further supports the simultaneous optimization of the two min and max players in general minimax settings. Our results indicate that simultaneous training of the two players not only can provide a faster training, but also can learn a model with better generalization performance. Our generalization analysis, therefore, revisits the notion of *implicit competitive regularization* introduced by (Schäfer et al., 2019) for simultaneous gradient methods in training GANs.

Finally, we discuss the results of our numerical experiments and compare the generalization performance of GDA and PPM algorithms in convex concave settings and single-step and multi-step gradient-based methods in non-convex non-concave GAN problems. Our numerical results also suggest that in general non-convex non-concave problems the models learned by simultaneous optimization algorithms can generalize better than the models learned by non-simultaneous optimization methods. We can summarize the main contributions of this paper as follows:

- Extending the algorithmic stability framework for analyzing generalization in minimax settings,
- Analyzing the generalization properties of minimax models learned by GDA and PPM algorithms in convex concave problems,
- Studying the generalization of stochastic GDA and GDmax learners in non-convex non-concave problems,
- Providing numerical results on the role of optimization algorithms in the generalization performance of learned minimax models.

2. Related Work

Generalization in GANs: Several related papers have studied the generalization properties of GANs. Arora et al. (2017) study the generalization behavior of GANs' learned models and prove a uniform convergence generalization bound in terms of the number of the discriminator's parameters. Wu et al. (2019) connect the algorithmic stability notion to differential privacy in GANs and numerically analyze the generalization behavior of GANs. References (Zhang et al., 2017; Bai et al., 2018) show uniform convergence bounds for GANs by analyzing the Rademacher complexity of the players. Feizi et al. (2020) provide a uniform convergence bound for the W2GAN problem. Unlike the mentioned related papers, our work provides algorithm-dependent generalization bounds by analyzing the stability of gradient-based optimization algorithms. Also, the related works (Arora & Zhang, 2017; Thanh-Tung et al., 2019) conduct empirical studies of generalization in GANs using birthday paradox-based and gradient penalty-based approaches, respectively.

Generalization in adversarial training: Understanding generalization in the context of adversarial training has recently received great attention. Schmidt et al. (2018) show that in a simplified Gaussian setting generalization in adversarial training requires more training samples than standard non-adversarial learning. Farnia et al. (2018); Yin et al. (2019); Khim & Loh (2018); Wei & Ma (2019); Attias et al. (2019) prove uniform convergence generalization bounds for adversarial training schemes through Pac-Bayes (McAllester, 1999; Neyshabur et al., 2017b), Rademacher analysis, margin-based, and VC analysis approaches. Zhai et al. (2019) study the value of unlabeled samples in obtaining a better generalization performance in adversarial training. We note that unlike our work the generalization analyses in the mentioned papers prove uniform convergence results. In another related work, Rice et al. (2020) empirically study the generalization performance of adversarially-trained models and suggest that the generalization behavior can significantly change during training.

Stability-based generalization analysis: Algorithmic stability and its connections to the generalization properties of learning algorithms have been studied in several related

works. As a concurrent work, Lei et al. (2021) analyze the generalization of stochastic gradient methods for solving min-max optimization problems through the algorithmic stability approach. In a related paper, Zhang et al. (2020c) analyze the generalization of saddle points in strongly-concave-concave problems using a stability-based approach. On the other hand, our generalization bounds are algorithm-dependent and apply to convex-concave and nonconvex-nonconcave settings. Shalev-Shwartz et al. (2010) discuss learning problems where learnability is feasible considering algorithmic stability, while it is infeasible with uniform convergence. Hardt et al. (2016) bound the generalization risk of the stochastic gradient descent learner by analyzing its algorithmic stability. Feldman & Vondrak (2018; 2019); Bousquet et al. (2020) provide sharper stability-based generalization bounds for standard learning problems. While the above works focus on standard learning problems with a single learner, we use algorithmic stability to analyze generalization in minimax settings with two players.

Connections between generalization and optimization in deep learning: The connections between generalization and optimization in deep learning have been studied in several related works. Analyzing the double descent phenomenon (Belkin et al., 2019; Nakkiran et al., 2019; Mei & Montanari, 2019), the effect of overparameterization on generalization (Allen-Zhu et al., 2019; Arora et al., 2019; Li & Liang, 2018), and the sharpness of local minima (Keskar et al., 2016; Dinh et al., 2017; Neyshabur et al., 2017a) have been performed in the literature to understand the implicit regularization of gradient methods in deep learning (Neyshabur et al., 2014; Zhang et al., 2016; Ma et al., 2018; Chatterjee, 2020). Schäfer et al. (2019) extend the notion of implicit regularization to simultaneous gradient methods in GAN settings and discuss an optimization-based perspective to this regularization mechanism. However, we focus on the generalization aspect of the implicit regularization mechanism. Also, Nagarajan & Kolter (2019) suggest that uniform convergence bounds may be unable to explain generalization in supervised deep learning.

Analyzing convergence and stability of minimax optimization algorithms: A large body of related papers (Heusel et al., 2017; Sanjabi et al., 2018; Lin et al., 2019; Fiez et al., 2019; Nouiehed et al., 2019; Hsieh et al., 2019; Du & Hu, 2019; Mazumdar et al., 2019; Thekumparampil et al., 2019; Mazumdar et al., 2020; Zhang et al., 2020b) study convergence properties of first-order and second-order minimax optimization algorithms. Also, the related works (Daskalakis et al., 2017; Gidel et al., 2018; Liang & Stokes, 2019; Mokhtari et al., 2020; Zhang & Wang, 2020; Zhang et al., 2020a) analyze the convergence behavior of optimistic methods and extra gradient (EG) methods as approximations of the proximal point method.

3. Preliminaries

In this paper, we focus on two standard families of minimax optimization algorithms: Gradient Descent Ascent (GDA) and Proximal Point Method (PPM). To review the update rules of these algorithms, consider the following minimax optimization problem for minimax objective $f(\mathbf{w}, \boldsymbol{\theta})$ and feasible sets $\mathcal{W}, \Theta$:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\boldsymbol{\theta}\in\Theta} f(\mathbf{w}, \boldsymbol{\theta}). \tag{1}$$

Then, for stepsize values α_w, α_θ, the followings are the GDA's and GDmax's update rules:

$$\begin{aligned} G_{\text{GDA}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) &:= \begin{bmatrix}\mathbf{w} - \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta})\\ \boldsymbol{\theta} + \alpha_\theta \nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta})\end{bmatrix},\\ G_{\text{GDmax}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) &:= \begin{bmatrix}\mathbf{w} - \alpha_w \nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta})\\ \operatorname{argmax}_{\widetilde{\boldsymbol{\theta}}\in\Theta} f(\mathbf{w}, \widetilde{\boldsymbol{\theta}})\end{bmatrix} \end{aligned} \tag{2}$$

In the above, $\operatorname{argmax}_{\theta\in\Theta} f(\mathbf{w}, \boldsymbol{\theta})$ is the optimal maximizer for $\mathbf{w}$. Also, given stepsize parameter η the update rule of PPM is as follows:

$$\begin{aligned} G_{\text{PPM}}(\begin{bmatrix}\mathbf{w}\\ \boldsymbol{\theta}\end{bmatrix}) := \operatorname*{argmin}_{\widetilde{\mathbf{w}}\in\mathcal{W}} \operatorname*{argmax}_{\widetilde{\boldsymbol{\theta}}\in\Theta} \big\{ & f(\widetilde{\mathbf{w}}, \widetilde{\boldsymbol{\theta}}) \\ & + \frac{1}{2\eta}\|\widetilde{\mathbf{w}} - \mathbf{w}\|_2^2 - \frac{1}{2\eta}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \big\}, \end{aligned} \tag{3}$$

In the Appendix, we also consider and analyze the PPmax algorithm that is a proximal point method fully solving the maximization subproblem at every iteration. Throughout the paper, we commonly use the following assumptions on the Lipschitzness and smoothness of the minimax objective.

Assumption 1. *$f(\mathbf{w}, \boldsymbol{\theta})$ is jointly L-Lipschitz in $(\mathbf{w}, \boldsymbol{\theta})$ and L_w-Lipschitz in $\mathbf{w}$ over $\mathcal{W} \times \Theta$, i.e., for every $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ we have*

$$\begin{aligned} &\big|f(\mathbf{w}, \boldsymbol{\theta}) - f(\mathbf{w}', \boldsymbol{\theta}')\big| \le L\sqrt{\|\mathbf{w} - \mathbf{w}'\|_2^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2},\\ &\big|f(\mathbf{w}, \boldsymbol{\theta}) - f(\mathbf{w}', \boldsymbol{\theta})\big| \le L_w\|\mathbf{w} - \mathbf{w}'\|_2. \end{aligned}$$

Assumption 2. *$f(\mathbf{w}, \boldsymbol{\theta})$ is continuously differentiable and ℓ-smooth on $\mathcal{W} \times \Theta$, i.e., $\big[\nabla_{\mathbf{w}} f(\mathbf{w}, \boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} f(\mathbf{w}, \boldsymbol{\theta})\big]$ is ℓ-Lipschitz on $\mathcal{W} \times \Theta$.*

We focus on several classes of minimax optimization problems based on the convexity properties of the objective function. Note that a differentiable function $g(\mathbf{u})$ is called convex in $\mathbf{u}$ if it satisfies the following inequality for every $\mathbf{u}_1, \mathbf{u}_2$:

$$g(\mathbf{u}_2) \ge g(\mathbf{u}_1) + \nabla g(\mathbf{u}_1)^\top (\mathbf{u}_2 - \mathbf{u}_1).$$

Furthermore, g is called μ-strongly-convex if for every $\mathbf{u}_1, \mathbf{u}_2$ it satisfies

$$g(\mathbf{u}_2) \ge g(\mathbf{u}_1) + \nabla g(\mathbf{u}_1)^\top (\mathbf{u}_2 - \mathbf{u}_1) + \frac{\mu}{2}\|\mathbf{u}_2 - \mathbf{u}_1\|_2^2.$$

Also, g is called concave and μ-strongly-concave if $-g$ is convex and μ-strongly-convex, respectively.

Definition 1. *Consider convex feasible sets $\mathcal{W}, \Theta$ in minimax problem (1). Then,*

- *The problem is called convex concave if $f(\cdot, \boldsymbol{\theta})$ and $f(\mathbf{w}, \cdot)$ are respectively convex and concave functions for every $\mathbf{w}, \boldsymbol{\theta}$.*
- *The problem is called μ-strongly-convex strongly-concave if $f(\cdot, \boldsymbol{\theta})$ and $f(\mathbf{w}, \cdot)$ are respectively μ-strongly-convex and μ-strongly-concave functions for every $\mathbf{w}, \boldsymbol{\theta}$.*
- *The problem is called non-convex μ-strongly-concave if $f(\mathbf{w}, \cdot)$ is μ-strongly-concave for every $\mathbf{w}$.*

4. Stability-based Generalization Analysis in Minimax Settings

Consider the following optimization problem for a minimax learning task:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\boldsymbol{\theta}\in\Theta} R(\mathbf{w}, \boldsymbol{\theta}) := \mathbb{E}_{\mathbf{Z}\sim P_Z}\big[f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{Z})\big] \tag{4}$$

The above minimax objective represents a cost function $f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{Z})$ for minimization and maximization variables $\mathbf{w}, \boldsymbol{\theta}$ and data variable $\mathbf{Z}$ that is averaged under the underlying distribution $P_{\mathbf{Z}}$. We call the objective function $R(\mathbf{w}, \boldsymbol{\theta})$ the true minimax risk. We also define $R(\mathbf{w})$ as the worst-case minimax risk over the maximization variable $\boldsymbol{\theta}$:

$$R(\mathbf{w}) := \max_{\boldsymbol{\theta}\in\Theta} R(\mathbf{w}, \boldsymbol{\theta}) \tag{5}$$

In the context of GANs, the worst-case risk $R(\mathbf{w})$ represents a divergence measure between the learned and true distributions, and in the context of adversarial training it represents the learner's risk under adversarial perturbations. Since the learner does not have access to the underlying distribution $P_{\mathbf{Z}}$, we estimate the minimax objective using the empirical samples in dataset $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ which are drawn according to $P_{\mathbf{Z}}$. We define the empirical minimax risk as:

$$R_S(\mathbf{w}, \boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z}_i). \tag{6}$$

Then, the worst-case empirical risk over the maximization variable $\boldsymbol{\theta}$ is defined as

$$R_S(\mathbf{w}) := \max_{\boldsymbol{\theta}\in\Theta} R_S(\mathbf{w}, \boldsymbol{\theta}). \tag{7}$$

We define the minimax generalization risk $\epsilon_{\text{gen}}(\mathbf{w})$ of minimization variable $\mathbf{w}$ as the difference between the worst-case true and empirical risks:

$$\epsilon_{\text{gen}}(\mathbf{w}) := R(\mathbf{w}) - R_S(\mathbf{w}). \tag{8}$$

The above generalization score measures the difference of empirical and true worst-case minimax risks. For a randomized algorithm A which outputs random outcome $A(S) = (A_w(S), A_\theta(S))$ for dataset S we define A's expected generalization risk as

$$\epsilon_{\text{gen}}(A) := \mathbb{E}_{S,A}\big[R(A_w(S)) - R_S(A_w(S))\big]. \tag{9}$$

We further define A's expected minimax generalization risk as the worst-case expected difference between the true and empirical minimax risks:

$$\epsilon^{\text{mm}}_{\text{gen}}(A) := \max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{S,A}\big[R(A_w(S), \boldsymbol{\theta}) - R_S(A_w(S), \boldsymbol{\theta})\big]. \tag{10}$$

Definition 2. *A randomized minimax optimization algorithm A is called ϵ-uniformly stable in minimization if for every two datasets $S, S' \in \mathcal{Z}^n$ which differ in only one sample, for every $\mathbf{z} \in \mathcal{Z}, \boldsymbol{\theta} \in \Theta$ we have*

$$\mathbb{E}_A\big[\, f(A_w(S), \boldsymbol{\theta}; \mathbf{z}) - f(A_w(S'), \boldsymbol{\theta}; \mathbf{z}) \,\big] \leq \epsilon. \tag{11}$$

We further call A ϵ-uniformly stable in the minimization solution if it satisfies the following for every $S, S' \in \mathcal{Z}^n$:

$$\mathbb{E}_A\big[\big\|A_w(S) - A_w(S')\big\|_2\big] \leq \epsilon. \tag{12}$$

Considering the above definition, we show the following theorem that connects the definition of uniform stability to the generalization risk of the learned minimax model.

Theorem 1. *(a) Assume minimax learner A is ϵ-uniformly stable in minimization. Then, A's expected minimax generalization risk is bounded as $\epsilon^{\text{mm}}_{\text{gen}}(A) \leq \epsilon$.*

(b) Assume minimax learner A is ϵ-uniformly stable in minimization. If the maximization problem over $\boldsymbol{\theta} \in \Theta$ can be swapped with the expectation over $\mathbf{Z}$, A's expected generalization risk will be bounded as $\epsilon_{\text{gen}}(A) \leq \epsilon$.

(c) Assume that minimax learner A is ϵ-uniformly stable in the minimization solution and the minimax objective is μ-strongly-concave in $\boldsymbol{\theta}$ over a convex feasible set Θ and satisfies Assumptions 1,2. Then, defining the condition number $\kappa := \ell/\mu$, A's expected generalization risk is bounded as $\epsilon_{\text{gen}}(A) \leq \sqrt{\kappa^2 + 1}\, L\, \epsilon$.

Proof. We defer the proof to the Appendix. □

Note that the condition in the above part (b) on swapping the maximization and expectation typically holds in standard adversarial training problems (Madry et al., 2017) where the maximization subproblem decouples across samples and can be independently solved for every data point. Next, we apply the above results to analyze generalization for convex concave and non-convex non-concave minimax learning problems.

5. Generalization Analysis for Convex Concave Minimax Problems

Analyzing convergence rates for convex concave minimax problems is well-explored in the optimization literature. Here, we use the algorithmic stability framework to bound the expected generalization risk in convex concave minimax learning problems. We start by analyzing the generalization risk in strongly-convex strongly-concave problems. The following theorem applies the stability framework to bound the expected generalization risk under this scenario.

Theorem 2. *Let minimax learning objective $f(\cdot,\cdot;\mathbf{z})$ be μ-strongly-convex strongly-concave and satisfy Assumption 2 for every $\mathbf{z}$. Assume that Assumption 1 holds for convex-concave $\widetilde{f}(\mathbf{w},\boldsymbol{\theta};\mathbf{z}) := f(\mathbf{w},\boldsymbol{\theta};\mathbf{z}) + \frac{\mu}{2}(\|\boldsymbol{\theta}\|_2^2 - \|\mathbf{w}\|_2^2)$ and every $\mathbf{z}$. Then, defining $\kappa = \ell/\mu$ full-batch and stochastic GDA and GDmax algorithms with stepsize $\alpha_w = \alpha_\theta \leq \frac{\mu}{\ell^2}$ will satisfy the following bounds over T iterations:*

$$\epsilon_{\text{gen}}(\text{GDA}) \leq \frac{2L^2\sqrt{\kappa^2+1}}{(\mu - \frac{\alpha_w \ell^2}{2})n}, \epsilon_{\text{gen}}(\text{GDmax}) \leq \frac{2L^2\sqrt{\kappa^2+1}}{\mu n}.$$

Proof. We defer the proof to the Appendix. □

Note that regarding Assumption 1 in the above theorem, we suppose the assumption holds for the deregularized $\widetilde{f}$, because a strongly-convex strongly-concave objective cannot be Lipschitz over an unbounded feasible set. We still note that the theorem's bounds will hold for the original f if in Assumption 1 we define f's Lipschitz constants over bounded feasible sets $\mathcal{W}, \Theta$.

Given sufficiently small stepsizes for GDA, Theorem 2 suggests a similar generalization performance between GDA and GDmax. For general convex concave problems, it is well-known in the minimax optimization literature that the GDA algorithm can diverge from an optimal saddle point solution. As we show in the following remark, the generalization bound suggested by the stability framework will also grow exponentially with the iteration count in this scenario.

Remark 1. *Consider a convex concave minimax objective $f(\cdot,\cdot;\mathbf{z})$ satisfying Assumptions 1 and 2. Given constant stepsizes $\alpha_w = \alpha_\theta = \alpha$, the GDA's generalization risk over T iterations will be bounded as:*

$$\epsilon_{\text{gen}}(GDA) \leq O\big(\frac{\alpha L L_w (1+\alpha^2\ell^2)^{T/2}}{n}\big).$$

In particular, the bound's exponential dependence on T is tight for the GDA's generalization risk in the special case of $f(\mathbf{w},\boldsymbol{\theta};\mathbf{z}) = \mathbf{w}^\top(\mathbf{z}-\boldsymbol{\theta})$.

Proof. We defer the proof to the Appendix. □

On the other hand, proximal point methods have been shown to resolve the convergence issues of GDA methods in convex concave problems (Mokhtari et al., 2019; 2020). Here, we also show that these algorithms enjoy a generalization risk growing at most linearly with T.

Theorem 3. *Consider a convex-concave minimax learning objective $f(\cdot,\cdot;\mathbf{z})$ satisfying Assumptions 1 and 2 for every $\mathbf{z}$. Then, full-batch and stochastic PPM with parameter η will satisfy the following bound over T iterations:*

$$\epsilon_{\text{gen}}^{\text{mm}}(\text{PPM}) \leq \frac{2\eta L L_w T}{n}.$$

Furthermore, under the swapping condition in Theorem 1-b we also have

$$\epsilon_{\text{gen}}(\text{PPM}) \leq \frac{2\eta L L_w T}{n}.$$

Proof. We defer the proof to the Appendix. □

The above generalization bound allows us to analyze the true worst-case minimax risk of PPM learners in convex concave problems. To this end, we decompose the true worst-case risk into the sum of the stability and empirical worst-case risks and optimize the sum of these two error components' upper-bounds. Note that Theorem 3 bounds the generalization risk of PPM in terms of stepsize parameter η and number of iterations T. Therefore, we only need to bound the iteration complexity of PPM's convergence to an ϵ-approximate saddle point. To do this, we show the following theorem that extends (Mokhtari et al., 2019)'s result for PPM to stochastic PPM.

Theorem 4. *Given a differentiable minimax objective $f(\mathbf{w},\boldsymbol{\theta};\mathbf{z})$ the average iterate updates $\bar{\mathbf{w}}^{(T)} := \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$, $\bar{\boldsymbol{\theta}}^{(T)} := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}$ of stochastic PPM (SPPM) with setpsize parameter η will satisfy the following for a saddle point $[\mathbf{w}_S^*, \boldsymbol{\theta}_S^*]$ of the empirical risk under dataset S:*

$$\mathbb{E}_A\big[R_S(\bar{\mathbf{w}}^{(T)})\big] - R_S(\mathbf{w}_S^*) \leq \frac{\big\|[\mathbf{w}^{(0)},\boldsymbol{\theta}^{(0)}] - [\mathbf{w}_S^*,\boldsymbol{\theta}_S^*]\big\|_2^2}{2\eta T}.$$

Proof. We defer the proof to the Appendix. □

The above convergence result suggests that the expected empirical worst-case risk of applying T iterations of stochastic PPM will be at most $O(1/\eta T)$. In addition, Theorem 3 shows that using that number of iterations the generalization risk will be bounded by $O(\eta T/n)$. Minimizing the sum of these two error components, the following corollary bounds the excess risk suffered by the PPM algorithm.

Corollary 1. *Consider a convex concave minimax objective that satisfies the swapping condition in Theorem 1b*

and a proximal point method with constant parameter η. Given that $\|\mathbf{w}^{(0)}-\mathbf{w}^*\|^2+\|\boldsymbol{\theta}^{(0)}-\boldsymbol{\theta}^*\|^2\leq D^2$ holds with probability 1 for optimal saddle solution $(\mathbf{w}^*,\boldsymbol{\theta}^*)$ of the minimax risk, it will take $T_{\text{PPM}}=\sqrt{\frac{nD^2}{2\eta^2LL_w}}$ iterations for the average iterate $\bar{\mathbf{w}}^{(T)}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$ of full-batch and stochastic PPM to have the following bounded excess risk:

$$\mathbb{E}_{S,A}\big[R(\bar{\mathbf{w}}^{(T_{\text{PPM}})})\big]-R(\mathbf{w}^*)\leq\sqrt{\frac{2D^2LL_w}{n}}.$$

Proof. We defer the proof to the Appendix. In the Appendix, we prove a similar bound for full-batch and stochastic PP-max as well. □

6. Generalization Analysis for Non-convex Non-concave Minimax Problems

In the previous section, we showed that in convex-concave minimax problems simultaneous and non-simultaneous optimization algorithms have similar generalization error bounds which are different by a constant factor L/L_w. However, here we demonstrate that this result does not generalize to general non-convex non-concave problems. We first study the case of non-convex strongly-concave minimax learning problems, where we can analytically characterize the generalization bounds for both stochastic GDA and GDmax algorithms. The following theorem states the results of applying the algorithmic stability framework to bound the generalization risk in such minimax problems.

Theorem 5. *Let learning objective $f(\mathbf{w},\boldsymbol{\theta};\mathbf{z})$ be non-convex μ-strongly-concave and satisfy Assumptions 1 and 2. Also, we assume that $f_{\max}(\mathbf{w};\mathbf{z}):=\max_{\boldsymbol{\theta}\in\Theta}f(\mathbf{w},\boldsymbol{\theta};\mathbf{z})$ is bounded as $0\leq f_{\max}(\mathbf{w};\mathbf{z})\leq 1$ for every $\mathbf{w},\mathbf{z}$. Then, defining $\kappa:=\ell/\mu$ we have*

1. The stochastic GDA (SGDA) algorithm with stepsizes $\alpha_{w,t}=c/t$, $\alpha_{\theta,t}=cr^2/t$ for constants $c>0, 1\leq r\leq\kappa$ satisfies the following bound over T iterations:

$$\epsilon_{\text{gen}}(\text{SGDA})\leq\frac{1+\frac{1}{(r+1)c\ell}}{n}\times\big(12cL^2(r+1)\sqrt{\kappa^2+1}\big)^{\frac{1}{(r+1)c\ell+1}}T^{\frac{(r+1)c\ell}{(r+1)c\ell+1}}.$$

2. The stochastic GDmax (SGDmax) algorithm with stepsize $\alpha_{w,t}=c/t$ for constant $c>0$ satisfies the following bound over T iterations:

$$\epsilon_{\text{gen}}(\text{SGDmax})\leq\frac{1+\frac{1}{(\kappa+1)\ell c}}{n}\times\big(2cL^2(\kappa^2+1)\big)^{\frac{1}{(\kappa+1)\ell c+1}}T^{\frac{(\kappa+1)\ell c}{(\kappa+1)\ell c+1}}.$$

Proof. We defer the proof to the Appendix. □

The above result shows that the generalization risks of stochastic GDA and GDmax change with the number of iterations and training set size as:

$$\epsilon_{\text{gen}}(\text{SGDA})\approx\mathcal{O}\big(T^{\frac{\ell(r+1)c}{\ell(r+1)c+1}}/n\big),$$
$$\epsilon_{\text{gen}}(\text{SGDmax})\approx\mathcal{O}\big(T^{\frac{\ell(\kappa+1)c}{\ell(\kappa+1)c+1}}/n\big). \tag{13}$$

Therefore, considering a maximization to minimization stepsize ratio of $r^2<\kappa^2$ will result in a better generalization bound for stochastic GDA compared to stochastic GDmax over a fixed and sufficiently large number of iterations.

Next, we consider general non-convex non-concave minimax problems and apply the algorithmic stability framework to bound the generalization risk of the stochastic GDA algorithm. Note that the maximized value of a non-strongly-concave function is in general non-smooth. Consequently, the stability framework does not result in a bounded generalization risk for the GDmax algorithm in general non-convex non-concave problems.

Theorem 6. *Let $0\leq f(\cdot,\cdot;\mathbf{z})\leq 1$ be a bounded non-convex non-concave objective satisfying Assumptions 1 and 2. Then, the SGDA algorithm with stepsizes $\max\{\alpha_{w,t},\alpha_{\theta,t}\}\leq c/t$ for constant $c>0$ satisfies the following bound over T iterations:*

$$\epsilon^{\text{mm}}_{\text{gen}}(\text{SGDA})\leq\frac{1+\frac{1}{\ell c}}{n}\big(2cLL_w\big)^{\frac{1}{\ell c+1}}T^{\frac{\ell c}{\ell c+1}}. \tag{14}$$

Moreover, under the swapping condition in Theorem 1b we also have

$$\epsilon_{\text{gen}}(\text{SGDA})\leq\frac{1+\frac{1}{\ell c}}{n}\big(2cLL_w\big)^{\frac{1}{\ell c+1}}T^{\frac{\ell c}{\ell c+1}}. \tag{15}$$

Proof. We defer the proof to the Appendix. □

Theorem 6 also shows that the SGDA algorithm with vanishing stepsize values will have a bounded generalization risk of $O(T^{\frac{\ell c}{\ell c+1}}/n)$ over T iterations. On the other hand, the stochastic GDmax algorithm could not enjoy a bounded algorithmic stablility degree in non-convex non-concave problems, since the optimal maximization value behaves non-smoothly in general.

7. Numerical Experiments

Here, we numerically examine the theoretical results of the previous sections. We first focus on a Gaussian setting for analyzing strongly-convex strongly-concave and convex concave minimax problems. Then, we empirically study generative adversarial networks (GANs) as non-convex non-concave minimax learning tasks.

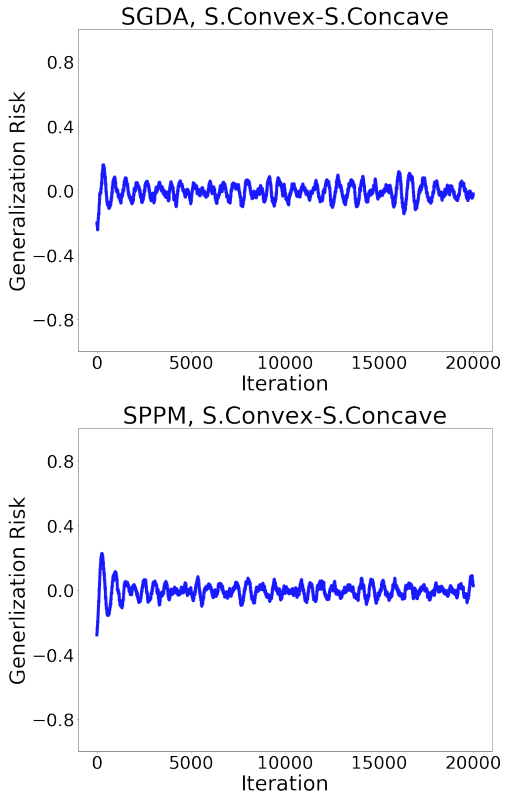

Figure 1. Generalization risk vs. iteration in the strongly-convex strongly-concave setting optimized by (top) stochastic GDA and (bottom) stochastic PPM.

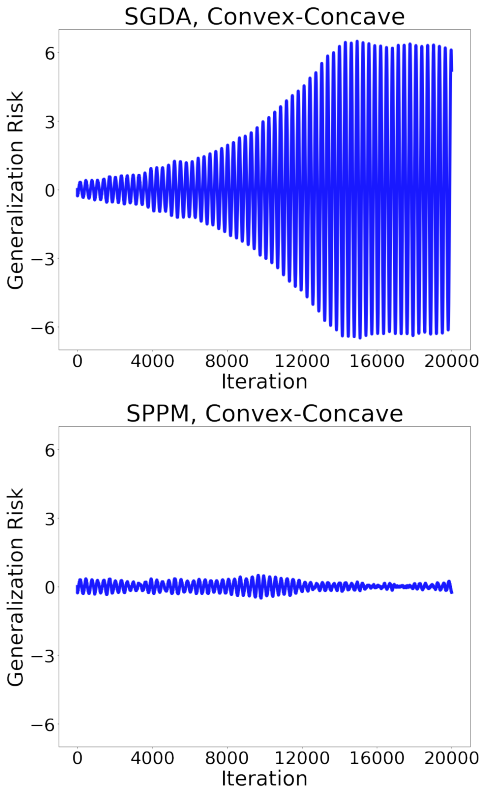

Figure 2. Generalization risk vs. iteration in the convex concave bilinear setting optimized by (top) stochastic GDA and (bottom) stochastic PPM.

7.1. Convex Concave Minimax Problems

To analyze our generalization results for convex concave minimax settings, we considered an isotropic Gaussian data vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$ with zero mean and identity covariance. In our experiments, we chose $\mathbf{Z}$'s dimension to be $d = 50$. We drew $n = 1000$ independent samples from the underlying Gaussian distribution to form a training dataset $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. For the μ-strongly-convex strongly-concave scenario, we considered the following minimax objective:

$$f_1(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z}) = \mathbf{w}^\top(\mathbf{z} - \boldsymbol{\theta}) + \frac{\mu}{2}\big(\|\mathbf{w}\|_2^2 - \|\boldsymbol{\theta}\|_2^2\big). \quad (16)$$

In our experiments, we used $\mu = 0.1$ and constrained the optimization variables to satisfy the norm bounds $\|\mathbf{w}\|_2, \|\boldsymbol{\theta}\|_2 \leq 100$ which we enforced by projection after every optimization step. Note that for the above minimax objective we have

$$\epsilon_{\text{gen}}(\mathbf{w}) = \mathbf{w}^\top(\mathbb{E}[\mathbf{Z}] - \mathbb{E}_S[\mathbf{Z}]), \quad (17)$$

where $\mathbb{E}[\mathbf{Z}] = \mathbf{0}$ is the underlying mean and $\mathbb{E}_S[\mathbf{Z}] := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$ is the empirical mean.

To optimize the empirical minimax risk, we applied stochastic GDA with stepsize parameters $\alpha_w = \alpha_\theta = 0.02$ and stochastic PPM with parameter $\eta = 0.02$ each for $T = 20,000$ iterations. Figure 1 shows the generalization risk values over the optimization achieved by the stochastic GDA (top) and PPM (bottom) algorithms. As shown in this figure, the absolute value of generalization risk remained bounded during the optimization for both the learning algorithms. In our experiments, we also observed a similar generalization behavior with full-batch GDA and PPM algorithms. We defer the results of those experiments to the supplementary document. Hence, our experimental results support Theorem 2's generalization bounds.

Regarding convex concave minimax problems, as suggested by Remark 1 we considered the following bilinear minimax objective in our experiments:

$$f_2(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z}) = \mathbf{w}^\top(\mathbf{z} - \boldsymbol{\theta}). \quad (18)$$

We constrained the norm of optimization variables as $\|\mathbf{w}\|_2, \|\boldsymbol{\theta}\|_2 \leq 100$ which we enforced through projection after every optimization iteration. Similar to the strongly-convex strongly-concave objective (16), for the above minimax objective we have the generalization risk in (17) with $\mathbb{E}[\mathbf{Z}]$ and $\mathbb{E}_S[\mathbf{Z}]$ being the true and empirical mean vectors.

We optimized the minimax objective (18) via stochastic and full-batch GDA and PPM algorithms. Figure 2 demonstrates

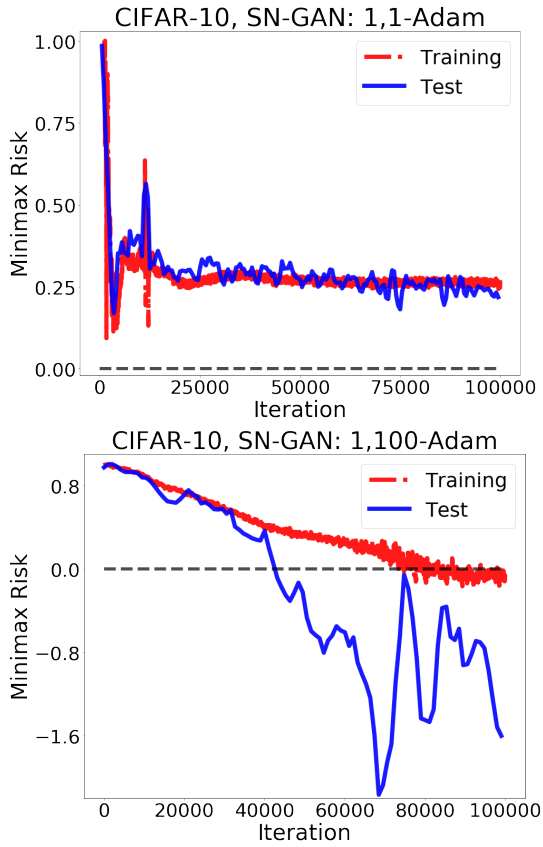

Figure 3. Minimax risk vs. iteration in the non-convex non-concave SN-GAN problem on CIFAR-10 data optimized by (top) 1,1 Adam descent ascent and (bottom) 1,100 Adam descent ascent

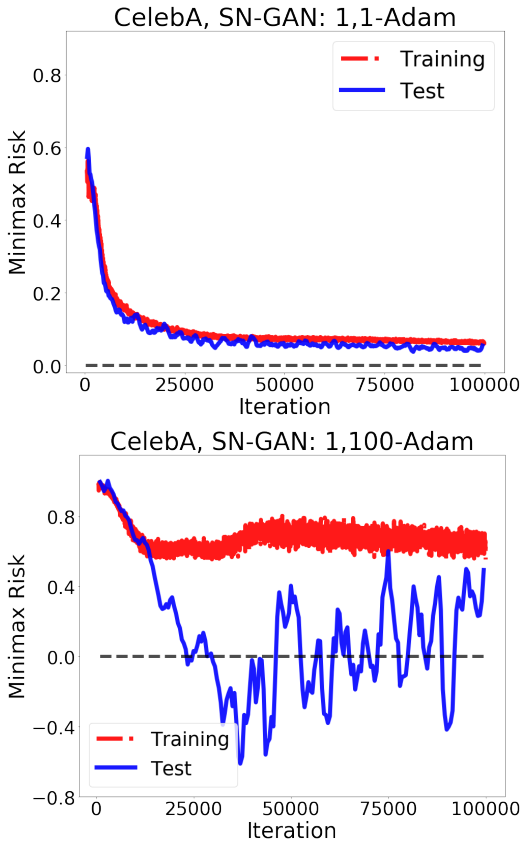

Figure 4. Minimax risk vs. iteration in the non-convex non-concave SN-GAN problem on CelebA data optimized by (top) 1,1 Adam descent ascent and (bottom) 1,100 Adam descent ascent.

the generalization risk evaluated at different iterations of applying stochastic GDA and PPM algorithms. As suggested by Remark 1, the generalization risk of stochastic GDA grew exponentially over the first 15,000 iterations before the variables reached the boundary of their feasible sets and then the generalization risk oscillated with a nearly constant amplitude of 6.2. On the other hand, we observed that the generalization risk of the stochastic PPM algorithm stayed bounded and below 0.5 for all the 20,000 iterations (Figure 2-bottom). Therefore, our numerical experiments also indicate that while in general convex concave problems the stochastic GDA learner can potentially suffer from a poor generalization performance, the PPM algorithm has a bounded generalization risk as shown by Theorem 3.

7.2. Non-convex Non-concave Problems

To numerically analyze generalization in general non-convex non-concave minimax problems, we experimented the performance of simultaneous and non-simultaneous optimization algorithms in training GANs. In our GAN experiments, we considered the standard architecture of DC-GANs (Radford et al., 2015) with 4-layer convolutional neural net generator and discriminator functions. For the minimax objective, we used the formulation of vanilla GAN (Goodfellow et al., 2014) that is

$$f(\mathbf{w}, \boldsymbol{\theta}; \mathbf{z}) = \log(D_{\mathbf{w}}(\mathbf{z})) + \mathbb{E}_{\boldsymbol{\nu}}\big[\log(1 - D_{\mathbf{w}}(G_{\boldsymbol{\theta}}(\boldsymbol{\nu})))\big].$$

For computing the above objective, we used Monte-Carlo simulation using 100 fresh latent samples $\boldsymbol{\nu}_i \sim \mathcal{N}(\mathbf{0}, I_{r=128})$ to approximate the expected value over generator's latent variable $\boldsymbol{\nu}$ at every optimization step. We followed all the experimental details from (Gulrajani et al., 2017)'s standard implementation of DC-GAN. Furthermore, we applied spectral normalization (Miyato et al., 2018) to regularize the discriminator function and assist reaching a near optimal solution for discriminator via boundedly many iterations needed for non-simultaneous optimization methods. We trained the spectrally-normalized GAN (SN-GAN) problem over CIFAR-10 (Krizhevsky et al., 2009) and CelebA (Liu et al., 2018) datasets. We divided the CIFAR-10 and CelebA datasets to 50,000, 160,000 training and 10,000, 40,000 test samples, respectively.

To optimize the minimax risk function, we used the standard Adam algorithm (Kingma & Ba, 2014) with batch-size 100.

For simultaneous optimization algorithms we applied 1,1 Adam descent ascent with the parameters $\mathrm{lr} = 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.9$ for both minimization and maximization updates. To apply a non-simultaneous algorithm, we used 100 Adam maximization steps per minimization step and increased the maximization learning rate to 5×10^{-4}. We ran each GAN experiment for T =100,000 iterations.

Figures 3, 4 show the estimates of the empirical and true minimax risks in the CIFAR-10 and CelebA experiments, respectively. We used 2000 randomly-selected samples from the training and test sets for every estimation task. As seen in these figures, for the experiments applying simultaneous 1,1 Adam optimization the empirical minimax risk generalizes properly from training to test samples (Figures 3,4-top). In contrast, in both the experiments with non-simultaneous methods after 30,000 iterations the empirical minimax risk suffers from a considerable generalization gap from the true minimax risk (Figures 3,4-bottom). The gap between the training and test minimax risks grew between iterations 30,000-60,000. The test minimax risk fluctuated over the subsequent iterations, which could be due to the insufficiency of 100 Adam ascent steps to follow the optimal discriminator solution at those iterations.

The numerical results of our GAN experiments suggest that non-simultaneous algorithms which attempt to fully solve the maximization subproblem at every iteration can lead to large generalization errors. On the other hand, standard simultaneous algorithms used for training GANs enjoy a bounded generalization error which can help the training process find a model with nice generalization properties. We defer further experimental results to the supplementary document.

8. Conclusion

In this paper, we study the generalization properties of standard gradient-based min-max optimization algorithm from the lens of algorithmic stability. We establish generalization error bounds for the simultaneous and non-simultaneous update optimization algorithms. In the strongly-convex strongly-concave case, our bounds indicate a similar generalization behavior between simultaneous and non-simultaneous optimization algorithms, whereas in the non-convex strongly-concave scenario our bounds suggest a superior performance for simultaneous stochastic GDA than non-simultaneous stochastic GDmax provided that the min-max stepsize ratio is below the condition number of the problem. As potential future directions, analyzing the tightness of the shown bounds and extending the stability-based analysis to standard GAN and adversarial training settings can be further explored.

Acknowledgements

This work is supported by the MIT-Air Force AI Accelerator (AIAA) under grant FA8750-19-2-1000. Also, the authors would like to thank the anonymous reviewers for their constructive feedback.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019.

Arora, S. and Zhang, Y. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pp. 162–183, 2019.

Bai, Y., Ma, T., and Risteski, A. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of machine learning research*, 2(Mar):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.

Chatterjee, S. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. *arXiv preprint arXiv:2002.10657*, 2020.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

Du, S. S. and Hu, W. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 196–205. PMLR, 2019.

Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

Feizi, S., Farnia, F., Ginart, T., and Tse, D. Understanding gans in the lqg setting: Formulation, generalization and stability. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Feldman, V. and Vondrak, J. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pp. 9747–9757, 2018.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *arXiv preprint arXiv:1902.10710*, 2019.

Fiez, T., Chasnov, B., and Ratliff, L. J. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

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.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pp. 2810–2819, 2019.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lei, Y., Yang, Z., Yang, T., and Ying, Y. Stability and generalization of stochastic gradient methods for minimax problems. *arXiv preprint arXiv:2105.03793*, 2021.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Liang, T. and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 907–915, 2019.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15: 2018, 2018.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pp. 3345–3354, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mazumdar, E., Ratliff, L. J., and Sastry, S. S. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

Mazumdar, E. V., Jordan, M. I., and Sastry, S. S. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

McAllester, D. A. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. Convergence rate of o(1/k) for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. *arXiv preprint arXiv:1906.01115*, 2019.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 11615–11626, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in neural information processing systems*, pp. 5947–5956, 2017a.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017b.

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

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rice, L., Wong, E., and Kolter, J. Z. Overfitting in adversarially robust deep learning. *arXiv preprint arXiv:2002.11569*, 2020.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7091–7101, 2018.

Schäfer, F., Zheng, H., and Anandkumar, A. Implicit competitive regularization in gans. *arXiv preprint arXiv:1910.05852*, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

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

Thanh-Tung, H., Tran, T., and Venkatesh, S. Improving generalization and stability of generative adversarial networks. *arXiv preprint arXiv:1902.03984*, 2019.

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

Wei, C. and Ma, T. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019.

Wu, B., Zhao, S., Chen, C., Xu, H., Wang, L., Zhang, X., Sun, G., and Zhou, J. Generalization in generative adversarial networks: A novel perspective from privacy protection. In *Advances in Neural Information Processing Systems*, pp. 307–317, 2019.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, G. and Wang, Y. On the suboptimality of negative momentum for minimax optimization. *arXiv preprint arXiv:2008.07459*, 2020.

Zhang, G., Bao, X., Lessard, L., and Grosse, R. A unified analysis of first-order methods for smooth games

via integral quadratic constraints. *arXiv preprint arXiv:2009.11359*, 2020a.

Zhang, G., Wu, K., Poupart, P., and Yu, Y. Newton-type methods for minimax optimization. *arXiv preprint arXiv:2006.14592*, 2020b.

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

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.