What is a Good Metric to Study Generalization of Minimax Learners?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Minimax optimization has served as the backbone of many machine learning (ML) problems. Although the *convergence behavior* of optimization algorithms has been extensively studied in minimax settings, their *generalization* guarantees in the stochastic setting, i.e., how the solution trained on empirical data performs on the unseen testing data, have been relatively underexplored. A fundamental question remains elusive: *What is a good metric to study generalization of minimax learners?* In this paper, we aim to answer this question by first showing that *primal risk*, a universal metric to study generalization in minimization, fails in simple examples of minimax problems. Furthermore, another popular metric, the *primal-dual risk*, also fails to characterize the generalization behavior for minimax problems with nonconvexity, due to non-existence of saddle points. We thus propose a new metric to study generalization of minimax learners: the *primal gap*, to circumvent these issues. Next, we derive generalization bounds for the primal gap in nonconvex-concave settings. As byproducts of our analysis, we also solve two open questions: establishing generalization bounds for primal risk and primal-dual risk in the strong sense, i.e., without strong concavity or assuming that the maximization and expectation can be interchanged, while either of these assumptions was needed in the literature. Finally, we leverage this new metric to compare the generalization behavior of two popular algorithms – gradient descent-ascent (GDA) and gradient descent-max (GDMax) in stochastic minimax optimization.

1 Introduction

Stochastic minimax optimization, a classical and fundamental problem in operations research and game theory, involves solving the following problem:

$$\min_{w \in W} \max_{\theta \in \Theta} E_{z \sim P_z}[f(w, \theta; z)].$$

More recently, such minimax formulations have received increasing attention in machine learning, with significant applications in generative adversarial networks (GANs) [Goodfellow et al., 2014], adversarial learning [Madry et al., 2017], and reinforcement learning [Chen and Wang, 2016, Dai et al., 2018]. Most existing works have focused on the *optimization* aspect of the problem, i.e., studying the rates of convergence, robustness, and optimality of several algorithms, in various minimax settings including convex-concave [Nemirovski et al., 2009, Monteiro and Svaiter, 2010], nonconvex-concave [Lin et al., 2020, Rafique et al., 2018], and certain special nonconvex-nonconcave [Nouiehed et al., 2019, Yang et al., 2020] problems.

However, the optimization aspect is not the only venue for understanding the success of minimax optimization in machine learning. In particular, as in classical supervised learning, which is usually studied as a *minimization* problem [Hastie et al., 2009], the out-of-sample *generalization* performance

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

is a key metric for evaluating the learned models. The study of generalization guarantees in minimax optimization (and related machine learning problems) has not received significant attention until recently [Arora et al., 2017, Feizi et al., 2020, Yin et al., 2019, Lei et al., 2021, Farnia and Ozdaglar, 2021, Zhang et al., 2021b]. Specifically, existing works along this line have investigated two types of generalization guarantees: *uniform* convergence generalization bounds, and *algorithm-dependent* generalization bounds. The former is more general and irrespective of the optimization algorithms being used, while the latter is usually finer and really explains what happens in practice, when optimization algorithms play an indispensable role. In fact, the former might not be able to explain generalization in deep learning [Nagarajan and Kolter, 2019], making the latter a more favorable metric for understanding the success of minimax optimization in machine learning.

Algorithm-dependent generalization for minimax optimization has been studied recently in [Farnia and Ozdaglar, 2021, Lei et al., 2021, Xing et al., 2021, Yang et al., 2022, Li and Liu, 2021]. These papers build on the algorithmic stability framework developed in [Bousquet and Elisseeff, 2002]. In particular, these works have studied *primal risk* and/or (variants of) *primal-dual risk* under different convexity and smoothness assumptions of the objective. Primal risk (see formal definition in §2) is a natural extension of risk from minimization problems. Primal-dual risk, on the other hand, is defined similarly but based on the duality gap of the solution, which makes sense when the saddle-point exists (i.e., $\min$ and $\max$ can be interchanged). Based on these metrics, [Farnia and Ozdaglar, 2021, Lei et al., 2021, Li and Liu, 2021], compare the performance of specific algorithms, e.g., gradient descent-ascent (GDA) and gradient descent-max (GDMax). We provide a more thorough literature review and comparison of results in Appendix A.

Although these metrics are natural extensions of generalization metrics from the *minimization* setting, they might not be the most suitable ones for studying generalization in stochastic *minimax* optimization, especially in the *nonconvex* settings that is pervasive in machine/deep learning applications. In particular, we are interested in the following fundamental question:

What is a good metric to study generalization of minimax learners?

In this paper, we make an initial attempt to answering this question, by identifying the inadequacies of existing metrics, and proposing a new metric, the *primal gap* that overcomes these inadequacies. We then provide generalization bounds for the newly proposed metric, and discuss how it captures information not included in the other existing metrics. We summarize our contributions as follows.

Contributions. First, we introduce an example through which we identify the inadequacies of *primal risk*, a well-studied metric for minimax generalization, in capturing the generalization behavior of *nonconvex-concave* minimax problems. Second, to address the issue, we propose a new metric – the *primal gap*, which provably avoids the issue in the example, and derive its generalization bounds. Next, we leverage this new metric to compare the generalization behavior of GDA and GDMax, two popular algorithms for minimax optimization and GAN training, and answer the question of *when does GDA generalize better than GDMax?* Moreover, we also address two open questions in the literature: establishing generalization bounds for primal risk and primal-dual risk without strong concavity or assuming that the maximization and expectation can be interchanged, while either of these assumptions was needed in the literature [Farnia and Ozdaglar, 2021, Lei et al., 2021, Xing et al., 2021, Yang et al., 2022] (see Section §E due to space limitation). Finally, under certain assumptions of the max learner, our results also generalize to the nonconvex-nonconcave setting.

2 Preliminaries

2.1 Problem formulation

In this paper, we consider the following (stochastic) minimax problem:

$$\min_{w \in W} \max_{\theta \in \Theta} E_{z \sim P_z} f(w, \theta; z). \tag{1}$$

We make the following assumption on the sets W and Θ throughout the paper.

Assumption 1. *W and Θ are convex, closed sets, and we further assume that W is compact with $\|w\| \leq M(W), \forall w \in W$. Here $M(W)$ is a constant dependent on the set W.*

Let $r(w, \theta) = E_{z \sim P_z} f(w, \theta; z)$. For a training dataset $S = \{z_1, \cdots, z_n\}$ with $n > 1$ i.i.d. samples drawn from P_z, we define $r_S(w, \theta) = \frac{1}{n} \sum_{i=1}^{n} f(w, \theta; z_i)$. Next, we define the following quantity:

Definition 1 (Primal Risk (Empirical/Population))**.** ***Primal population risk*** *is given by* $r(w) = \max_{\theta\in\Theta} E_{z\sim P_z} f(w,\theta;z)$*, and the* ***primal empirical risk*** *is given by:* $r_S(w) = \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} f(w,\theta;z_i)$.

Throughout this paper, we use (w_S, θ_S) to denote a solution of the min-max problem: $\min_{w\in W}\max_{\theta\in\Theta} r_S(w,\theta)$. Notice that (w_S,θ_S) need not be a saddle point of r_S. Furthermore, we use (w^*,θ^*) to denote a solution of $\min_{w\in W}\max_{\theta\in\Theta} r(w,\theta)$. Once again, notice that (w^*,θ^*) may not be a saddle point of r.

The goal in Problem (1) is to minimize the primal population risk $r(w)$. Note that this function can be decomposed as

$$r(w) = r_S(w) + (r(w) - r_S(w)). \tag{2}$$

In practice, we only have access to $r_S(w,\theta)$ and our goal is to design algorithms for minimizing $r(w)$ using dataset S. Suppose A is a learning algorithm initialized at $(w,\theta) = (0,0)$. We define (w_S^A, θ_S^A) to be the output of Algorithm A using dataset S.

From Equation (2), it is clear if we ensure $r_S(w_S^A)$ as well as $r(w_S^A) - r_S(w_S^A)$ are small, this would guarantee that $r(w_S^A)$ is small, which is the goal of Problem (1). Note that we can always ensure that $r_S(w_S^A)$ is small by using a good optimization Algorithm A (if the problem is tractable). The main goal in the study of generalization is therefore to estimate the generalization error $r(w_S^A) - r_S(w_S^A)$.

Definition 2. *The generalization error for the primal risk is defined as:*

$$\zeta_{gen}^{P}(A) = E_S E_A[r(w_S^A) - r_S(w_S^A)]. \tag{3}$$

Here the expectations are taken over the randomness in the dataset S, as well as any randomness used in the Algorithm A.

This metric has been used to study generalization in stochastic minimization problems, i.e., when the maximization set Θ is a singleton, as well as several recent works in minimax optimization (see [Hardt et al., 2016, Farnia and Ozdaglar, 2021]).

We are interested in when the solution to the empirical problem w_S^A has good *generalization behavior*, i.e., when $E[r(w_S^A) - \min_{w\in W} r(w)]$ is small - w_S^A is an approximate minimizer of the primal population risk r. In the next subsection, we briefly describe why the generalization error of the primal risk $\zeta_{gen}^{P}(A)$ is a good measure to study generalization behavior in minimization problems.

2.1.1 $\zeta_{gen}^{P}(A)$ for minimization problems

Consider a minimization problem of the form

$$\min_{w\in W} E_{z\sim P_z}[g(w;z)]. \tag{4}$$

In this case, we define the (minimization) primal risk (population and empirical version respectively) as: $m(w) = E_{z\sim P_z} g(w;z)$, and $m_S(w) = \frac{1}{n}\sum_{i=1}^{n} g(w;z_i)$. The generalization error $\zeta_{gen}^{P,min}(A)$ for the (minimization) primal risk is the same as in Definition 2 using the (minimization) primal risk.

Assume that the generalization error of the primal risk for an Algorithm A is small, say $\zeta_{gen}^{P,min}(A) \le \epsilon$. This implies that (from Definition 2): $E[m(w_S^A)] \le E[m_S(w_S^A)] + \epsilon$. Note that the expectation is with respect to S and A. Now, in order to show that w_S^A has good generalization behavior, we first see that:

$$E[m(w_S^A) - \min_{w\in W} m(w)] \le E[m_S(w_S^A)] + \epsilon - \min_{w\in W} m(w). \tag{5}$$

However, note that for minimization problems, since $E[m_S] = m$, we have [1]: $\min_{w\in W} m(w) \ge E[\min_{w\in W} m_S(w)]$, which gives us:

$$\begin{aligned} E[m(w_S^A) - \min_{w\in W} m(w)] &\le E[m_S(w_S^A)] + \epsilon - E[\min_{w\in W} m_S(w)] \\ &= E[m_S(w_S^A) - \min_{w\in W} m_S(w)] + \epsilon = \epsilon. \end{aligned}$$

[1] Here we use the fact that $E_z[\min_x f(x,z)] \le \min_x E_z[f(x,z)]$.

Therefore, for minimization problems, if the generalization error for primal risk is small, we have good generalization behavior of the solution to the empirical risk minimization. In the next subsection, we highlight some results in the literature which discusses bounds on generalization error of the primal risk. These results depend on the concept of algorithmic stability which we discuss next.

2.2 Stability of algorithms

Stability analysis is a powerful tool to analyse the generalization behavior of algorithms (see [Bousquet and Elisseeff, 2002]). In this section, we will review some definitions and theoretical results about stability bounds existing in the current literature. More specifically, in this paper, we adopt the following definition of stability:

Definition 3. *Suppose that A is a randomized algorithm for solving the minimax problem. We define (w_S^A, θ_S^A) as the output of Algorithm A using dataset S. We say S and S' are neighboring dataset if they defer only in one sample. An Algorithm A is defined to be ϵ-stable if $\|w_S^A - w_{S'}^A\| \leq \epsilon$, $\|\theta_S^A - \theta_{S'}^A\| \leq \epsilon$ for any neighboring datasets S and S'.*

From [Hardt et al., 2016], we have the following basic result for the generalization error of $r_S(w, \theta)$:

Theorem 1 ([Hardt et al., 2016])**.** *Consider the (stochastic) minimization problem defined in* (4)*. Suppose $g(\cdot; z)$ is L-Lipschitz continuous, i.e., $\|g(w_1; z) - g(w_2; z)\| \leq L\|w_1 - w_2\|, \forall w_1, w_2 \in W$. Then, for an ϵ-stable Algorithm A, we have $|E_S E_A[r(w_S^A) - r_S(w_S^A)]| \leq L\epsilon$.*

However, this theorem cannot be directly extended to analyze the generalization behavior of minimax learners because we have an additional maximization step before taking expectation.

In this paper, we first show that good bounds for $\zeta_{gen}^P(A)$ is not sufficient to conclude the generalization behaviour in minimax problems. We then propose a new measure, the primal gap $\zeta_{gen}^{PG}(A)$ which is more insightful to understand generalization behavior in minimax problems. We provide bounds for $\zeta_{gen}^{PG}(A)$ in the nonconvex-concave setting (which includes the more restrictive nonconvex - strongly concave case) and as a byproduct, we also derive bounds for $\zeta_{gen}^P(A)$ for the nonconvex-concave case, which to the best of our knowledge, is an open problem.

3 Primal Gap: A New Metric to Study Generalization

The main idea which leads to the success of $\zeta_{gen}^P(A)$ as a metric to study generalization for minimization learners is that $E[\min_{w \in W} r_S(w)] \leq \min_{w \in W} r(w)$, which is no longer the case in the minimax case. In fact, in this section, we show that a good bound for the generalization error of primal risk does not imply good generalization behavior in minimax problems. We then propose a new metric, the primal gap, which circumvents these issues and provides insights into the generalization behavior in minimax problems.

3.1 Is $\zeta_{gen}^P(A)$ a good metric?

We provide an example where the generalization error of the primal risk is small, but the final solution to the empirical problem has poor generalization behavior. In this example, the minimizer of $r_S(w)$ is suboptimal for $r(w)$ with high probability, and $E_S[r(w_S) - r(w^*)]$ is large.

Example 1 (Analytical example)**.** *Let $y \sim N(0, 1/\sqrt{n})$ be a Gaussian random variable in $\mathbb{R}$. Define the truncated Gaussian variable $z \sim P_z$ as follows: $z = y$ if $|y| < \lambda \log n/\sqrt{n}$ and $z = \lambda \log n/\sqrt{n}$ if $y \geq \lambda \log n/\sqrt{n}$. Let $f(w, \theta; z) = \frac{1}{2}w^2 - \left(\frac{1}{2n^2}\theta^2 - z\theta + 1\right) w$, where $w \in W = [0, 1]$, $\theta \in \Theta = [-\lambda n, \lambda n]$ with a sufficiently large $\lambda > 0$, and $z_i \sim P_z$ be i.i.d truncated Gaussian variables. Then, we have $r_S(w, \theta) = \frac{1}{2}w^2 - \left(\frac{1}{2n^2}\theta^2 - \frac{\sum_{i=1}^n z_i}{n}\theta + 1\right) w$, and*

$$r(w, \theta) = \frac{1}{2}w^2 - \left(\frac{1}{2n^2}\theta^2 + 1\right) w. \tag{6}$$

Note that this leads to the primal population risk function: $r(w) = \frac{1}{2}w^2 - w$.

It is not hard to see that we always have $r_S(w) \geq r(w)$. Note that this means $\zeta_{gen}^P(A) < 0$, and thus we have good generalization of primal risk. However, we can prove that for large enough λ,

$$E_S[r(w_S) - r(w^*)] \geq 0.02. \tag{7}$$

This means that w_S has a constant error compared to w^* in terms of the population risk, despite that the generalization error of the primal risk is small. This phenomenon is due to that $\min_{w\in W} r_S(w) - \min_{w\in W} r(w) > c$ for some $c > 0$, and hence minimizing $r_S(w)$ is very different from minimizing $r(w)$.

This example shows that the generalization error of primal risk is not a good measure to study generalization in minimax learners. The main drawback is that $E_S[\min_w r_S(w) - \min_{w\in W} r(w)] \leq 0$ no longer hold for minimax problems while it always holds for minimization problems. We now introduce another more practical example, from GAN training, to further illustrate this point.

Example 2 (GAN-training example)**.** *Suppose that we have a real distribution P_r in $\mathbb{R}^d$ which can be represented as $G^*(y)$ with $y \in \mathbb{R}^k$ drawn from a standard Gaussian distribution P_0 and a mapping $G^* : \mathbb{R}^k \to \mathbb{R}^d$. For an arbitrary generator G, we define P_G to be the distribution of the random variable $G(y)$ with $y \sim P_0$. So our goal is to find a generator G such that $P_G = P_r$. The generative adverserial network (GAN) is a popular tool for solving this problem. Consider a Generative Adversarial Network (GAN) with generator G and discriminator D. The goal of GAN training is to find a pair of generator G and a discriminator D that solves the min-max problem:*

$$\begin{aligned}\min_G \max_D \quad & \{E_{x\sim P_r}\Phi(D(x)) + E_{x\sim P_G}[\Phi(1 - D(x))]\} \\ &= \min_G \max_D \quad \{E_{x\sim P_r}\Phi(D(x)) + E_{y\sim P_0}[\Phi(1 - D(G(y)))]\},\end{aligned}$$

where $\Phi : \mathbb{R} \to \mathbb{R}$ is concave, monotonically increasing and $\Phi(u) = -\infty$ for $u \leq 0$. To connect to the minimax formulation in (1)*, we note that $z = (x, y)$, and $P_z = P_r \times P_0$. Also, we denote*

$$r(G, D) = E_{x\sim P_r}\Phi(D(x)) + E_{y\sim P_0}[\Phi(1 - D(G(y)))]$$

to be the population risk. We now give the empirical version of this problem. Let $S_1 = \{x_1, \cdots, x_n\}$ and $S_2 = \{y_1, \cdots, y_n\}$. Let $S = S_1 \cup S_2$ and $r_S(G, D) = \frac{1}{n}\left(\sum_{i=1}^n \Phi(D(x_i) + \Phi(1 - D(G(y_i)))\right)$. We assume that P_{G_w} has the same support set as P_r. Moreover, we assume that $\|w - w^*\| \leq 0.5$ and $G_w(y)$ is 1-Lipschitz w.r.t. w for any y. Then we have (see Theorem B.1 in Arora et al. [2017])

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \geq (1 - \delta)\,(2\Phi(1) - 2\Phi(1/2))$$

which is bounded away from 0.

Note that in this example, we also have $E_S[\min_w r_S(w) - \min_w r(w)] > 0$, implying that $\zeta^P_{gen}(A)$ might not be a good metric. To address this issue, we next define a new metric to study the generalization of minimax learners.

3.2 Primal gap

The empirical and population versions of the primal gap are defined as follows:

Definition 4 (Primal gap (empirical/population))**.** *The **population primal gap** is defined as $\Delta(w) = r(w) - \min_{w\in W} r(w)$, and the **empirical primal gap** is defined as $\Delta_S(w) = r_S(w) - \min_{w\in W} r_S(w)$.*

Notice that these two primal gaps can always take 0 at $w_S = \arg\min_{w\in W} r_S(w)$ and $w^* = \arg\min_{w\in W} r(w)$ respectively even if the saddle point of problem 2 does not exist. Next, we define the expected generalization error of this primal gap as follows:

Definition 5. *The generalization error for the primal gap is defined as*

$$\zeta^{PG}_{gen}(A) = E_S E_A[\Delta(w^A_S) - \Delta_S(w^A_S)].$$

Notice that we have the decomposition

$$\begin{aligned}&E_S E_A[r(w^A_S) - \min_{w\in W} rw] \\ &= E_S E_A[r_S(w^A_S) - \min_{w\in W} r_S(w)] + E_S E_A[(r(w^A_S) - \min_{w\in W} r(w)) - (r_S(w^A_S) - \min_{w\in W} r_S(w))] \\ &= E_S E_A[r_S(w^A_S) - \min_{w\in W} r_S(w)] + \zeta^{PG}_{gen}(A).\end{aligned}$$

If $E_S E_A[r_S(w^A_S) - \min_{w\in W} r_S(w)]$ is small and $\zeta^{PG}_{gen}(A)$ is small (or large), then $E_S E_A[r(w^A_S) - \min_{w\in W} r(w)]$ is small (or large).

Next, we provide bounds on $\zeta^{PG}_{gen}(A)$ for stable algorithms A, and show that in Example 1, $\zeta^{PG}_{gen}(A)$ cannot be small (unlike $\zeta^P_{gen}(A)$).

3.3 Relationship between generalization and stability

We provide bounds for the generalization error of the primal gap (Definition 5) for $\epsilon-$stable algorithms A. We will focus on the nonconvex-concave case where the following assumptions are made throughout the rest of the paper.

Assumption 2. *The function f in Problem* (1) *is nonconvex-concave, i.e. $f(w, \cdot; z)$ is a concave function for all $w \in W$ and for all z.*

Next we define the notion of *capacity*, which will play a key role in the bounds we derive for $\zeta_{gen}^{PG}(A)$.

Definition 6 (Capacity)**.** *For any $w \in W$ and any constraint set Θ, we define*

$$\Theta(w) = \arg\max_{\theta \in \Theta} r(w, \theta) \qquad \Theta_S(w) = \arg\max_{\theta \in \Theta} r_S(w).$$

We define the capacities C_e and C_p as:

$$C_e(\Theta) = \max_S \max_{w \in W} \operatorname{dist}(0, \Theta_S(w)) \qquad C_p(\Theta) = \max_{w \in W} \operatorname{dist}(0, \Theta(w)).$$

For the specific constraint set in Problem 1, we denote the capacities as C_e and C_p respectively.

The norm of the parameter is usually viewed as the metric for the complexity of learning models. In fact, the norm of the optimal solution determines the Rademacher complexity of the function class in statistical learning theory [Vapnik, 1999]. Moreover, in deep learning, minimum-norm solution of overparameterized neural networks is well-known to enjoy better generalization performance [Zhang et al., 2021a]. Hence, we view the capacity constant C_e and C_p as natural metrics to capture the model complexity for the best response of the max learner, when using the empirical data set and population data respectively.

Now, we are ready to discuss the relationship between the stability bound and the generalization error of algorithms in nonconvex-concave minimax problems. All proofs have been deferred to the appendix. We make the following assumptions throughout the paper:

Assumption 3. *The gradient of f is ℓ-Lipschitz-continuous for all z, i.e.,*

$$\|\nabla f(w_1, \theta_1; z) - \nabla f(w_2, \theta_2; z)\| \leq \ell(\|w_1 - w_2\| + \|\theta_1 - \theta_2\|), \forall w_1, w_2 \in W, \quad \forall \theta_1, \theta_2 \in \Theta.$$

Moreover, fixing $w \in W$, the partial gradient $\nabla_\theta f(w, \cdot; z)$ is $\ell_{\theta\theta}$-Lipschitz continuous with respect to θ for all z, i.e., $\|\nabla_\theta f(w, \theta_1; z) - \nabla_\theta f(w, \theta_2; z)\| \leq \ell_{\theta\theta}\|\theta_1 - \theta_2\|, \forall w \in W, \quad \forall \theta_1, \theta_2 \in \Theta$.

Assumption 4. *For any $\Theta_1 \subseteq \Theta$, we assume that f is $L(\Theta_1)$-Lipschitz-continuous with respect to $w \in W, \theta \in \Theta_1$ for all z, i.e., $\|f(w_1, \theta_1; z) - f(w_2, \theta_2; z)\| \leq L(\Theta_1)(\|w_1 - w_2\| + \|\theta_1 - \theta_2\|), \quad \forall w_1, w_2 \in W, \ \forall \theta_1, \theta_2 \in \Theta_1$, and the gradient $\nabla f(w, \theta; z)$ is uniformly bounded as $\|\nabla_{w,\theta} f(w, \theta; z)\| \leq L(\Theta_1)$ for all z and $w \in W, \theta \in \Theta_1$. Moreover, $f(w^*, \cdot; z)$ is L_θ^* Lipschitz continuous with respect to θ where $w^* \in \arg\min_{w \in W} r(w)$. We also define $L := L(B(0, 2C_p + 1) \cap \Theta)$ and $L_r := L(B(0, r) \cap \Theta)$, where $B(v, r)$ denotes the l_2-ball with radius r centered at v.*

Note that we can decompose the generalization error of the primal gap as follows:

$$\begin{aligned}\zeta_{gen}^{PG}(A) &:= E_S E_A[\Delta(w_S^A) - \Delta_S(w_S^A)] = E_S E_A[r(w_S^A) - r_S(w_S^A)] + E_S[\min_{w \in W} r_S(w) - \min_{w \in W} r(w)] \\ &= \zeta_{gen}^{P}(A) + E_S[\min_{w \in W} r_S(w) - \min_{w \in W} r(w)].\end{aligned}$$

Next, we provide a bound on the generalization error for the primal risk $\zeta_{gen}^{P}(A)$. To the best of our knowledge, this is the first bound for $\zeta_{gen}^{P}(A)$ in the nonconvex-concave setting.

Lemma 1. *The generalization error of the primal risk of an $\epsilon-$stable Algorithm A for a min-max problem with concave maximization problem can be bounded by $\zeta_{gen}^{P}(A) \leq \sqrt{8C_p^2 \ell L} \cdot \sqrt{\epsilon}$.*

We show that this dependence on ϵ is tight in the Appendix Section C.2.

Since we already have the generalization error for the primal risk $E_S E_A[r(w_S^A) - r_S(w_S^A)]$ from Lemma 1, we only need to estimate

$$\zeta_{gen}^{PM}(A) = E_S E_A[\min_{w \in W} r_S(w) - \min_{w \in W} r(w)] = E_S[\min_{w \in W} r_S(w) - \min_{w \in W} r(w)] \quad \text{[Primal Min Error]}. \tag{8}$$

The following theorem gives the generalization bound of the primal gap using the upper bound from Lemma 1 and bounding the Primal Min Error in Equation (8). Note that in minimization case, we always have $\zeta_{gen}^{PM} \leq 0$. But this is not always true in minimax case as shown in Example 1. So we need to estimate this error in minimax problems.

Theorem 2. *Suppose Algorithm A is ϵ-stable. The generalization bound of the primal gap is given by*

$$\zeta_{gen}^{PG}(A) \leq \sqrt{8\ell L C_p^2 \epsilon} + 4L_\theta^* C_e/\sqrt{n}.$$

Note that the first term is from the generalization of the primal risk, as shown in Lemma 1, and the second term is an upper bound for the Primal Min Error in Equation (8). We show the tightness of the bound in Section C.4.

3.4 Revisiting Example 1

Recall Example 1 in the Section 3.1. In this example, we have that the primal risk generalizes well, but the solution w_S does not, thereby indicating that the generalization error of the primal risk may not be a good metric to study generalization behavior of minimax problem. Note that in Example 1, we have $\zeta_{gen}^{PG}(A) \geq 0.02$ (implying bad generalization) while $\zeta_{gen}^{P}(A) \leq 0$ (implying generalize well). We already know that this algorithm can not generalize well. Therefore, $\zeta_{gen}^{PG}(A)$ can capture the generalization better than $\zeta_{gen}^{P}(A)$. Therefore, the primal min error ζ_{gen}^{PM} also plays an important role in analysing the generalization for minimax problems. As shown in the appendix (Proposition 3), we have

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \geq 0.005. \tag{9}$$

On the other hand, it is easy to compute that $L_\theta^* = \lambda \log n/\sqrt{n}$ and $C_e = \lambda n \log n$. Therefore, by Theorem 2, we have an upper bound for the primal min error (see Equation (8)): $E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \leq 4L_\theta^* C_e/\sqrt{n} = 4\log n$, which is tight up to a $\log$ factor according to (9). Therefore, the primal gap has a constant generalization error which is consistent with the observation that the solution to the empirical problem does not have good generalization behavior.

3.5 Nonconvex-nonconcave case

In this section, we extend our results to the nonconvex-nonconcave setting. We will show that under certain assumptions on the inner maximization problem, we can derive generalization error bounds for the primal risk and primal gap in terms of algorithmic stability. Due to space limitation, we defer the details and justifications of the assumptions (specifically Assumptions 6 and 7) to Appendix C.5.

Furthermore, we assume that $f(\cdot, \cdot; z)$ is L-Lipschitz[2] continuous in $W \times \Theta$. This, along with Assumptions 6 and 7, allows us to derive the generalization error bounds of the primal risk and primal gap in terms of algorithmic stability.

Lemma 2. *Suppose that Assumption 6 holds. If a minimax learning Algorithm A is an ϵ-stable algorithm, we have $\zeta_{gen}^{P}(A) \leq L\epsilon + \sqrt{L\lambda_p}\sqrt{\epsilon}$.*

Similarly, we can derive the generalization bound for the primal gap given the above assumptions.

Theorem 3. *Suppose Assumptions 6 and 7 hold. Then we have $\zeta_{gen}^{PG}(A) \leq \zeta_{gen}^{P}(A) + \sqrt{L\lambda_e}/\sqrt{n}$.*

The proof of these two results can be found in Appendix C.5.

4 Comparison of GDA and GDMax

In Section 3.3, we provide generalization bounds for the primal gap for any $\epsilon-$ stable algorithm. In this section, we focus on two algorithms in particular – GDA and GDMax. These two algorithms are described in Algorithms 1 and 2 in Appendix D.

[2]Note that this is different from the L defined for the nonconvex-concave case. Here L captures the Lipschitz constant over the whole constraint set. In the nonconvex-concave case, $L = L(B(0, 2C_p + 1))$.

We note that though analyzing the optimization properties of GDA/stochastic GDA for solving the empirical minimax problem is an important topic, our focus in this paper is on studying the generalization behavior of these algorithms. We assume that the empirical version of minimax problem we consider can be solved by GDA and GDMax, i.e., we assume that GDA and GDMax satisfy the following assumption:

Assumption 5. *Let A be a minimax learner, such as GDA or GDMax. Then we assume that A has the following convergence rate: $E_A[r_S(w^t) - \min_{w\in W} r_S(w)] \leq (\phi_A(M(W)) + \phi_A(C_e))/\psi_A(t)$, where $M(W)$ is the maximum of the norms of w, $\phi_A(s)$, $\psi_A(s)$ are nonnegative, increasing functions that tend to infinity as $s \to \infty$.*

For simplicity, throughout this section, we assume that $\|f(w, \theta; z)\| \leq 1$ for all w, θ, and z.

The next theorem provides a bound for the population primal gap $\Delta(w_S^A) := r(w_S^A) - \min_{w\in W} r(w)$. Note that the goal of any algorithm is to make this gap as small as possible.

For an Algorithm A and subsets $W_0 \subseteq W, \Theta_0 \subseteq \Theta$, we define $A(W_0, \Theta_0)$ as the algorithm which restricts Algorithm A to solve (1) under constraint sets W_0 and Θ_0. Specifically, $A(W, \Theta)$ is just A.

Theorem 4. *Let $w_S^{A,t}, \theta_S^{A,t}$ be the t-th iterate generated by Algorithm A using dataset S. Assume that $\{\theta_S^{A,t}\} \subseteq \Theta_0 = \Theta_\theta^A$ for $t \leq T$ with probability $1 - \delta$ (due to the randomness in S) and $B(0, C_p) \subseteq \Theta_\theta^A$. Here $B(v, r)$ denotes the l_2-ball with radius r centered at v. Let $A_0 = A(W, \Theta_0)$. Then after T iterations of Algorithm A, the population primal gap can be bounded as:*

$$E_S[r(w_S^{A,T}) - \min_{w\in W} r(w)] \leq (\phi_{A_0}(M_w) + \phi_{A_0}(C_e(\Theta_\theta^A)))/\psi_{A_0}(T) + \zeta_{gen}^P(A_0) + 4L_\theta^* C_e(\Theta_\theta^A)/\sqrt{n} + \delta,$$

where $\zeta_{gen}^P(A_0) = E_S E_A[r(w_S^{A_0,T}) - r_S(w_S^{A_0,T})]$ is the generalization error of the primal risk of Algorithm A_0.

Remark 1. *Theorem 4 builds a closer connection between generalization behavior and the dynamics of minimax learner A. More restricted the the max learner is, better the the generalization behavior. We make this clear in the comparison of GDA and GDMax by analyzing the three terms in Theorem 4.*

4.1 Analyzing the term $\zeta_{gen}^P(A_0)$

First, we study the generalization bound of the primal risk, i.e., ζ_{gen}^P in Theorem 4. For GDA, we can estimate ζ_{gen}^P by using Lemma 1. Therefore, it suffices to estimate the stability of GDA. We do this and bound $\zeta_{gen}^P(A_0)$ for GDA in the following lemma:

Lemma 3. *Let $c_0 = \max\{\alpha_0, \beta_0\}$, If we use diminishing stepsizes $\alpha_t = \alpha_0/t$ and $\beta_t = \beta_0/t$ for GDA for T iterations, we have the stability bound $\epsilon^{GDA} \leq 2L_{\Theta_\theta^{GDA}} T^{c_0\ell}/(n\ell)$. Therefore, for GDA with diminishing stepsizes $\alpha_0/t, \beta_0/t$ run for T iterations (denoted by GDA_T), the generalization error of the primal risk can be bounded by: $\zeta_{gen}^P(GDA_T) \leq (L_{\Theta_\theta^{GDA}})^{3/2}\sqrt{8C_p^2/\ell}\sqrt{T^{c_0\ell}/n}$.*

However for GDMax, we can not compute a uniform stability bound that vanishes as n goes to infinity. In fact, we can show from the following simple example that ζ_{gen}^P(GDMax) can be a constant independent of n, which means that for the case where $r(w, \theta)$ is nonconvex-concave, the generalization bound for primal risk of GDMax can be undesirable.

Example 3 (Constant generalization error of primal risk for GDMax)**.** *Consider a dataset S with n elements. Define the objective function: $f(w, \theta; z) = \left(\frac{w}{n^2} - z\right)\theta - \frac{\theta^2}{2n}$, where $w \in W = [-n\sqrt{n}, n\sqrt{n}]$, $\theta \in \Theta = \mathbb{R}$ and z is drawn from the uniform distribution over $\{-1/\sqrt{n}, 1/\sqrt{n}\}$. We have*

$$r_S(w) = \frac{n^2}{2}\left(\frac{w}{n^2} - \frac{1}{n}\sum_{i=1}^n z_i\right)^2,$$

and $r(w) = \frac{w^2}{2n^2}$. Therefore, $\min_{w\in W} r(w) = 0$. From the definition of the function f and the sets W and Θ, we have $\ell = 1/n^2$, $L = \mathcal{O}(1/\sqrt{n})$.

Note that one step of GDMax can attain the minimizer of $r_S(w)$ (since it is a one dimensional quadratic problem), i.e., $w_S = n\sum_{i=1}^n z_i$ and $r_S(w_S) = 0$. Furthermore, we have $E_S r(w_S) =$

$E[\frac{(\sum_{i=1}^n z_i)2}{2}] = 1/2 > 0$. *Therefore,* $\zeta_{gen}^P(\textit{GDMax}) = E[r(w_S) - r_S(w_S)] = 1/2 > 0$ *cannot be made small.*

Therefore, from Lemma 3 and Example 3, we see that the bound for the expected Population primal gap contains the term ζ_{gen}^P which cannot be bounded for GDMax, whereas can be bounded for GDA which leads us to conclude that GDA generalizes better than GDMax for such problems. However, it is possible to bound ζ_{gen}^P(GDMax) in certain problems, and in this case the two other terms in Theorem 4 become crucial. We analyze them next.

4.2 Analyzing the other two terms of Theorem 4

As shown in Example 1, sometimes the GDMax can have a good generalization bound of the primal risk. Therefore, we need to analyze the other two terms in Theorem 4, i.e., $(\phi_A(M_w) + \phi_A(C_e(\Theta_\theta^A)))/\psi_A(T)$ and $L_\theta^* C_e(\Theta_\theta^A)/\sqrt{n}$. For these two terms, since L_θ^* is fixed, the constant $C_e(\Theta_\theta^A)$ is the key term which differentiates the performance of different algorithms.

By definition, the constant $C_e(\Theta_\theta^{GDMax})$ for GDMax is nearly C_e (See Definition 6). Therefore, the population primal gap after T steps of GDMax is dominated by C_e if C_e is large. However, the set Θ_θ^{GDA} for GDA can be much smaller than Θ, which implies that $C_e(\Theta_\theta^{GDA})$ can be much smaller than C_e. This phenomenon can be seen from Example 1: If we perform one step of GDMax with primal stepsize 1, we can attain $w^1 = w_S$. Then $E_S[r(w_S^1) - \min_{w\in W} r(w)] \geq 0.005$ from (7). For GDA, we can see that $w^1 = 1$ after one step of GDA with stepsize 1. Therefore, GDA generalizes better than GDMax. Generally, we have the following estimate of $C_e(\Theta_\theta^{GDA})$.

Lemma 4. *Let* $L_0 = \max_z \|\nabla f(w_0, \theta_0; z)\|$. *Let* $c_0 = \max\{\alpha_0, \beta_0\}$, *If we use diminishing stepsizes* $\alpha_t = \alpha_0/t$ *and* $\beta_t = \beta_0/t$ *for GDA, then after* T *steps we have* $\|\theta^t\| \leq T^{c_0\ell} L_0/\ell$ *for* $t \in [T]$.

Therefore, if C_e is much larger than C_p, using GDA with $C_p \leq T^{c_0\ell} L_0/\ell \leq C_e$ is better than GDMax. We make this more concrete in the context of GAN training next.

4.3 GAN training

We now study the specific case of GAN training to explore why GDA might generalize better than GDMax. This is numerically verified in the literature, such as Farnia and Ozdaglar [2021]. Specifically, we revisit Example 2, and consider a special case: D is restricted to be a over-parametrized linear function with respect to θ. Define the descriminator $D(x) = \Phi^T(x)v + b_0$, where $\Phi(x) = [\Phi_1(x), \cdots, \Phi_m(x)]^T \in \mathbb{R}^m$ is the feature matrix and $b_0 \in \mathbb{R}$. Also suppose that G is parametrized by w and $G^* = G_{w^*}$. Then the GAN problem can be written as $\min_{w\in W} \max_{\theta\in\Theta} r(w, \theta)$, where

$$r(w, \theta) = E_{x\sim P_r}[\phi(v^T\Phi(x) + b_0)] + E_{y\sim P_0}[\phi(1 - v^T\Phi(G_w(y)) - b_0)].$$

Here $\theta = (v, b_0)$. Assume that $\sqrt{\sigma_{\max}\left(E_{x\sim P_{G_w}}\Phi(x)\Phi^T(x)\right)} \leq \bar{\sigma}_{\max}/\sqrt{m}$, where $\sigma_{\max}(\cdot)$ denotes the largest singular value of a matrix and $\bar{\sigma}_{\max} > 0$ is a constant. Also assume that $E_{x\sim P_{G_w}}\Phi(x)\Phi^T(x)$ is full rank. Also, we assume that $|\phi'(\lambda)| \leq L_\phi$ for any $\lambda \in [0, 1]$. Therefore, we have $E[\|\nabla_\theta f(w, \theta; z)\|^2] \approx L_\phi^2 \bar{\sigma}_{\max}^2$. Then it is reasonable to assume that $\|\nabla f\| \leq \mathcal{O}(1)$.

Lemma 5. *Suppose* $\Phi(x)$ *is sub-Gaussian and the matrix*

$$Q_S = [\Phi(x_1) \quad \Phi(x_2) \cdots \quad \Phi(x_n) \quad \Phi(G_w(y_1)) \quad \cdots \quad \Phi(G_w(y_n))]$$

is full column rank ($m > n$) *with probability* 1. *Then with probability at least* $1 - C\delta$ *with some constant* C, *we have* $\|\theta_S(w^*)\| \geq \Omega(\sqrt{n})$, *where* $\theta_S(w^*) \in \arg\max_{\theta\in\Theta} r_S(w^*, \theta)$.

Now, for $\theta \in \arg\max_{\theta'\in\Theta} r(w^*, \theta')$, it can be easily seen that $v = 0, b_0 = 1/2$ in this case. Therefore, $C_p \approx 1/2$. Finally, combining the previous discussion on GDA in Lemma 4, and using the fact that C_e is large from Lemma 5, we see from Theorem 4 that GDA can generalize better than GDMax. More detailed discuss of the GAN-training example and Lemma 5 can be found in Section D.

References

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pages 224–232. PMLR, 2017.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183. PMLR, 2019.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in GANs. *arXiv preprint arXiv:1806.10586*, 2018.

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

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064, 1997.

Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3573–3579. IEEE, 2020.

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

Farzan Farnia, Jesse M Zhang, and David Tse. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding GANs in the LQG setting: Formulation, generalization and stability. *IEEE Journal on Selected Areas in Information Theory*, 1(1):304–311, 2020.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 33:20766–20778, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020b.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of the proximal point method for nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*, 2020.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *33rd International Conference on Machine Learning, ICML 2016*, pages 1868–1877. International Machine Learning Society (IMLS), 2016.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction*, volume 2. Springer, 2009.

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

Minhui Huang, Kaiyi Ji, Shiqian Ma, and Lifeng Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684*, 2022.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. *arXiv preprint arXiv:2105.03793*, 2021.

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

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

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050:9, 2017.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. 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*, pages 1497–1507. PMLR, 2020a.

Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Renato DC Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Y Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

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

Dmitrii M Ostrovskii, Babak Barazandeh, and Meisam Razaviyayn. Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*, 2021a.

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

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex concave min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 31, 2018.

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

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

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2019.

Bingzhe Wu, Shiwan Zhao, Chaochao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. Generalization in generative adversarial networks: A novel perspective from privacy protection. *Advances in Neural Information Processing Systems*, 32, 2019.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. *arXiv preprint arXiv:2112.05604*, 2021.

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

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

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

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

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. *arXiv preprint arXiv:1711.02771*, 2017.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Materials for
"What is a Good Metric to Study Generalization of Minimax Learners?"

A Related Work

Algorithms for minimax optimization. There is a vast literature on algorithms for minimax optimization. The most popular algorithms include the Extragradient (EG), the Optimistic Gradient Descent Ascent (OGDA) and the Gradient Descent Ascent(GDA) and their variants. The EG algorithms introduced in Korpelevich [1976], has been analyzed in several papers including Monteiro and Svaiter [2010], Mokhtari et al. [2020a,b], Golowich et al. [2020b] for (strongly)convex-(strongly)concave problems. Another popular algorithm is the optimistic gradient descent ascent (OGDA) introduced in Popov [1980] and has been analyzed in several recent works including Daskalakis et al. [2017], Hsieh et al. [2019], Golowich et al. [2020a]. Once again, all these works focus on the (strongly)convex-(strongly)concave setting. Stochastic versions of these algorithms in similar settings have also been analyzed in several papers including Nemirovski et al. [2009], Hsieh et al. [2019], Fallah et al. [2020]. A few papers including Lin et al. [2020], Zhang et al. [2020], Huang et al. [2022], Zhang et al. [2021c], Ostrovskii et al. [2021b], Kong et al. [2019], Zhang et al. [2020] analyze gradient based algorithms in the nonconvex-(strongly)concave cases. Some papers including Rafique et al. [2018], Yang et al. [2021], Ostrovskii et al. [2021a], Grimmer et al. [2020] analyze special cases of nonconvex-nonconcave (like nonconvex-PL) for algorithms like GDA and its variants. However, in this paper, we are interested in the generalization performance of these algorithms. We summarize below the most related literature that studies the generalization behavior in minimax optimization problems.

Algorithm-independent generalization. Specific to the machine learning problems of GAN and adversarial training, there have been several papers studying the uniform convergence generalization bounds. [Arora et al., 2017] establish a uniform convergence generalization bound which depends on the number of discriminator parameters. [Wu et al., 2019] connect the stability-based theory to differential privacy ([Shalev-Shwartz et al., 2010]) in GANs and numerically study the generalization behavior in GANs. [Zhang et al., 2017, Bai et al., 2018] analyze the Rademacher complexity of the players to show the uniform convergence bounds for GANs. In the simpler Gaussian setting, [Feizi et al., 2020] and [Schmidt et al., 2018] derive bounds for GANs and adversarial training, respectively. The uniform convergence bounds for adversarial training have also been studied under several statistical learning frameworks, e.g., PAC-Bayes [Farnia et al., 2018], Rademacher complexity [Yin et al., 2019], margin-based [Wei and Ma, 2019], and VC analysis [Attias et al., 2019]. Recently, [Zhang et al., 2021b] investigate the generalization of empirical saddle point (ESP) solution in strongly-convex-concave problems using a stability-based approach. Note that these results are not specific to the optimization algorithms being used.

Algorithm-dependent generalization. Algorithm specific generalization bounds for minimax optimization have attracted increasing attention. Based on the algorithmic stability framework in [Bousquet and Elisseeff, 2002], [Farnia and Ozdaglar, 2021] has established generalization bounds of standard gradient descent-ascent and proximal point algorithms under the strongly-convex-strongly-concave setting, and those of stochastic GDA and GDMax under the nonconvex-strongly concave setting. Concurrently, [Lei et al., 2021] derives high-probability generalization bounds for both strongly-convex-strongly-concave and convex-strongly-concave settings, with possibly nonsmooth objectives, also through the lens of algorithmic stability. Recently, Li and Liu [2021] improves the high probability generalization bounds in terms of different stability measures and gives a comprehensive discussion on the generalization bound for different algorithms. These works hinged on the metrics of *primal risk* and *primal-dual risk*. As shown in the present work, the former is not necessarily suitable to characterize the generalization behavior of minimax optimization, while the latter might not make sense when the saddle point does not exists, i.e., the $\max$ and $\min$ cannot be interchanged, a common scenario in machine learning (see Jin et al. [2020]). Following this line of work, [Xing et al., 2021] provide generalization bounds specifically for adversarial training, which is essentially the primal risk, also using the algorithmic stability framework. Recently, [Yang et al., 2022] study the generalization of stochastic GDA under differential privacy constraints.

Reference	Assumption	Metric	Rate
[Farnia and Ozdaglar, 2021]	NC-μ-SC	PR	$L\sqrt{\kappa^2+1}\epsilon$
[Lei et al., 2021]	NC-μ-SC	PR	$L(1+\kappa)\epsilon$
[Lei et al., 2021]	μ-SC-SC	PD	$\sqrt{2}L(1+\kappa)\epsilon$
This work (Theorem 2)	NC-C	PG	$\sqrt{4L\ell C_p^2}\cdot\sqrt{\epsilon}+\epsilon L+4L_\theta^* C_e/\sqrt{n}$
This work (Lemma 1)	NC-C	PR	$\sqrt{4L\ell C_p^2}\cdot\sqrt{\epsilon}+\epsilon L$
This work (Theorem 7)	C-C	PD	$\left(\sqrt{4L\ell C_p^2}+\sqrt{4L\ell(C_p^w)^2}\right)\sqrt{\epsilon}+2\epsilon L$

Table 1: Generalization bounds for ϵ-stable algorithms. PR stands for Primal Risk, PD stands for the primal-dual risk and PG stands for the primal gap. NC-μ-SC stands for nonconvex-μ strongly concave. μ-SC-SC stands for μ strongly convex-μ strongly concave. NC-C stands for nonconvex-concave. C-C convex-concave. L is the Lipschitz constant of the function f. κ stands for the condition number L/μ. The constants in the theorems have been defined in the appropriate sections. Note that there are other results in [Farnia and Ozdaglar, 2021, Lei et al., 2021] for cases where the expectation and $\max$ operator can be interchanged. This case is almost identical to the minimization problem and we thus do not include it in the table.

A.1 Existing related results

From [Farnia and Ozdaglar, 2021], we have the following theorem showing the connection between stability and generalization for minimax problems.

Theorem 5 ([Farnia and Ozdaglar, 2021])**.** *Consider an Algorithm A which is ϵ-stable. We have the following two claims:*

1. *If the maximization and the expectation can be swapped when computing $r(w)$, then*
$$E_S E_A[\zeta_{gen}^P(A)] \leq \epsilon.$$
2. *If $f(\cdot,\cdot;z)$ is nonconvex-strongly-concave and f is μ-strongly-concave with respect to θ, then*
$$E_S E_A[\zeta_{gen}^P(A)] \leq L\sqrt{\kappa^2+1}\epsilon.$$

Remark 2. *In [Lei et al., 2021], the authors proved a generalization bound in a weak sense, i.e., they consider the weak duality gap:*

$$(\max_{\theta\in\Theta} E_S E_A r(w_S^A,\theta) - \min_{w\in W} E_S E_A r(w,\theta_S^A)) - (\max_{\theta\in\Theta} E_S E_A r_S(w_S^A,\theta) - \min_{w\in W} E_S E_A r_S(w,\theta_S^A)).$$

However, notice that the expectation is inside the min and max operators. It does not deal with the coupling of the maximization and expectation.

Remark 3. *According to Theorem 5, the generalization bound for ζ_{gen}^P scales with the condition number κ_θ, and therefore cannot give useful bounds in the absence of strong concavity (when $\kappa_\theta \to \infty$).*

Remark 4. *The generalization bounds for ζ_{gen}^P of algorithms for problems in terms of stability without strong concavity is still open to the best of our knowledge. As mentioned in [Lei et al., 2021], finding generalization bounds without the strong concavity assumption is an interesting open problem.*

B Analysis of Example 1

In this section, we analyze the toy example given in Example 1.

Proposition 1. *For the risk function and data distribution given in Example 1, we have*
$$E_S[r(w) - r_S(w)] \leq 0$$
for any $w \in W$.

Proof. For a fixed w, $r(w) = w^2/2 - w$. On the other hand,
$$\begin{aligned} r_S(w) &= \max_{\theta\in\Theta} r(w,\theta) && (10)\\ &\geq r_S(w,0) && (11)\\ &= r(w). && (12)\end{aligned}$$
Therefore, we have the desired result. □

Next, we prove that $|\sum_{i=1}^n z_i|$ will stay in the interval $[0.5, \lambda]$ with high probability.

Lemma 6. *For large enough $\lambda > 2$, we have*

$$\Pr\left(\left|\sum_{i=1}^n z_i\right| \in [0.5, \lambda]\right) > 0.4, \qquad \Pr\left(\left|\sum_{i=1}^n z_i\right| \in [2, \lambda]\right) > 0.01.$$

Proof. Let $y_i \sim N(0, 1/\sqrt{n}), i = 1, \cdots, n$ be n i.i.d. variables. Then $\sum_{i=1}^n y_i \sim N(0, 1)$. According to the table of Normal distribution, we have $\Pr(|\sum_{i=1}^n y_i| \in [0.5, \lambda]) \geq 0.41$. By the definition of z_i, we have

$$\Pr(|\sum_{i=1}^n z_i| \in [0.5, \lambda]) \geq \Pr(|\sum_{i=1}^n y_i| \in [0.5, \lambda], |y_i| < 3\log n/\sqrt{n}) + \Pr(\max_{i\in[n]}(|y_i|) \geq 3\log n/\sqrt{n}).$$

For the first term, we have

$$\begin{aligned}
& \Pr(|\sum_{i=1}^n y_i| \in [0.5, \lambda], |y_i| < 3\log n/\sqrt{n}) && (13)\\
\geq\ & \Pr(|\sum_{i=1}^n y_i| \in [0.5, \lambda]) - \Pr(\max_{i\in[n]}(|y_i|) \geq 3\log n/\sqrt{n}) && (14)\\
\geq\ & 0.41 - \sum_{i=1}^n \Pr(|y_i| \geq 3\log n/\sqrt{n}) && (15)\\
\geq\ & 0.41 - ne^{-\gamma 9\log^2 n} && (16)\\
\geq\ & 0.41 - 1/n^{\lambda\gamma - 1}. && (17)
\end{aligned}$$

Taking λ sufficiently large yields the desired result, where the first inequality is because of the union bound and the second inequality is due to the tail bound of Normal distribution. Therefore, $\Pr(|\sum_{i=1}^n z_i| \in [0.5, \lambda]) > 0.4$ for sufficiently large n. The second statement follows similarly, noting from the table of Normal distribution that $\Pr(|\sum_{i=1}^n y_i| \in [0.5, \lambda]) \geq 0.046$. □

Proposition 2. *For sufficiently large $\lambda > 0$, we have*

$$E_S[r(w_S) - \min_{w\in W} r(w)] \geq 0.001.$$

Proof. If $|\sum_{i=1}^n z_i| \in [0.5, \lambda]$, we have

$$w_S = \max(0, 1 - (\sum_{i=1}^n z_i)^2/2) \leq 0.9.$$

In this case, we have

$$r(w_S) - \min_{w\in W} r(w) \geq 0.005, \tag{18}$$

by direct calculation. Therefore, we have

$$\begin{aligned}
& E_S[r(w_S) - \min_{w\in W} r(w)] && (19)\\
\geq\ & \Pr(|\sum_{i=1}^n z_i| \in [0.5, \lambda]) \cdot 0.05 + \Pr(|\sum_{i=1}^n z_i| \notin [0.5, \lambda]) \cdot 0 && (20)\\
\geq\ & 0.02, && (21)
\end{aligned}$$

where the first inequality is because of (18) and the fact that $r(w_S) - \min_{w\in W} r(w) \geq 0$ for any S. □

Proposition 3. *For sufficiently large $\lambda > 0$, we have:*

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \geq 0.005$$

for Example 1.

Proof. If $|\sum_{i=1}^n z_i| \geq \lambda > 2$, we have $w_S = 0$ and hence $r_S(w_S) = 0$. If $|\sum_{i=1}^n z_i| \leq \lambda$, we have

$$r_S(w_S) - r(w^*) \geq r_S(w_S) - r(w_S) = w_S(\sum_{i=1}^n z_i)^2/2 \geq 0.$$

Therefore, $\min_{w\in W} r_S(w) \geq \min_{w\in W} r(w)$ for any S. By Lemma 6, we can prove that $\Pr(|\sum_{i=1}^n z_i| \in [2,\lambda]) \geq 0.01$ for sufficiently large λ. Notice that for $|\sum_{i=1}^n z_i| \in [2,\lambda]$, $r_S(w_S) - \min_{w\in W} r(w) = 1/2$. Therefore, we have

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \geq \Pr(|\sum_{i=1}^n z_i| \in [2,\lambda]) \cdot 1/2 \geq 0.005.$$

This completes the proof. □

C Deferred Results and Proofs in Section 3

C.1 Proof of Lemma 1

In this subsection, we assume that A is an ϵ-stable algorithm. For any $w \in W$, let $\Theta_S(w) = \arg\max_{\theta\in\Theta} r_S(w,\theta)$ and $\Theta(w) = \arg\max_{\theta\in\Theta} r(w,\theta)$ be the solution sets of the problems. Let $\theta(w)$ be any element in $\Theta(w)$. Then

$$\begin{aligned} E_A E_S[r(w_S^A) - r_S(w_S^A)] &= E_A E_S[r(w_S^A, \theta(w_S^A)) - r_S(w_S^A, \theta_S(w_S^A))] \\ &\leq E_A E_S[r(w_S^A, \theta(w_S^A)) - r_S(w_S^A, \theta(w_S^A))], \end{aligned}$$

where the inequality is because $r_S(w_S^A, \theta_S(w_S^A)) \geq r_S(w_S^A, \theta)$ for any θ. Let f be μ-strongly concave with respect to θ. We denote the condition number by $\kappa_\theta = \ell_{\theta\theta}/\mu$.

In the strongly concave case, $\Theta(w)$ has a unique element $\theta(w)$, which is κ_θ-Lipschitz continuous with respect to w (see [Lin et al., 2020]).

Then, defining $\tilde{f}(w,z) = f(w,\theta(w);z)$, the minimax problem reduces to the usual minimization problem on the function $\tilde{f}$. The stability and the Lipschitz continuity of $\theta(w)$ with respect to w yield the generalization bound of $L\sqrt{\kappa^2+1}\epsilon$. This is the result shown in Theorem 1 of [Farnia and Ozdaglar, 2021].

However, if the maximization problem is not strongly concave, we lose the Lipschitz continuity and the uniqueness. To overcome this difficulty, we define an approximate maximizer $\bar{\theta}(w)$ to $r(w,\theta)$. Concretely speaking, we define $\bar{\theta}(w)$ to be the point after s steps of gradient ascent for the function $r(w,\cdot)$ with a stepsize $1/\ell_{\theta\theta}$ and being initialized at 0. Then we have the following lemma:

Lemma 7. *For any $w \in W$, we have*

1. $\|\bar{\theta}(w) - \bar{\theta}(w')\| \leq s\frac{\ell}{\ell_{\theta\theta}}\|w - w'\|$.
2. $r(w) - r(w,\bar{\theta}(w)) \leq \ell_{\theta\theta}C_p^2/s$.

Proof. To prove the first part, let $\theta_0 = \theta_0' = 0$. Define θ_t, θ_t' recursively as follows:

$$\theta_{t+1} = \theta_t + \nabla_\theta r(w,\theta_t)/\ell_{\theta\theta}$$

and

$$\theta_{t+1}' = \theta_t' + \nabla_\theta r(w',\theta_t')/\ell_{\theta\theta}.$$

We prove $\|\theta_t - \theta'_t\| \leq t\frac{\ell}{\ell_{\theta\theta}}\|w - w'\|$ by induction. For $t = 0$, $\theta_0 - \theta'_0 = 0$. Assume the induction hypothesis $\|\theta_{t-1} - \theta'_{t-1}\| \leq (t-1)\frac{\ell}{\ell_{\theta\theta}}\|w - w'\|$ holds. We have

$$\begin{aligned}
\|\theta_t - \theta'_t\| &= \|(\theta_{t-1} + \nabla_\theta r(w, \theta_{t-1})/\ell_{\theta\theta}) - (\theta'_{t-1} + \nabla_\theta r(w, \theta'_{t-1})/\ell_{\theta\theta}) \\
&\qquad + (\nabla_\theta r(w, \theta'_{t-1}) - \nabla_\theta r(w', \theta'_{t-1}))/\ell_{\theta\theta}\| \\
&\leq \|(\theta_{t-1} + \nabla_\theta r(w, \theta_{t-1})/\ell_{\theta\theta}) - (\theta'_{t-1} + \nabla_\theta r(w, \theta'_{t-1})/\ell_{\theta\theta})\| \\
&\qquad + \|(\nabla_\theta r(w, \theta'_{t-1}) - \nabla_\theta r(w', \theta'_{t-1}))/\ell_{\theta\theta}\| \\
&\leq \|\theta_{t-1} - \theta'_{t-1}\| + \ell\|w - w'\|/\ell_{\theta\theta} \\
&\leq (t-1)\frac{\ell}{\ell_{\theta\theta}}\|w - w'\| + \frac{\ell}{\ell_{\theta\theta}}\|w - w'\| \\
&= t\frac{\ell}{\ell_{\theta\theta}}\|w - w'\|,
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from non-expansiveness of gradient ascent for concave functions and the ℓ-Lipschitz continuity of ∇r, and the third inequality follows from the induction hypothesis.

Therefore, letting $t = s$ completes the proof of the first part. The second part of this lemma is just the convergence result for gradient ascent on smooth concave functions (see e.g., [Nesterov, 2013]). □

Consider a virtual algorithm $\bar{A}$: for any S, the algorithm returns $w = w_S^A$ and $\theta = \bar{\theta}(w_S^A)$.

Lemma 8. *The stability of this virtual algorithm is* $\epsilon\sqrt{\left(s\frac{\ell}{\ell_{\theta\theta}}\right)^2 + 1}$.

Proof. It is direct from the first part of Lemma 7. □

Then we have the generalization bound of $r_S(w, \theta)$:

Lemma 9. *We have*

$$E_S E_A[r(w_S^A, \bar{\theta}(w_S^A)) - r_S(w_S^A, \bar{\theta}(w_S^A))] \leq \epsilon L\sqrt{\left(s\frac{\ell}{\ell_{\theta\theta}}\right)^2 + 1}.$$

Proof. For any z, by Assumption 4, we have

$$\|f(w_S^{\bar{A}}, \theta_S^{\bar{A}}; z) - f(w_{S'}^{\bar{A}}, \theta_{S'}^{\bar{A}}; z)\| \leq \epsilon L\sqrt{\left(s\frac{\ell}{\ell_{\theta\theta}}\right)^2 + 1}.$$

The result follows directly from the standard stability theory in [Hardt et al., 2016]. □

Now we are ready to derive the generalization bound of the Primal Risk for an Algorithm A with ϵ-stability. First, we have

$$\begin{aligned}
E_S E_A[r(w_S^A) - r_S(w_S^A)] &\leq E_S E_A[r(w_S^A) - r_S(w_S^A, \bar{\theta}(w_S^A))] \\
&\leq E_S E_A[(r(w_S^A, \bar{\theta}(w_S^A) + \ell_{\theta\theta}C_p^2/s) - r_S(w_S^A, \bar{\theta}(w_S^A))] \\
&= E_S E_A[r(w_S^A, \bar{\theta}(w_S^A)) - r_S(w_S^A, \bar{\theta}(w_S^A))] + \ell_{\theta\theta}C_p^2/s \\
&\leq \epsilon L\sqrt{\left(s\frac{\ell}{\ell_{\theta\theta}}\right)^2 + 1} + \ell_{\theta\theta}C_p^2/s,
\end{aligned}$$

where the first inequality is because $r_S(w_S^A) = \max_\theta r_S(w_S^A, \theta)$, the second inequality is because of the second part of Lemma 7 and the last inequality is because of Lemma 9. Optimizing over s, the generalization error is bounded by $\zeta_{gen}^P(A) \leq \sqrt{8L\ell C_p^2} \cdot \sqrt{\epsilon}$. This completes the proof. □

C.2 Tightness of the bound for Primal Risk

Consider the following risk function: $f(w,\theta;z) = \sqrt{n/\epsilon}((w/(n\sqrt{n}\epsilon) - z)\theta - \theta^2/(2n\sqrt{n}\epsilon))$, where $w \in W = [-\lambda\epsilon\sqrt{n}\log n, \lambda\epsilon\sqrt{n}\log n]$ and $\theta \in \Theta = \mathbb{R}$. The sample z is drawn from the uniform distribution over $\{-1/\sqrt{n}, 1/\sqrt{n}\}$. Then we have $r(w) = \sqrt{n/\epsilon}(w^2/(2\epsilon n\sqrt{n}))$ and $r_S(w) = \sqrt{n/\epsilon}((\epsilon n\sqrt{n})(w/(\epsilon n\sqrt{n}) - \sum_{i=1}^n z_i/n)^2/2)$. Now we have $\ell = \frac{\sqrt{n/\epsilon}}{(\epsilon n\sqrt{n})}$, $C_p = \lambda\epsilon\sqrt{n}\log n$ and $L = \frac{\sqrt{n/\epsilon}}{\sqrt{n}}$. If we perform one-step of GDMax with stepsize $1/\ell_{r_S}$ where $\ell_{r_S} = \sqrt{n/\epsilon}/(\epsilon n\sqrt{n})$, then we attain $w_S = \arg\min_{w\in W} r_S(w)$. The stability bound of the GDMax is ϵ. Therefore, the generalization error of the primal risk is estimated as:

$$\zeta_{gen}^P(GDMax) \le \sqrt{8L\ell C_p^2}\sqrt{\epsilon} = 8\lambda\log n\sqrt{\epsilon}.$$

On the other hand, $\sum_{i=1}^n z_i/n \in [-\lambda\log n/n, \lambda\log n/n]$ holds with probability at least $1 - C/n^\lambda$ by Hoefding inequality. Let $\bar{w}_S = \epsilon\sqrt{n}(\sum_{z_i\in S} z_i)$. Then with probability at least $1 - C/n^\lambda$, $w_S = \bar{w}_S$ Notice that $E_S[r(\bar{w}_S) - r_S(\bar{w}_S)] = E_S[\sqrt{n}\cdot(\sqrt{\epsilon}n\sqrt{n})\cdot(\sum_{z_i\in S} z_i/n)^2] = \sqrt{\epsilon}$. It is not hard to show that $|r(\bar{w}_S)| \le n\sqrt{\epsilon}$, $r_S(\bar{w}_S) = 0$, $|r(w_S)| \le 2n\sqrt{\epsilon}$ and $|r_S(w_S)| \le 2n\sqrt{\epsilon}$. Then we have

$$E_S[r(\bar{w}_S) - r_S(\bar{w}_S)] - E_S[r(w_S) - r_S(w_S)] \le 5Cn\sqrt{\epsilon}/n^\lambda.$$

Therefore, $E[r(w_S) - r_S(w_S)] \ge \sqrt{\epsilon}/2$ for sufficiently large λ and n. Then in this example we have

$$\sqrt{\epsilon}/2 \le \zeta_{gen}^P(A) \le 8\lambda\log n\sqrt{\epsilon}.$$

For $\epsilon \le 1/n^{\tau+1}$, we have

$$\log n \le \frac{1}{\tau+1}\log(1/\epsilon).$$

Therefore, the estimate $\zeta_{gen}^P \le \lambda\sqrt{\epsilon}\log(1/\epsilon)/(\tau+1)$ is tight up to a $\log(1/\epsilon)$ factor.

C.3 Proof of Theorem 2

Recall that the empirical primal gap is defined as

$$\Delta_S(w) = r_S(w) - \min_{w\in W} r_S(w)$$

and the population primal gap is given by

$$\Delta(w) = r(w) - \min_{w\in W} r(w).$$

Suppose we are given an ϵ-stable Algorithm A. We then want to derive the generalization error

$$\zeta_{gen}^{PG}(A) = E_S E_A[\Delta(w_S^A) - \Delta_S(w_S^A)].$$

Since we already have the generalization error for the primal risk $E_S E_A[r(w_S^A) - r_S(w_S^A)]$ in Theorem 1, we only need to estimate

$$E_S E_A[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] = E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)]$$

to get a generalization bound on the primal gap.

Lemma 10. *Let $w^* \in \arg\min_{w\in W} r(w)$. Suppose that $f(w^*,\cdot;z)$ is L_θ^* Lipschitz continuous with respect to θ. Then we have*

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \le \sqrt{2L_\theta^* C_e^2}/\sqrt{n}.$$

Proof. We use similar techniques as in the proof of Lemma 1.

Step 1. We define an approximate maximizer $\tilde{\theta}_S$ of the function $r_S(w^*,\cdot)$. $\tilde{\theta}_S$ is attained by performing s steps of gradient ascent to $r_S(w^*,\cdot)$ with stepsize $1/\ell_{\theta\theta}$ and being initialized at 0.

Similar to Lemma 7, we have the following lemma:

Lemma 11. *We have the following properties:*

1. $\|\tilde{\theta}_S - \tilde{\theta}_{S'}\| \leq 2sL_\theta^*/(n\ell_{\theta\theta})$.
2. $r_S(w^*) - r_S(w^*, \tilde{\theta}_S) \leq \ell_{\theta\theta} C_e^2/s$.

Proof. The proof is similar to the proof of Lemma 7. To prove the first part, let $\tilde{\theta}_0 = \tilde{\theta}'_0 = 0$. Define $\tilde{\theta}_t, \tilde{\theta}'_t$ recursively as follows:

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t + \nabla_\theta r_S(w^*, \tilde{\theta}_t)/\ell_{\theta\theta}$$

and

$$\tilde{\theta}'_{t+1} = \tilde{\theta}'_t + \nabla_\theta r_{S'}(w^*, \tilde{\theta}'_t)/\ell_{\theta\theta}.$$

We prove $\|\tilde{\theta}_t - \tilde{\theta}'_t\| \leq L_\theta^*/(n\ell_{\theta\theta})$ by induction. For $t = 0$, $\tilde{\theta}_0 - \tilde{\theta}'_0 = 0$. Assume the induction hypothesis $\|\tilde{\theta}_{t-1} - \tilde{\theta}'_{t-1}\| \leq (t-1)L_\theta^*/(n\ell_{\theta\theta})$ holds. We have

$$\begin{aligned}
\|\tilde{\theta}_t - \tilde{\theta}'_t\| &= \|(\tilde{\theta}_{t-1} + \nabla_\theta r_S(w^*, \tilde{\theta}_{t-1})/\ell_{\theta\theta}) - (\tilde{\theta}'_{t-1} + \nabla_\theta r_S(w^*, \tilde{\theta}'_{t-1})/\ell_{\theta\theta}) \\
&\qquad + (\nabla_\theta r_S(w^*, \tilde{\theta}'_{t-1}) - \nabla_\theta r_{S'}(w^*, \tilde{\theta}'_{t-1}))/\ell_{\theta\theta}\| \\
&\leq \|(\tilde{\theta}_{t-1} + \nabla_\theta r_S(w^*, \tilde{\theta}_{t-1})/\ell_{\theta\theta}) - (\tilde{\theta}'_{t-1} + \nabla_\theta r_S(w^*, \tilde{\theta}'_{t-1})/\ell_{\theta\theta})\| \\
&\qquad + \|(\nabla_\theta r_S(w^*, \tilde{\theta}'_{t-1}) - \nabla_\theta r_{S'}(w^*, \tilde{\theta}'_{t-1}))/\ell_{\theta\theta}\| \\
&\leq \|\tilde{\theta}_{t-1} - \tilde{\theta}'_{t-1}\| + \ell\|w - w'\|/\ell_{\theta\theta} \\
&\leq (t-1)\frac{2L_\theta^*}{n\ell_{\theta\theta}} + \frac{2L_\theta^*}{n\ell_{\theta\theta}} \\
&= t\frac{2L_\theta^*}{n\ell_{\theta\theta}},
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from non-expansiveness of gradient ascent for concave functions and the L_θ^*-Lipschitz continuity of $f(w^*, \cdot; z)$, and the third inequality follows from the induction hypothesis.

Therefore, letting $t = s$ completes the proof of the first part. The second part of this lemma is just the convergence result for gradient ascent on smooth concave functions (see e.g., [Nesterov, 2013]). □

We then define the virtual algorithm $\tilde{A}$ given by $w_S^{\tilde{A}} = w^*$ and $\theta_S^{\tilde{A}} = \tilde{\theta}_S$. Since the output argument w of $\tilde{A}$ is always w^*, the stability of $\tilde{A}$ only depends on $\tilde{\theta}_S$. Then the stability bound of this virtual algorithm is given in the following lemma:

Lemma 12. *The stability of Algorithm $\tilde{A}$ is given by $\epsilon_{sta}(\tilde{A}) = 2sL_\theta^*/(n\ell_{\theta\theta})$.*

Then by the standard stability theory in [Hardt et al., 2016], we have

$$|E_S E_A[r_S(w^*, \tilde{\theta}_S) - r(w^*, \tilde{\theta}_S)]| \leq 2s(L_\theta^*)^2/(n\ell_{\theta\theta}). \tag{22}$$

Step 2. We have

$$\begin{aligned}
E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] &\overset{\text{(i)}}{=} E_S[r_S(w_S) - r(w^*, \theta^*)] \\
&\overset{\text{(ii)}}{\leq} E_S[r_S(w^*) - r(w^*, \theta^*)] \\
&\overset{\text{(iii)}}{\leq} E_S[r_S(w^*, \tilde{\theta}_S) - r(w^*, \theta^*)] + \ell_{\theta\theta} C_e^2/s \\
&\overset{\text{(iv)}}{\leq} E_S[r_S(w^*, \tilde{\theta}_S) - r(w^*, \tilde{\theta}_S)] + \ell_{\theta\theta} C_e^2/s,
\end{aligned}$$

where (i) follows from the definition of w^*, θ^*, (ii) follows since w_S minimizes $r_S(w)$, (iii) follows from Lemma 11, and (iv) follows from the optimality of θ^* given w^*. Then by (22), we have

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \leq E_S[r_S(w^*, \tilde{\theta}_S) - r(w^*, \tilde{\theta}_S)] + \ell_{\theta\theta} C_e^2/s \tag{23}$$

$$\leq 2sL_\theta^*/(n\ell_{\theta\theta}) + \ell_{\theta\theta} C_e^2/s \tag{24}$$

$$\leq 4L_\theta^* C_e/\sqrt{n} \tag{25}$$

which completes the proof. □

The final statement of the theorem follows from Lemma 10 and Theorem 1. □

C.4 Tightness of Bound for Primal Min Error

We can construct an example with C_e,L_θ^* independent of n and the upper bound for ζ_{gen}^{PM} is tight up a $\log$ factor.

Consider the following example:

- Let $f(w,\theta;z) = \frac{1}{M}(w^2/2 + w(\log^2 n\theta^2/(2K^2) + n\log nz\theta/K + 1))$, where $w \in W = [-1,1]$, $\theta \in \Theta = [-\lambda K, \lambda K]$ for some arbitrary constants $K > 0$ and $M > 0$. z is drawn from a truncated Gaussian distribution. Concretely speaking, let $y \sim N(0, 1/\sqrt{n})$. Then $z = y$ if $|y| \leq \lambda \log n/\sqrt{n}$ and $z = \lambda \log n/\sqrt{n}$ if $y \geq \lambda \log n/\sqrt{n}$. Notice that $Mf(w,\theta;z) = h(w,\theta';z)$, where $\theta' = n\log n\theta/K$, $h(w,\theta';z) = w^2/2 + w((\theta')^2/(2n^2) + z\theta' + 1)$, $\theta' = n\log n\theta/K \in [-\lambda n\log n, \lambda n \log n]$. Notice that h is just the risk function in Example 1 in the paper. Therefore, we can estimate the lower bound of the Primal Min Error corresponding to f using the result in Example. The lower bound of Primal Min Error of the problem in Example 1 (corresponding to h) is 0.005. Then the Primal-Min Error (corresponding to f) $\zeta_{gen}^{PM}(A) = E_S \min_{w\in W} r_S(w) - \min_{w\in W} r(w) \geq 0.005/M$. On the other hand, it is not hard to have $L_\theta^* = \lambda\sqrt{n}\log^2 n/MK$, $C_e = \lambda K$. Therefore, $L_\theta^* C_e = \lambda^2\sqrt{n}\log^2 n/M$. Let $M = \sqrt{n}\log^2 n$. Then $L_\theta^* = \lambda/K$.
  If K does not depend on n, L_θ^*, C_e do not depend on n. Therefore, the Primal-Min Error $\zeta_{gen}^{PM}(A)$ satisfies $0.005\lambda^2(L_\theta^* C_e)/(\log^2 n\sqrt{n}) = 0.005(L_\theta^* C_e)/(ML_\theta^* C_e) \leq \zeta_{gen}^{PM}(A) \leq (L_\theta^* C_e)/\sqrt{n}$, which is tight up to a $\log^2 n$.
- If we let $M = 1$ and $K = 1$, then $\zeta_{gen}^{PM}(A) \leq \lambda^2 \log^2 n$ as discussed in the paper. This upper bound means that we can not attain an arbitrary accuracy $\delta > 0$. The lower bound, i.e., $\zeta_{gen}^{PM}(A) \geq 0.005\lambda^4$, also implies that we can not attain an arbitrary accuracy.
- If we let $M = 1/(\lambda^2 \log^2 n)$ and $K = 1$, $L_\theta^* C_e/\sqrt{n} = 1$. This upper bound implies that we can not let ζ_{gen}^{PM} smaller than arbitrariy required accuracy. Return to the lower bound, i.e., $\zeta_{gen}^{PM}(A) \geq 0.005/(\lambda^2 \log^2 n)$. If we want to attain an accuracy δ, the required sample complexity is $2^{\lambda/\sqrt{\delta}}$, which is larger than a polynomial size and hence is still viewed as intractable. In this sense, the upper bound and the lower bound do not make a major difference.
- Combining the two points above, we can conclude that in terms of sample complexity, our bound is tight (up to logarithmic factors).

C.5 Results in §3.5 for the Nonconvex-Nonconcave Case

For the nonconvex-nonconcave setting, we assume that the maximization problem can be solved by gradient ascent for any w. Note that this includes the case where the maximization problem has the gradient-domination structure (see for example Rafique et al. [2018]).

We first make the following assumptions on the inner maximization problem:

Assumption 6. *For any $\gamma > 0$, there exists an algorithm which outputs $\theta_P^\gamma(w)$, for the inner maximization problem $\max_{\theta\in\Theta} r(w,\theta)$, satisfying the following conditions:*

1. $r(w) - r(w, \theta_P^\gamma(w)) \leq \gamma$.
2. $\|\theta_P^\gamma(w) - \theta_P^\gamma(w')\| \leq \frac{\lambda_p}{\gamma}\|w - w'\|$ *with some constant* $\lambda_p > 0$ *for all* $w, w' \in W$.

Assumption 7. *For any $\gamma > 0$, there exists an algorithm which outputs $\theta_E^\gamma(S)$, for the inner maximization problem $\max_{\theta\in\Theta} r_S(w^*,\theta)$, satisfying the following conditions:*

1. $r_S(w^*) - r_S(w^*, \theta_E^\gamma(S)) \leq \gamma$.
2. *For any neighboring dataset S, S', we have $\|\theta_E^\gamma(S) - \theta_E^\gamma(S')\| \leq \frac{\lambda_e}{n\gamma}$ with some constant $\lambda_e > 0$.*

The following lemma gives sufficient conditions for these two assumptions to hold.

Lemma 13. *Consider constants $D_e \geq \gamma$ and $D_p \geq \gamma$.*

1. *Suppose that gradient ascent with diminishing stepsizes c_0/t for the problem $\max_{\theta\in\Theta} r(w,\theta)$ has convergence rate $r(w) - r(w,\theta^s) \leq D_p/s$. Then we define $\theta_p^\gamma(w)$ by performing $s = D_p/\gamma$ steps of gradient ascent. Then, $\theta_p^\gamma(w)$ satisfies Assumption 6.*

2. *Suppose that gradient ascent with constant stepsize c_0 for the problem $\max_{\theta\in\Theta} r(w,\theta)$ has convergence rate $r(w) - r(w,\theta^s) \leq D_p\eta^s$ for some constant $0 < \eta < 1$. Then we define $\theta_p^\gamma(w)$ by $s = \log(D_p/\gamma)/\log(1/\eta)$ steps of gradient ascent. Then, $\theta_p^\gamma(w)$ satisfies Assumption 6.*

3. *Suppose gradient ascent with diminishing stepsizes c_0/t for the problem $\max_{\theta\in\Theta} r_S(w,\theta)$ has convergence rate $r_S(w) - r_S(w,\theta^s) \leq D_p/s$. Then we define $\theta_e^\gamma(S)$ by performing $s = D_e/\gamma$ steps of gradient ascent. Then, $\theta_e^\gamma(S)$ satisfies Assumption 7.*

4. *Suppose gradient ascent with constant stepsize c_0 for the problem $\max_{\theta\in\Theta} r_S(w,\theta)$ has convergence rate $r_S(w) - r_S(w,\theta^s) \leq D_e\eta^s$ for some constant $0 < \eta < 1$. Then we define $\theta_e^\gamma(w)$ by $s = \log(D_e/\gamma)/\log(1/\eta)$ steps of gradient ascent. Then, $\theta_e^\gamma(S)$ satisfies Assumption 7.*

Proof. We only prove the first part of this lemma and the others can be proved similarly. Let $s = [D_p/\gamma] + 1$, where $[r]$ denotes the largest integer no more than r. To prove the first part, let $\theta_0 = \theta_0' = 0$. Define θ_t, θ_t' recursively as follows:

$$\theta_{t+1} = \theta_t + c_0\nabla_\theta r(w,\theta_t)/t$$

and

$$\theta_{t+1}' = \theta_t' + c_0\nabla_\theta r(w',\theta_t')/t.$$

We prove $\|\theta_t - \theta_t'\| \leq t\frac{\ell}{\ell_{\theta\theta}}\|w - w'\|$ by induction. For $t = 0$, $\theta_0 - \theta_0' = 0$. Assume the induction hypothesis $\|\theta_{t-1} - \theta_{t-1}'\| \leq (t-1)\frac{\ell}{\ell_{\theta\theta}}\|w - w'\|$. We have

$$\begin{aligned}
\|\theta_t - \theta_t'\| &= \|(\theta_{t-1} + c_0\nabla_\theta r(w,\theta_{t-1})/t) - (\theta_{t-1}' + c_0\nabla_\theta r(w,\theta_{t-1}')/t) && (26)\\
&\quad + c_0(\nabla_\theta r(w,\theta_{t-1}') - \nabla_\theta r(w',\theta_{t-1}'))/t\| && (27)\\
&\leq \|(\theta_{t-1} + c_0\nabla_\theta r(w,\theta_{t-1})/t) - (\theta_{t-1}' + c_0\nabla_\theta r(w,\theta_{t-1}')/t)\| && (28)\\
&\quad + c_0\|(\nabla_\theta r(w,\theta_{t-1}') - \nabla_\theta r(w',\theta_{t-1}'))/t\| && (29)\\
&\leq (1 + c_0\ell_{\theta\theta}/t)\|\theta_{t-1} - \theta_{t-1}'\| + c_0\ell\|w - w'\|/t. && (30)
\end{aligned}$$

Here the first inequality follows from the triangle inequality, the second inequality follows from the $\ell_{\theta\theta}-$Lipschitz continuity of $\nabla_\theta r$ and ℓ-Lipschitz continuity of ∇r. Therefore, we have

$$\|\theta_t - \theta_t'\| \leq (1 + c_0\ell_{\theta\theta}/t)\|\theta_{t-1} - \theta_{t-1}'\| + c_0\ell\|w - w'\|/t.$$

Let $\delta_t = \|\theta_t - \theta_t'\|$. Then by the above recursion, we have

$$\delta_t + \ell/\ell_{\theta\theta}\|w - w'\| \leq \prod_{i=1}^{t}(1 + c_0\ell_{\theta\theta}/i)\ell\|w - w'\|/\ell_{\theta\theta}.$$

Using the inequalities $e^a \geq 1 + a$ and $\sum_{i=1}^t 1/i \leq \log t$, we have

$$\delta_t \leq \frac{t\ell}{\ell_{\theta\theta}}\|w - w'\|.$$

Letting $t = s$ yields

$$\|\theta_p^\gamma(w) - \theta_p^\gamma(w')\| \leq \frac{s\ell}{\ell_{\theta\theta}}\|w - w'\|.$$

Since $D_p > \gamma$, we have

$$s \leq [D_p/\gamma] + 1 \leq 2D_p/\gamma.$$

Hence,

$$s\frac{\ell}{\ell_{\theta\theta}} \cdot \gamma \leq 2D_p\ell/\ell_{\theta\theta}.$$

Setting $\lambda_p = 2D_p\ell/\ell_{\theta\theta}$ yields the desired result. □

Algorithm 1 GDA

Input: initial iterate $(w_S^0, \theta_S^0) = (0, 0)$, stepsizes α_t, β_t, projection operators P_W and P_Θ;
for $t = 0, \ldots, T-1$ **do**
  $w_S^{t+1} = P_W\left(w_S^t - \alpha_t \nabla_w r_S(w, \theta)\right)$
  $\theta_S^{t+1} = P_\Theta\left(\theta_S^t + \beta_t \nabla_\theta r_S(w, \theta)\right)$
end for

Algorithm 2 GDMax

Input: initial iterate $(w_S^0, \theta_S^0) = (0, 0)$, stepsizes α_t, projection operators P_W and P_Θ;
for $t = 0, \ldots, T-1$ **do**
  $w_S^{t+1} = P_W\left(w_S^t - \alpha_t \nabla_w r_S(w, \theta)\right)$
  $\theta_S^{t+1} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, r_S(w_S^{t+1}, \theta)$
end for

Remark 5. *Note that for some practical nonconvex optimization problems in machine learning, gradient descent indeed converges to the global minima at a reasonably fast rate, e.g., in training deep overparametrized neural networks [Du et al., 2019], robust least squares problems [El Ghaoui and Lebret, 1997], phase retrieval and matrix completion [Ma et al., 2019]. Our Assumptions 6 and 7 can be viewed as an abstract summary of some benign properties of gradient descent for certain nonconvex optimization problems.*

Furthermore, we assume that $f(\cdot, \cdot; z)$ is L-Lipschitz[3] continuous in $W \times \Theta$. This, along with Assumptions 6 and 7, allows us to derive the generalization error bounds of the primal risk and primal gap in terms of algorithmic stability.

Lemma 14. *Suppose that Assumption 6 holds. If a minimax learning Algorithm A is an ϵ-stable algorithm, we have*

$$\zeta_{gen}^P(A) \leq L\epsilon + \sqrt{L\lambda_p}\sqrt{\epsilon}.$$

Proof. This is similar to the proof of Lemma 1. We first define the virtual algorithm $\bar{A}$ which outputs $(w_S^A, \theta_p^\gamma(w_S^A))$. By Assumption 6, it can be easily seen that $\bar{A}$ is $(1+\lambda_p/\gamma)\epsilon$-stable. Then by Theorem 1, we have

$$E_S E_A[r(w_S^A, \theta_p^\gamma(w_S^A)) - r_S(w_S^A, \theta_p^\gamma(w_S^A))] \leq L(1 + \lambda_p/\gamma)\epsilon.$$

This gives us:

$$\begin{aligned} E_S E_A[r(w_S^A) - r_S(w_S^A)] &\leq E_S E_A[r(w_S^A, \theta_p^\gamma(w_S^A)) - r_S(w_S^A, \theta_p^\gamma(w_S^A))] + \gamma \\ &\leq L\epsilon + L\lambda_p\epsilon/\gamma + \gamma. \end{aligned}$$

Taking $\gamma = \sqrt{L\lambda_p}\sqrt{\epsilon}$, we have

$$\zeta_{gen}^p(A) \leq L\epsilon + \sqrt{L\lambda_p}\sqrt{\epsilon}.$$

This completes the proof. □

Similarly, we can derive the generalization bound for the primal gap given the above assumptions.

Theorem 6. *Suppose Assumptions 6 and 7 hold. Then we have*

$$\zeta_{gen}^{PG}(A) \leq \zeta_{gen}^P(A) + \sqrt{L\lambda_e}/\sqrt{n}.$$

The proof of this theorem is similar to the proof of Lemma 14 and Theorem 2 and hence omitted.

[3]Note that this is different from the L defined for the nonconvex-concave case. Here L captures the Lipschitz constant over the whole constraint set. In the nonconvex-concave case, $L = L(B(0, 2C_p + 1))$.

D Proof in Section 4

D.1 Proof of Theorem 4

First, we have

$$\begin{aligned}
&E_S E_{A_0}[r(w_S^{A_0,T}) - \min_{w\in W} r(w)] \\
&= E_S E_{A_0}[r_S(w_S^{A_0,T}) - \min_{w\in W} r_S(w)] + E_S E_{A_0}[r(w_S^{A_0,T}) - r_S(w_S^{A_0,T})] \\
&\qquad + E_S E_{A_0}[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)]. \qquad (31)
\end{aligned}$$

Furthermore, by Assumption 5 and Theorem 2, we have

$$E_S E_{A_0}[r(w_S^{A_0,T}) - \min_{w\in W} r(w)] \le (\phi_{A_0}(M_w) + \phi_{A_0}(C_e(\Theta_0)))/\psi_{A_0}(T) + \zeta_{gen}^P(A_0) + L_\theta^* C_e(\Theta_0)/\sqrt{n}.$$

Next, notice that the output of A_0 is equal to the output of A with probability at least $1-\delta$ and $\|r(w)\| \le 1$. Therefore, we have

$$|E_S E_A[r(w_S^{A,T})] - E_S E_{A_0}[r(w_S^{A_0,T})]| \le \delta,$$

which gives the desired result. □

D.2 Proof of Lemma 3

Define $\delta_t = \|(w_S^t, \theta_S^t) - (w_{S'}^t, \theta_{S'}^t)\|$. We have

$$\delta_{t+1} \le (1 + c_0\ell/t)\delta_t + 2c_0 L_{\Theta_\theta^{GDA}}/nt.$$

Therefore,

$$\begin{aligned}
\delta_{t+1} + \frac{2L_{\Theta_\theta^{GDA}}}{\ell n} &\le (1 + c_0\ell/t)\left(\delta_t + \frac{2L_{\Theta_\theta^{GDA}}}{\ell n}\right) \\
&\le \frac{2L_{\Theta_\theta^{GDA}}}{\ell n} T^{c_0\ell}, \qquad (32)
\end{aligned}$$

which completes the proof. □

D.3 Proof of Lemma 4

For a fixed dataset S, let $g_t = \nabla r_S(w^t, \theta^t)$ and $d_t = \|(w^0, \theta^0) - (w^t, \theta^t)\|$. Then we have $g_t \le L_0 + d_t\ell$ and $d_{t+1} \le d_t + c_0 g_t/t$. Substituting the first inequality into the second one, we have

$$d_{t+1} \le d_t + c_0 d_t/t + L_0 c_0/t,$$

which gives us

$$d_{t+1} + L/\ell \le (1 + c_0\ell/t)(d_t + L_0/\ell).$$

Multiplying this inequality from 0 to $T-1$ yields

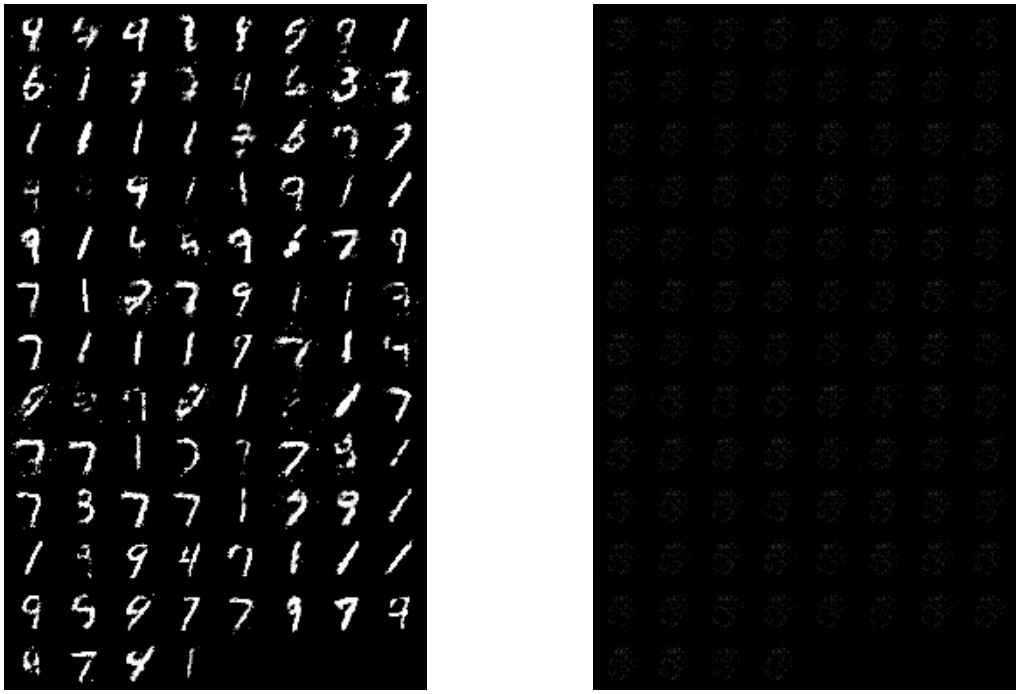
$$d_T \le T^{c_0\ell} L_0/\ell,$$

which completes the proof. □

D.4 Proof of Lemma 5

Let $u = [1, 1, \cdots, 1, 0, \cdots, 0]^T \in \mathbb{R}^{2n}$. Then $\theta_S(w)$ satisfies $Q_S^T \theta_S(w) = u - b_0 e$, where $e = [1, 1, \cdots, 1]^T \in \mathbb{R}^{2n}$. It can be easily seen that $\|u - b_0 e\| \ge \sqrt{n}/2$.

We can also show that $\sigma_{\max}(Q_S) \le 2\sigma_{\max} \cdot \sigma_{\max}(P)$, where $P \in \mathbb{R}^{2n\times m}$ is full row-rank and independent rows. Moreover, every row of P has covariance matrix $I_m/\sqrt{m}$. Then by random matrix theory (see Vershynin [2010]), we have $\sigma_{\max}(P) \le \mathcal{O}(\sqrt{m}/\sqrt{m} - C\sqrt{n}/\sqrt{m} + \log(1/\delta)/\sqrt{m}) = \mathcal{O}(1)$ with probability $1 - C\delta$. Therefore, we have $\theta_S(w) \ge \Omega(\sqrt{n})$. □



(a) GDA (b) GDMax

Figure 1: Comparison of the results on MNIST generated by GDA and GDMax.

D.5 Experiments on GAN-training

In this section, we provide some numerical results to corroborate our theoretical findings.

D.5.1 Setup

We train a GAN on MNIST data using two algorithms – GDA and GDMax. Since the stability is improved by using adaptive methods like Adam, we use Adam-descent-ascent (ADA) and Adam-descent-max (ADMax) instead. ADA simultaneously trains the generator and the discriminator, while ADMax trains the optimal discriminator for each generator step. We simulate this by taking 10 steps of ascent for every descent step. Figure 1 plots the images generated by GANs trained using these two algorithms. Finally, in Figure 2, we plot the norms of the discriminator trained by these two algorithms.

D.5.2 Results

Figure 1 plots the images generated by GANs trained using GDA and GDMax (using Adam instead of the simple gradient step). As predicted by the theory in Section D, we can see that GDA produces better images than the corresponding GAN trained using GDMax. Furthermore, the claim that $C_e >> C_p$ can be seen from Figure 2 where we see that the norm of the discriminator trained using GDMax is much larger than the norm of the discriminator trained using GDA. This follows from the results in Section 4.2. GDMax trains the discriminator to exactly distinguish between the empirical data generated by the true and fake distributions. Therefore, when they are nearly the same, their empirical distributions would be close as well. This would imply that the discriminator would need to have a very large slope (Lipschitz constant) to exactly distinguish between the two empirical datasets, and this in turn leads to a large discriminator norm (which captures the Lipschitz constant of the discriminator).

E Generalization Error for Primal-Dual Risk

If the saddle point exists, the *primal-dual risk* is often a good measure of generalization:

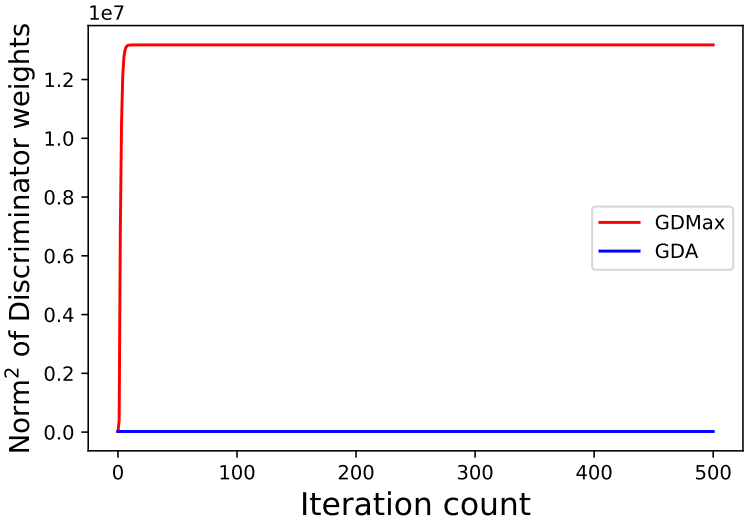


Figure 2: Comparison of the norm squares of discriminator weights.

Definition 7. *The empirical and population primal-dual (PD) risks are defined as:*

$$\Delta_S^{PD}(w,\theta) = \max_{\theta' \in \Theta} r_S(w,\theta') - \min_{w' \in W} r_S(w',\theta),$$

and

$$\Delta^{PD}(w,\theta) = \max_{\theta' \in \Theta} r(w,\theta') - \min_{w' \in W} r(w',\theta).$$

A point (w,θ) is called a saddle point of r_S (or r) if $\Delta_S^{PD}(w,\theta) = 0$ (or $\Delta^{PD}(w,\theta) = 0$). Furthermore, if a saddle point (w_S,θ_S) exists for $r_S(\cdot,\cdot)$, we have $w_S = \min_{w \in W} r_S(w)$. Moreover, if $w_S \in \arg\min_{w \in W} r_S(w)$ and $\theta_S \in \arg\max_{\theta \in \Theta} r_S(w_S,\theta)$, then (w_S,θ_S) is a saddle point of $r_S(\cdot,\cdot)$.

Notice that if we can find an approximate saddle point (w_S,θ_S) of $r_S(w,\theta)$, i.e., $\Delta_S^{PD}(w_S,\theta_S) < \epsilon$ and guarantee that $\Delta^{PD}(w_S,\theta_S) - \Delta_S^{PD}(w_S,\theta_S)$ is small, we can guarantee that $\Delta(w_S,\theta_S)$ is small and therefore (w_S,θ_S) is an approximate saddle point of $r(\cdot,\cdot)$. Hence if the saddle point exists for $r_S(\cdot,\cdot)$, the primal-dual risk can be a good measure for the generalization of a solution. We define the expected generalization error for the primal-dual risk as follows:

Definition 8. *The generalization error for the primal-dual risk is defined as*

$$\zeta_{gen}^{PD}(A) = E_S E_A[\Delta^{PD}(w_S^A,\theta_S^A) - \Delta_S^{PD}(w_S^A,\theta_S^A)].$$

E.1 The generalization of the primal-dual risk for convex-concave problems

Similar to Definition 6, we define the W-capacity as follows:

Definition 9 (W-Capacity)**.** *Let*

$$W_S(\theta) = \min_{w \in W} r_S(w,\theta), \text{ and } \quad W^*(\theta) = \min_{w \in W} r(w,\theta).$$

The W-capacities C_e^w and C_p^w are defined as

$$\begin{aligned} C_e^w &= \max_{S,\theta} \operatorname{dist}(0, W_S(\theta)) \\ C_p^w &= \max_{\theta} \operatorname{dist}(0, W^*(\theta)). \end{aligned} \tag{33}$$

Next, we also define the following:

Definition 10. *Let $f^-(\theta, w; z) = -f(w, \theta; z)$. We first have*

$$r^-(\theta, w) = E_{z\sim P_z}[f^-(\theta, w; z)], \qquad r_S^-(\theta, w) = \frac{1}{n}\sum_{i=1}^{n} f^-(\theta, w; z_i). \tag{34}$$

Furthermore, we define:

$$\begin{aligned} r^-(\theta) &= \max_{w\in W} r^-(\theta, w) = -(\min_{w\in W} r(w, \theta)) \\ r_S^-(\theta) &= \max_{w\in W} r_S^-(\theta, w) = -(\min_{w\in W} r_S(w, \theta)). \end{aligned} \tag{35}$$

Now, we have the following bound for the generalization error of the primal-dual risk, $\zeta_{gen}^{PD}(A)$ for an ϵ-stable Algorithm A :

Theorem 7. *Suppose that Algorithm A is ϵ-stable. The generalization error $\zeta_{gen}^{PD}(A)$ for convex-concave problem, i.e., when $f(\cdot, \cdot; z)$ is convex-concave for all z, is bounded by:*

$$\zeta_{gen}^{PD}(A) = L\sqrt{C_p^2\ell}\sqrt{\epsilon} + L\sqrt{(C_p^w)^2\ell}\sqrt{\epsilon}.$$

Proof. Notice that

$$\begin{aligned} \zeta_{gen}^{PD}(A) &= E_S E_A[\Delta^{PD}(w_S^A, \theta_S^A) - \Delta_S^{PD}(w_S^A, \theta_S^A)] && (36) \\ &= E_S E_A[r(w_S^A) - r_S(w_S^A)] + E_S E_A[r^-(\theta_S^A) - r_S^-(\theta_S^A)]. && (37) \end{aligned}$$

The two terms can be bounded by Lemma 1 respectively. By Lemma 1, we have

$$E_S E_A[r(w_S^A) - r_S(w_S^A)] \le L\sqrt{C_p^2\ell}\sqrt{\epsilon}$$

and

$$E_S E_A[r^-(\theta_S^A) - r_S^-(\theta_S^A)] \le L\sqrt{(C_p^w)^2\ell}\sqrt{\epsilon}.$$

Combining these two inequalities yields the desired result. □

E.2 $\zeta_{gen}^{PD}(T)$ for the proximal point algorithm

In this section, we only study the generalization behavior of the proximal point algorithm (PPA) ((See Equation (3) in [Farnia and Ozdaglar, 2021])). By Farnia and Ozdaglar [2021], the stability of T steps of PPA can be bounded as follows:

Lemma 15 ([Farnia and Ozdaglar, 2021])**.** *The stability of T steps of PPA can be bounded by $\epsilon \le (C_e + C_e^w)T/n$.*

Therefore, substituting the result of Lemma 15 in Theorem 7, we have the following generalization bound for T steps of PPA:

Theorem 8. *After T steps of PPA, the generalization error of the primal-dual gap can be bounded by:*

$$\zeta_{gen}^{PD}(T) = L_\Theta^{3/2}(\sqrt{C_p^2\ell} + \sqrt{(C_p^w)^2\ell})\sqrt{(C_e + C_e^w)T/n}.$$

E.3 The population primal-dual risk of PPA

Finally, we give the population primal-dual risk after T steps of PPA. By Mokhtari et al. [2020b], we have the following convergence result of PPA.

Lemma 16 ([Mokhtari et al., 2020b])**.** *Let (w_S^t, θ_S^t) be the iterates obtained after t iterations of proximal point algorithm on the function $r_S(\cdot, \cdot)$ and $\bar{w}_S^t = \frac{1}{t}\sum_{i=1}^t w_S^i, \bar{\theta}_S^t = \frac{1}{t}\sum_{i=1}^t \theta_S^i$ be the averaged iterates. Then we have*

$$\Delta_S^{PD}(\bar{w}_S^T, \bar{\theta}_S^T) \le \ell(C_e^2 + (C_e^w)^2)/T.$$

Combining Lemma 16 and Theorem 8, we have the following result:

Theorem 9. *Let (w_S^t, θ_S^t) be the iterates obtained after t iterations of proximal point algorithm on the function $r_S(\cdot, \cdot)$ and $\bar{w}_S^t = \frac{1}{t}\sum_{i=1}^t w_S^i, \bar{\theta}_S^t = \frac{1}{t}\sum_{i=1}^t \theta_S^i$ be the averaged iterates. Then, the expected population primal-dual risk at the point $(\bar{w}_S^t, \bar{\theta}_S^t)$ can be bounded by:*

$$E_S[\Delta^{PD}(w_S^t, \theta_S^t)] \le \ell(C_e^2 + (C_e^w)^2)/T + L_\Theta^{3/2}(\sqrt{C_p^2\ell} + \sqrt{(C_p^w)^2\ell})\sqrt{(C_e + C_e^w)T/n}.$$