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

Asu Ozdaglar [1] **Sarath Pattathil** [1] **Jiawei Zhang** [1] **Kaiqing Zhang** [1]

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 the minimax settings, their *generalization* guarantees in stochastic minimax optimization problems, i.e., how the solution trained on empirical data performs on 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 problems, which has also been adopted recently to study generalization in minimax ones, fails in simple examples. We thus propose a new metric to study generalization of minimax learners: the *primal gap*, defined as the difference between the primal risk and its minimum over all models, to circumvent the issues. Next, we derive generalization error bounds for the primal gap in nonconvex-concave settings. As byproducts of our analysis, we also solve two open questions: establishing generalization error bounds for primal risk and primal-dual risk, another existing metric that is only well-defined when the global saddle-point exists, 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]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. Correspondence to: Jiawei Zhang <jwzhang@mit.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

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 algorithms for solving an empirical version of the problem where it approximates the expectation by an average over a sampled dataset, 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 sufficient to achieve the success of stochastic 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 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 performance in deep learning, e.g., these bounds can increase with the training dataset size and easily become vacuous in practice (Nagarajan and Kolter, 2019), making the latter a more fa-

vorable 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). These papers build on the algorithmic stability framework developed in (Bousquet and Elisseeff, 2002), which are further investigated in (Hardt et al., 2016). 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 the definition of risk from minimization problems. Primal-dual risk, on the other hand, is defined similarly but based on the duality gap of the solution. It is know that it is well-defined and can be optimized to zero only when the global saddle-point exists (i.e., $\min$ and $\max$ can be interchanged). Based on these metrics, (Farnia and Ozdaglar, 2021; Lei et al., 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 in Section 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, where the global saddle-point might not exist. In particular, we are interested in the following fundamental question:

What is a good metric to study generalization of minimax learners[1]?

In this paper, we make an initial attempt to answering this question, by identifying the inadequacies of the existing metric, and proposing a new metric, the *primal gap* that overcomes these inadequacies. We then provide generalization error 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 generalization in stochastic minimax optimization, 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 error 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 error bounds for primal risk and primal-dual risk without strong concavity or assuming that the maximization and expectation can be interchanged, while at least one of these assumptions was needed in the literature (Farnia and Ozdaglar, 2021; Lei et al., 2021; Xing et al., 2021; Yang et al., 2022). Finally, under certain assumptions of the max learner, our results also generalize to the nonconvex-nonconcave setting.

[1]Hereafter, we use *learner* and *learning algorithm* interchangeably.

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 i.i.d. variables 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*[2]

$$r(w) = \max_{\theta\in\Theta} E_{z\sim P_z} f(w,\theta;z),$$

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 minimax problem: $\min_{w\in W}\max_{\theta\in\Theta} r_S(w,\theta)$. Notice that (w_S,θ_S) need not be a global 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

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

[2]Note that we slightly abuse the notation here by allowing r and r_S to have inputs that can be both w and (w,θ). The distinction will be clear from context.

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 of the primal risk, as defined below.

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 stochastic minimax optimization (see (Hardt et al., 2016; Farnia and Ozdaglar, 2021; Lei et al., 2021)).

We are interested in the question of 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 Section C, we briefly describe why the generalization error of the primal risk $\zeta_{gen}^P(A)$ is a good measure to study the generalization behavior in minimization problems.

Next, we highlight some results in the literature which discusses generalization error bounds of the primal risk. These results depend on the concept of algorithmic stability we use later.

2.2. Stability of algorithms

Stability analysis is a powerful tool to analyze 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 (ϵ-stable Algorithm)**.** *Suppose that A is a randomized algorithm for solving the stochastic 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 $E_A\|w_S^A - w_{S'}^A\| \leq \epsilon$ and $E_A\|\theta_S^A - \theta_{S'}^A\| \leq \epsilon$ for any neighboring datasets S and S'.*

(Hardt et al., 2016) gives the following basic result for the generalization error of $r_S(w)$.

Theorem 1 ((Hardt et al., 2016))**.** *Consider the (stochastic) minimization problem defined in 8. Suppose $g(\cdot; z)$ is $\bar{L}$-Lipschitz continuous, i.e., $\forall z$, it holds that $\|g(w_1; z) - g(w_2; z)\| \leq \bar{L}\|w_1 - w_2\|, \forall w_1, w_2 \in W$. Then, for an ϵ-stable Algorithm A, we have $|E_S E_A[r(w_S^A) - r_S(w_S^A)]| \leq \bar{L}\epsilon$.*

Section D discusses when Primal Risk be used as a good metric, for example when the expectation and maximization can be interchanged. Unfortunately, this is not the case for many minimax problems. If they are not interchangeable, it is unclear how to estimate the generalization error bound of the primal risk. In fact, whether primal risk is still a good metric for studying generalization behavior in such problems remains elusive.

In the next section, we will see how to estimate generalization error bound of primal risk for nonconvex-concave and even nonconvex-nonconcave problems. To the best of our knowledge, this is the first result which provides generalization error bounds for the primal risk without assuming the interchangeability or strong concavity of the inner maximization problems (see e.g., (Lei et al., 2021)). Furthermore, we will see that even in some simple minimax problems, the generalization error bound of the primal risk can fail to capture the generalization behavior of minimax learners. We then propose a new metric and use its generalization error to properly characterize the generalization behavior of minimax learners.

3. Primal Gap: A New Metric to Study Generalization

The key idea behind the success of $\zeta_{gen}^P(A)$ as a way to characterize to study generalization for minimization learners is that $E[r_S(w)] = r(w)$ for any w, which is no longer the case in the minimax case. In fact, we first show via example that a good bound for the generalization error of primal risk does not imply good generalization behavior for minimax learners.

3.1. Primal risk can fail for minimax learners

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*

$$r(w,\theta) = \frac{1}{2}w^2 - \left(\frac{1}{2n^2}\theta^2 + 1\right)w. \tag{4}$$

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* $\zeta^P_{gen}(A) \leq 0$, *and thus we have a small generalization error for primal risk. However, we can prove that for large enough* λ,

$$E_S[r(w_S) - r(w^*)] \geq 0.02. \tag{5}$$

This means that w_S *has a constant error compared to* w^* *in terms of the population risk, despite that its generalization error 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 $\min_w r_S(w)$ and $\min_w r(w)$ can be very different. 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* $P_G = P_r$. *GAN is a popular tool for solving this problem. Consider a GAN with generator* G, *parametrized by* w *and discriminator* D *parametrized by* θ. *The goal of GAN training is to find a pair of a generator* G *and a discriminator* D *that solves the minimax problem:*

$$\begin{aligned}&\min_G \max_D \quad \{E_{x\sim P_r}\phi(D(x)) + E_{x\sim P_G}[\phi(1 - D(x))]\}\\ &= \min_w \max_\theta \{E_{x\sim P_r}\phi(D_\theta(x)) + E_{y\sim P_0}[\phi(1 - D_\theta(G_w(y)))]\}\end{aligned}$$

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*

$$r(w,\theta) = E_{x\sim P_r}\phi(D_\theta(x)) + E_{y\sim P_0}[\phi(1 - D_\theta(G_w(y)))]$$

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* $r_S(w,\theta) = \frac{1}{n}\left(\sum_{i=1}^n \phi(D_\theta(x_i)) + \phi(1 - D_\theta(G_w(y_i)))\right)$. *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. *Here* w^* *denotes the parameter for which* $G_{w^*} = G^*$. *Then, combining Theorem B.1 in (Arora et al., 2017) and the Lipschitz continuity of* $G_w(y)$ *as well as* $\|w - w^*\| \leq 0.5$, *we have that the distance between the sets* S_1 *and* $\{G_w(y_1), G_w(y_2), \cdots, G_w(y_n)\}$ *will be larger than* 0.6 *with probability greater than* $1 - O(n^2/e^d)$. *Now, if* n *is only of polynomial size of* d, *the optimal discriminator for disjoint datasets outputs* 1 *on one dataset, and* 0 *on the other. On the other hand, when* $w = w^*$, *the optimal discriminator for the population problem outputs* $1/2$ *for any sample it receives. Combining these two results, we have:*

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

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 using $\zeta^P_{gen}(A)$ might not be a good way to characterize the generalization behavior in GAN training. To address this issue, we next define a new metric, the primal gap, and use its generalization error to study the generalization of minimax learners.

3.2. Primal gap to the rescue

The population and empirical 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 $w_S \in \arg\min_{w\in W} r_S(w)$ and $w^* \in \arg\min_{w\in W} r(w)$ 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*

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

Remark 1. *For Example 3, since the maximization and expectation can be interchanged, the minimax problem is*

equivalent to a minimization problem. Then we have

$$\begin{aligned} E_S[\min_w r_S(w)] &= E_S[\min_w \max_\theta E_{z\sim P_z(S)} f(w,\theta;z)] \\ &= E_S[\min_w E_{z\sim P_z(S)}[\max_\theta f(w,\theta;z)]] \\ &= E_S[\min_w E_{z\sim P_z(S)}[f_{\max}(w;z)]] \\ &\le E_S[E_{z\sim P_z(S)}[f_{\max}(w;z)]] \end{aligned}$$

for any w. Therefore, we have $E_S[\min_w r_S(w)] \le \min_w r(w)$. Consequently, we have $\zeta_{gen}^P \ge \zeta_{gen}^{PG}$, which means that good generalization bounds for the primal risk implies good generalization bounds for the primal gap. Therefore, if the maximization and expectation are interchangeable, primal risk is sufficient to study the generalization behavior because the generalization error of the primal risk is an upper bound of the generalization error of the primal gap in this case.

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

3.3. Relationship between generalization and stability

We provide bounds for the generalization error of the primal gap (Definition 5) for ϵ-stable Algorithm 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*

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

We define the capacities C_p and C_e as:

$$\begin{aligned} C_p(\Theta) &= \max_{w\in W} \operatorname{dist}(0,\Theta(w)) \\ C_e(\Theta) &= \max_S \max_{w\in W} \operatorname{dist}(0,\Theta_S(w)), \end{aligned}$$

where $\operatorname{dist}(p,\mathcal{S})$ denotes the distance between a point p to a set $\mathcal{S}$ in Euclidean space, i.e.,

$$\operatorname{dist}(p,\mathcal{S}) := \inf_{q\in\mathcal{S}} \|p-q\|_2.$$

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

The norm of the model parameter (its distance to 0) is usually viewed as the metric for the complexity of the model. 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, i.e., the power of the maximizer, 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., for all z*

$$\begin{aligned} &\|\nabla f(w_1,\theta_1;z) - \nabla f(w_2,\theta_2;z)\| \\ &\le \ell(\|w_1-w_2\| + \|\theta_1-\theta_2\|), \quad \forall w_1,w_2 \in W, \quad \forall \theta_1,\theta_2 \in \Theta. \end{aligned}$$

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., $\|\nabla_\theta f(w,\theta_1;z) - \nabla_\theta f(w,\theta_2;z)\| \le \ell_{\theta\theta}\|\theta_1-\theta_2\|, \forall w \in W, \quad \forall \theta_1,\theta_2 \in \Theta$.

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., $\|f(w_1,\theta_1;z) - f(w_2,\theta_2;z)\| \le L(\Theta_1)(\|w_1-w_2\| + \|\theta_1-\theta_2\|), \quad \forall w_1,w_2 \in W, \ \forall \theta_1,\theta_2 \in \Theta_1$, and the gradient $\nabla f(w,\theta;z)$ is uniformly bounded as $\|\nabla_{w,\theta} f(w,\theta;z)\| \le L(\Theta_1)$ 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:

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

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 (without strong concavity) setting.

Lemma 1. *The generalization error of the primal risk of an ϵ-stable Algorithm A for a minimax problem with concave maximization problem can be bounded by $\zeta_{gen}^P(A) \le \sqrt{4L\ell C_p^2}\cdot\sqrt{\epsilon} + \epsilon L$.*

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

$$E_S E_A\big[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)\big] = E_S\big[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)\big]. \quad (6)$$

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 (6).

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

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

The first term in the bound above is from the generalization bound of the primal risk, as shown in Lemma 1. Note that the bound in Lemma 1 only involves C_p, as the key in the analysis is to upper-bound the population risk $r(w_S^A)$, which requires bounding the power of the maximizer using the population capacity C_p. This reflects the intuition that the power of the maximizer should affect the generalization behavior of minimax learners, and the stronger the maximizer is, the harder for the learner to generalize. On the other hand, the bound in Theorem 2 additionally involve C_e, the empirical capacity. Technically, C_e (instead of C_p) appears since we need to bound $\min_w r_S(w)$ (defined on the empirical dataset) in the Primal Min Error term in (6). The appearance of C_e reflects the intuition that the difference between the maximizers of the empirical and population risks should make a difference in characterizing the generalization of minimax learners. This intuition cannot be captured by the generalization error of the primal risk, as in Lemma 1. Note that in the minimization case, the Primal Min Error can be upper-bounded directly by zero, and such a distinction disappears, making primal risk a valid metric.

3.4. Revisiting Example 1

Recall Example 1 in Section 3.1. In this example, we have that the primal risk has a small generalization error, but the solution w_S does not generalize well. In particular, as shown in the appendix (Proposition 3), we have

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \ge 0.005. \quad (7)$$

On the other hand, it is easy to compute that $L_\theta^* = \lambda \log n/\sqrt{n}$ and $C_e = \lambda n$. Therefore, by Theorem 2, we have an upper bound for the Primal Min Error (see Equation (6)): $E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \le 4L_\theta^* C_e/\sqrt{n} = 4\log n$, which is tight up to a log factor according to (7). 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.

Due to space limitations, we relegate the analysis for the nonconvex-nonconcave setting to Appendix, Section G, and the comparison of GDA and GDMax to the Appendix, Section H.

4. Conclusions

In this paper, we first demonstrate the shortcomings of one popular metric, the primal risk, in terms of characterizing the generalization behavior of minimax learners. We then propose a new metric, the primal gap, whose generalization error overcomes these shortcomings and captures the generalization behavior of algorithms that solve stochastic minimax problems. Finally, we use this newly proposed metric to study the generalization behavior of two different algorithms – GDA and GDMax, and study cases where GDA has a better generalization behavior than GDMax. Future directions include further investigation of the proposed new metric, the primal gap, and deriving its (tighter) generalization error bounds in other structured stochastic minimax optimization problems in machine learning.

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.

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.

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.

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 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 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) have established generalization bounds of standard gradient descent-ascent and proximal point algorithms under the convex-concave setting, and those of stochastic GDA and GDMax under the nonconvex-strongly concave setting. Concurrently, (Lei et al., 2021) derive high-probability generalization bounds for both convex-concave and weakly convex-weakly concave settings, with possibly nonsmooth objectives, also through the lens of algorithmic stability. Both 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 is known to be appropriate only when the saddle point exists, which is usually not the case in the nonconvex settings that are common in machine learning. 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.

B. Existing Related Results

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

Theorem 3 ((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*
$$E_S E_A[\zeta^P_{gen}(A)] \leq \epsilon.$$
2. *If $f(\cdot,\cdot;z)$ is nonconvex-strongly-concave and f is μ-strongly-concave with respect to θ, then*
$$E_S E_A[\zeta^P_{gen}(A)] \leq L\sqrt{\kappa^2+1}\epsilon.$$

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 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.

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 3, 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.*

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

Consider a stochastic optimization problem of the form

$$\min_{w\in W} \; E_{z\sim P_z}[g(w;z)]. \tag{8}$$

We define the (minimization) primal risk (population and empirical version respectively) as: $r(w) = E_{z\sim P_z} g(w;z)$, and $r_S(w) = \frac{1}{n}\sum_{i=1}^n g(w;z_i)$. 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): $E[r(w_S^A)] \le E[r_S(w_S^A)] + \epsilon$. 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:

$$E[r(w_S^A) - \min_{w\in W} r(w)] \le E[r_S(w_S^A)] + \epsilon - \min_{w\in W} r(w). \tag{9}$$

However, note that for minimization problems, since $E[r_S] = r$, we have that[3] $\min_{w\in W} r(w) \ge E[\min_{w\in W} r_S(w)]$, which gives us:

$$E[r(w_S^A) - \min_{w\in W} r(w)] \le E[r_S(w_S^A)] + \epsilon - E[\min_{w\in W} r_S(w)] = E[r_S(w_S^A) - \min_{w\in W} r_S(w)] + \epsilon = \epsilon.$$

Therefore, for minimization problems, if the generalization error for primal risk is small, the solution to the empirical risk minimization problem has good generalization behavior.

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

D. When is primal risk a valid metric for minimax learners?

According to the above discussions for minimization problems, we know that the primal risk is a valid metric to study generalization behavior in these problems, and furthermore, the generalization error bound of the primal risk can be estimated in terms of algorithmic stability. However, Theorem 1 cannot be directly extended to analyze the generalization behavior of minimax learners because we have an additional maximization step before taking expectation.

A natural question emerges: Under what conditions does primal risk serve as a valid metric to study generalization behavior of minimax problems. One sufficient condition is when the maximization step and expectation can be interchanged, i.e., when

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

for any distribution P_z. Letting $f_{\max}(w;z) = \max_{\theta\in\Theta} f(w,\theta;z)$, we further have

$$r(w) = \max_{\theta\in\Theta} E_{z\sim P_z} f(w,\theta;z) = E_{z\sim P_z}[\max_{\theta\in\Theta} f(w,\theta;z)] = E_{z\sim P_z} f_{\max}(w;z).$$

Therefore, the minimax problem in (1) is equivalent to the (stochastic) minimization problem with loss function $f_{\max}(w;z)$. Moreover, letting $P(S)$ be the uniform distribution over the dataset $S = \{z_1, \cdots, z_n\}$, we have

$$r_S(w) = \max_{\theta\in\Theta} E_{z\sim P(S)}[f(w,\theta;z)] E_{z\sim P(S)}[\max_{\theta\in\Theta} f(w,\theta;z)] = \frac{1}{n}\sum_{i=1}^{n} f_{\max}(w;z_i).$$

Therefore, $r_S(w)$ is just the empirical primal risk corresponding to the minimization problem with loss function $f_{\max}(w;z)$. Hence, Theorem 1 can be directly used to minimax problems where the maximization and expectation can be interchanged.

Theorem 4. *Suppose that $f(w,\theta;z)$ is $\bar{L}$-Lipschitz continuous with respect to w, i.e., $|f(w_1,\theta;z) - f(w_2,\theta;z)| \le \bar{L}\|w_1 - w_2\|$ for any $w_1, w_2 \in W, \theta \in \Theta$ and z. If an Algorithm A is ϵ-stable, we have*

$$E_S E_A[r(w_S^A) - r_S(w_S^A)] \le \bar{L}\epsilon.$$

Proof. From the previous analysis along with Theorem 1, it suffices to show that $f_{\max}(\cdot;z)$ is $\bar{L}$-Lipschitz continuous. In fact, we have

$$\begin{aligned} f_{\max}(w_1;z) - f_{\max}(w_2;z) &= f(w_1,\theta(w_1);z) - f(w_2,\theta(w_2);z) \\ &\le f(w_1,\theta(w_1);z) - f(w_2,\theta(w_1);z) \le \bar{L}\|w_1 - w_2\|, \end{aligned}$$

where $\theta(w) \in \arg\max_{\theta\in\Theta} f(w,\theta;z)$, the first inequality is because of the definition of $\theta(w)$ and the second inequality is because of the Lipschitz continuity of f with respect to w. Using the same argument, we can prove

$$f_{\max}(w_2;z) - f_{\max}(w_1;z) \le \bar{L}\|w_1 - w_2\|.$$

Therefore, we prove the $\bar{L}$-Lipschitz continuity of $f_{\max}(\cdot;z)$ and hence finish the proof. □

By the above discussion, we know that if maximization and expectation can be interchanged, the minimax problem can be reduced to a minimization problem and hence the primal risk is a valid metric for studying the generalization behavior of minimax learners and the generalization error can be estimated using the same method as for minimization problems. In practice, the adversarial-training problems can be such an example of minimax problems.

Example 3 (Adversarial-training)**.** *We consider the adversarial training problem (Madry et al., 2017). Suppose we have loss function $g(w;z)$ for a supervised learning problem. Here z denotes the training sample and w denotes the model parameter. Due to the noise in the data or due to an adversarial attack, for any sample z, we consider an uncertainty set $B(z,\epsilon_0)$ around it. The goal is to train a model that is robust to the data with possible perturbation in the uncertainty set. Let θ_z be some adversarial sample from the set $B(z,\epsilon_0)$ and let θ be an infinite dimensional vector (functional) with the component θ_z corresponding to the sample z. Define the function $\iota_B(v)$ to be the indicator function of the set B, i.e., $\iota_B(v) = 0$ if $v \in B$ and $\iota_B(v) = \infty$ otherwise. The goal of adversarial training is to solve the following minimax problem:*

$$\min_{w}\max_{\theta} \quad E_{z\sim P_z} f(w,\theta;z), \tag{10}$$

where $f(w,\theta;z) = g(w;\theta_z) + \iota_{B(z,\epsilon_0)}(\theta_z)$. For any distribution P_z over z's, we have

$$\begin{aligned}\max_\theta\ E_{z\sim P_z} f(w,\theta;z) &= \max_\theta\ E_{z\sim P_z}[g(w;\theta_z) + \iota_{B(z,\epsilon_0)}(\theta_z)] = E_{z\sim P_z}[\max_{\theta_z}\ (g(w;\theta_z) + \iota_{B(z,\epsilon_0)}(\theta_z))] \\ &= E_{z\sim P_z}[\max_\theta\ f(w,\theta;z)],\end{aligned}$$

where the second and the third equalities use the fact that $\theta_{z'}$ does not contribute to $f(w,\theta;z)$ if $z \neq z'$. Therefore, the expectation and maximization can be interchanged in adversarial training problems. This implies that the results of Theorem 4 can be applied and therefore primal risk is a valid metric to study the generalization behavior in such problems.

E. 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,

$$\begin{aligned} r_S(w) &= \max_{\theta\in\Theta} r(w,\theta) && (11)\\ &\geq r_S(w,0) && (12)\\ &= r(w). && (13)\end{aligned}$$

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 2. *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

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

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

$$r(w_S) - \min_{w\in W} r(w) \geq 0.005, \tag{14}$$

by direct calculation. Therefore, we have

$$\begin{aligned} & E_S[r(w_S) - \min_{w\in W} r(w)] && (15)\\ \geq\ & \Pr(|\sum_{i=1}^n z_i| \in [0.5, \lambda]) \cdot 0.05 + \Pr(|\sum_{i=1}^n z_i| \notin [0.5, \lambda]) \cdot 0 && (16)\\ \geq\ & 0.02, && (17)\end{aligned}$$

where the first inequality is because of (14) 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 2, 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. □

F. Proofs in Section 3

F.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 3. *For any $w \in W$, we have*[4]

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

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

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

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 4. *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 3. □

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

Lemma 5. *We have*

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

[4]For point 2, it holds when $s > 0$. For $s = 0$, we have the bound $r(w) - r(w,\bar{\theta}(w)) \le \ell_{\theta\theta}C_p^2$. We do not separate this degenerate case for ease of presentation.

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 error bound of the Primal Risk for an Algorithm A with ϵ-stability. First, we have

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

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 3 and the last inequality is because of Lemma 5. Optimizing over[5] s, the generalization error is bounded by $\zeta_{gen}^P(A) \leq \sqrt{4L\ell C_p^2} \cdot \sqrt{\epsilon} + \epsilon L$. This completes the proof. □

F.2. 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 error bound on the primal gap.

Lemma 6. *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*

$$E_S[\min_{w \in W} r_S(w) - \min_{w \in W} r(w)] \leq 4L_\theta^* C_e/\sqrt{n}.$$

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 3, we have the following lemma:

[5] Here we assume that the optimal s is a real number greater than 0. Constraining s to be an integer and also incorporating 0 does not change the result and we ignore this case here. See also Footnote 4.

Lemma 7. *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 3. 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 8. *The stability of Algorithm $\tilde{A}$ is given by $\epsilon_{sta}(\tilde{A}) = 2s(L_\theta^*)^2/(n\ell_{\theta\theta})$.*

Then by the standard stability theory in (Hardt et al., 2016), we have

$$|E_S E_A[r_S(w^*, \tilde{\theta}_S) - r(w^*, \tilde{\theta}_S)]| \leq 2s(L_\theta^*)^2/(n\ell_{\theta\theta}). \tag{18}$$

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 7, and (iv) follows from the optimality of θ^* given w^*. Then by (18), we have

$$E_S[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)] \le E_S[r_S(w^*, \tilde{\theta}_S) - r(w^*, \tilde{\theta}_S)] + \ell_{\theta\theta} C_e^2/s \tag{19}$$

$$\le 2s(L_\theta^*)^2/(n\ell_{\theta\theta}) + \ell_{\theta\theta} C_e^2/s \tag{20}$$

$$\le 4L_\theta^* C_e/\sqrt{n} \tag{21}$$

which completes the proof. □

The final statement of the theorem follows from Lemma 6 and Lemma 1. □

G. 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.

We make the following assumptions on the inner maximization problem:

Assumption 5. *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. $r(w) - r(w, \theta_P^\gamma(w)) \le \gamma$.
2. $\|\theta_P^\gamma(w) - \theta_P^\gamma(w')\| \le \frac{\lambda_p}{\gamma}\|w - w'\|$ *with some constant $\lambda_p > 0$ for all $w, w' \in W$.*

Assumption 6. *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. $r_S(w^*) - r_S(w^*, \theta_E^\gamma(S)) \le \gamma$.
2. *For any neighboring dataset S, S', we have $\|\theta_E^\gamma(S) - \theta_E^\gamma(S')\| \le \frac{\lambda_e}{n\gamma}$ with some constant $\lambda_e > 0$.*

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

Lemma 9. *Consider constants $D_e \ge \gamma$ and $D_p \ge \gamma$.*

1. *Suppose that gradient ascent with diminishing stepsizes c_0/t for the problem $\max_{\theta\in\Theta} r(w,\theta)$ has convergence rate $r(w) - r(w,\theta^s) \le D_p/s$. Then we define $\theta_p^\gamma(w)$ by performing $s = D_p/\gamma$ steps of gradient ascent. Then, $\theta_p^\gamma(w)$ satisfies Assumption 5.*
2. *Suppose that gradient ascent with constant stepsize c_0 for the problem $\max_{\theta\in\Theta} r(w,\theta)$ has convergence rate $r(w) - r(w,\theta^s) \le D_p\eta^s$ 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 5.*
3. *Suppose that gradient ascent with diminishing stepsizes c_0/t for the problem $\max_{\theta\in\Theta} r_S(w,\theta)$ has convergence rate $r_S(w) - r_S(w,\theta^s) \le D_p/s$. Then we define $\theta_e^\gamma(S)$ by performing $s = D_e/\gamma$ steps of gradient ascent. Then, $\theta_e^\gamma(S)$ satisfies Assumption 6.*
4. *Suppose that gradient ascent with constant stepsize c_0 for the problem $\max_{\theta\in\Theta} r_S(w,\theta)$ has convergence rate $r_S(w) - r_S(w,\theta^s) \le D_e\eta^s$ 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 6.*

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 (**?**). Our Assumptions 5 and 6 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[6] continuous in $W\times\Theta$. This, along with Assumptions 5 and 6, allows us to derive the generalization error bounds of the primal risk and primal gap in terms of algorithmic stability.

Lemma 10. *Suppose that Assumption 5 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 5. *Suppose Assumptions 5 and 6 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 10 and Theorem 2 and hence omitted.

G.1. Proof of Lemma 9

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

$$\begin{aligned}
\|\theta_t-\theta_t'\| &= \|(\theta_{t-1}+c_0\nabla_\theta r(w,\theta_{t-1})/t)-(\theta_{t-1}'+c_0\nabla_\theta r(w,\theta_{t-1}')/t) && (22)\\
&\quad+c_0(\nabla_\theta r(w,\theta_{t-1}')-\nabla_\theta r(w',\theta_{t-1}'))/t\| && (23)\\
&\leq \|(\theta_{t-1}+c_0\nabla_\theta r(w,\theta_{t-1})/t)-(\theta_{t-1}'+c_0\nabla_\theta r(w,\theta_{t-1}')/t)\| && (24)\\
&\quad+c_0\|(\nabla_\theta r(w,\theta_{t-1}')-\nabla_\theta r(w',\theta_{t-1}'))/t\| && (25)\\
&\leq(1+c_0\ell_{\theta\theta}/t)\|\theta_{t-1}-\theta_{t-1}'\|+c_0\ell\|w-w'\|/t. && (26)
\end{aligned}$$

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 $e^a\geq1+a$ 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

$$s\leq[D_p/\gamma]+1\leq2D_p/\gamma.$$

Hence,

$$s\frac{\ell}{\ell_{\theta\theta}}\cdot\gamma\leq2D_p\ell/\ell_{\theta\theta}.$$

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

[6]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, $L=L(B(0,2C_p+1))$.

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**
  $w_S^{t+1} = P_W\left(w_S^t - \alpha_t \nabla_w r_S(w, \theta)\right)$
  $\theta_S^{t+1} = \operatorname*{argmax}_{\theta \in \Theta} r_S(w_S^{t+1}, \theta)$
end for

G.2. Proof of Lemma 10

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 5, 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

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

H. Comparison of GDA and GDMax

In Section 3.3, we provide generalization bounds for the primal gap for any ϵ-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 I.

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 the stochastic minimax problem can be solved by GDA and GDMax, i.e., we assume that GDA and GDMax satisfy the following assumption:

Assumption 7. *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, and $\phi_A(s)$, $\psi_A(s)$ 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 A to solve (1) under constraint sets W_0 and Θ_0. Specifically, $A(W, \Theta)$ is just A.

Theorem 6. *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:*

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

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

Remark 6. *Theorem 6 builds a closer connection between generalization behavior and the dynamics of the minimax learner* A*. It shows that suitable restriction to the max learner can lead to better minimax learner, in terms of generalization. We make this clear in the comparison of GDA and GDMax by analyzing the three terms in Theorem 6.*

H.1. Analyzing the term I

First, we study the generalization error bound of the primal risk, i.e., ζ^P_{gen} in Theorem 6. For GDA, we can estimate ζ^P_{gen} by using Lemma 1. Therefore, it suffices to estimate the stability of GDA. We do this in the following lemma:

Lemma 11. *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* $\epsilon^{GDA} \leq 2L_{\Theta_\theta^{GDA}} T^{c_0\ell}/(n\ell)$.

Now, since we have a bound for $\zeta^P_{gen}(A)$ for ϵ-stable Algorithm A in Lemma 1, we can substitute the stability bound for GDA from Lemma 11 in this expression to get a bound on $\zeta^P_{gen}(GDA)$ for GDA. We do this in the next proposition. We can bound $\zeta^P_{gen}(A_0)$ for GDA by substituting the stability bound in Lemma 11 into Lemma 1 (letting $\epsilon = \epsilon^{GDA}$).

Proposition 4. *Let* $c_0 = \max\{\alpha_0, \beta_0\}$ *and assume that* $f(\cdot, \cdot; z)$ *is* $L_{\Theta_\theta^{GDA}}$*-Lipschitz-continuous inside the set* $W \times \Theta_\theta^{GDA}$*. 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:*

$$\zeta^P_{gen}(GDA_T) \leq (L_{\Theta_\theta^{GDA}})^{3/2} \sqrt{8C_p^2/\ell} \sqrt{T^{c_0\ell}/n} + 2L^2_{\Theta_\theta^{GDA}} T^{c_0\ell}/(n\ell).$$

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 $\zeta^P_{gen}(\text{GDMax})$ can be a constant that is independent of n, which means that for the case where $r(w, \theta)$ is nonconvex-concave, the generalization error of primal risk of GDMax can be undesirable.

Example 4 (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* $E_S r(w_S) = E[\frac{(\sum_{i=1}^n z_i)2}{2}] = 1/2 > 0$*. Thus,* $\zeta^P_{gen}(GDMax) = E[r(w_S) - r_S(w_S)] = 1/2 > 0$ *cannot be made small.*

Therefore, from Proposition 4 and Example 4, we see that the bound for the expected population primal gap contains the term ζ^P_{gen} which cannot be bounded for GDMax, whereas can be bounded for GDA which leads us to the conclusion that GDA generalizes better than GDMax for such problems. However, it is possible to bound $\zeta^P_{gen}(\text{GDMax})$ in certain problems, and in this case the other terms in Theorem 6 become crucial. We analyze them next.

H.2. Analyzing the term II

As shown in Example 1, sometimes GDMax can have a good generalization bound for the primal risk. Therefore, we need to analyze the other two terms in Theorem 6, 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 (5). 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 12. *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.

H.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 13. *Suppose $\Phi(x)$ is sub-Gaussian and the matrix*

$$Q_S = \begin{bmatrix}\Phi(x_1) & \Phi(x_2)\cdots & \Phi(x_n) & \Phi(G_w(y_1)) & \cdots & \Phi(G_w(y_n))\end{bmatrix}$$

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 12, and using the fact that C_e is large from Lemma 13, we see from Theorem 6 that GDA can generalize better than GDMax. More detailed discuss of the GAN-training example and Lemma 13 can be found in Section I.

I. Proofs in Section H

I.1. Proof of Theorem 6

First, we have

$$\begin{aligned}
&E_S E_{A_0}[r(w_S^{A_0,T}) - \min_{w\in W} r(w)] \\
&= E_S E_{A_0}[r_S(w_S^{A_0,T}) - \min_{w\in W} r_S(w)] + E_S E_{A_0}[r(w_S^{A_0,T}) - r_S(w_S^{A_0,T})] \\
&\qquad + E_S E_{A_0}[\min_{w\in W} r_S(w) - \min_{w\in W} r(w)].
\end{aligned} \tag{27}$$

Furthermore, by Assumption 7 and Theorem 2, we have

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

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

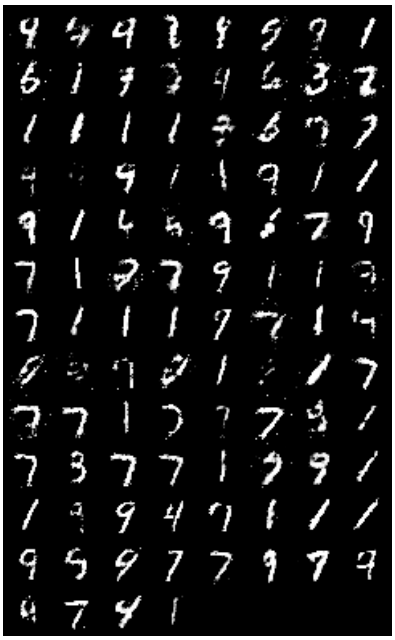
$$|E_S E_A[r(w_S^{A,T})] - E_S E_{A_0}[r(w_S^{A_0,T})]| \leq \delta,$$

which gives the desired result. □

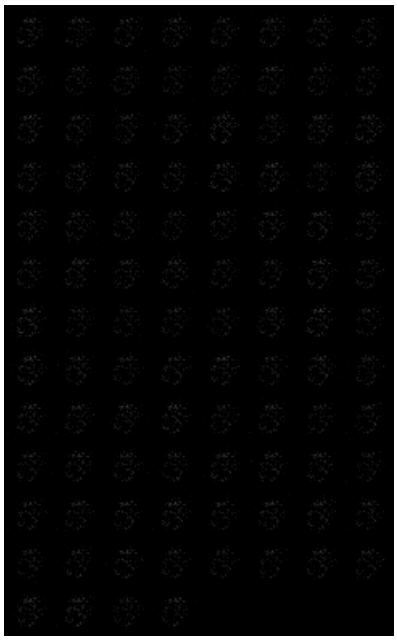
I.2. Proof of Lemma 11

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

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



(a) GDA



(b) GDMax

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

Therefore,

$$\delta_{t+1} + \frac{2L_{\Theta_\theta^{GDA}}}{\ell n} \leq (1 + c_0\ell/t)\left(\delta_t + \frac{2L_{\Theta_\theta^{GDA}}}{\ell n}\right) \leq \frac{2L_{\Theta_\theta^{GDA}}}{\ell n} T^{c_0\ell}, \tag{28}$$

which completes the proof. □

I.3. Proof of Lemma 12

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 $g_t \leq L_0 + d_t\ell$ and $d_{t+1} \leq d_t + c_0 g_t/t$. Substituting the first inequality into the second one, we have

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

which gives us

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

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

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

which completes the proof. □

I.4. Proof of Lemma 13

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 $\|u - b_0 e\| \geq \sqrt{n}/2$.

We can also show that $\sigma_{\max}(Q_S) \leq 2\sigma_{\max} \cdot \sigma_{\max}(P)$, 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 $\sigma_{\max}(P) \leq \mathcal{O}(\sqrt{m}/\sqrt{m} - C\sqrt{n}/\sqrt{m} + \log(1/\delta)/\sqrt{m}) = \mathcal{O}(1)$ with probability $1 - C\delta$. Therefore, we have $\theta_S(w) \geq \Omega(\sqrt{n})$. □

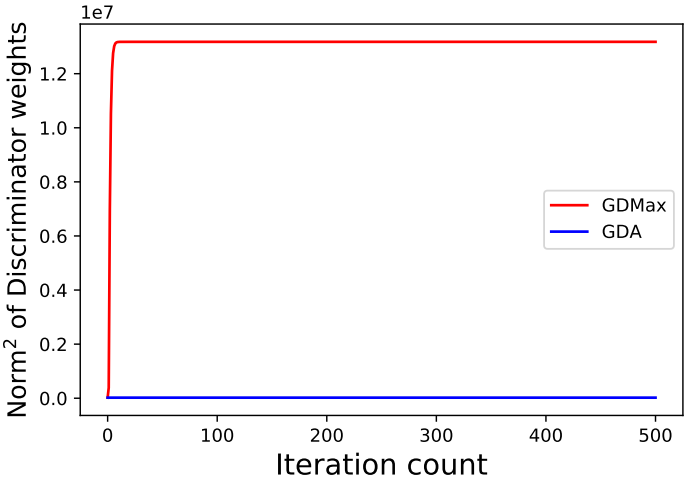


Figure 2. Comparison of the norm squares of discriminator weights.

I.5. Experiments on GAN-training

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

I.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.

I.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 I, 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 H.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).

J. Generalization Error for Primal-Dual Risk

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

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

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

and

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

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 generalization error of the primal-dual risk can be a good measure for the generalization of the solution to the empirical problem. 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*

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

J.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*

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

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

$$\begin{aligned}C_p^w&=\max_{\theta}\operatorname{dist}(0,W^*(\theta)\\ C_e^w&=\max_{S,\theta}\operatorname{dist}(0,W_S(\theta)).\end{aligned}\tag{29}$$

Next, we also define the following:

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

$$r^-(\theta,w)=E_{z\sim P_z}[f^-(\theta,w;z)],\qquad r_S^-(\theta,w)=\frac{1}{n}\sum_{i=1}^n f^-(\theta,w;z_i).\tag{30}$$

Furthermore, we define:

$$\begin{aligned}r^-(\theta)&=\max_{w\in W} r^-(\theta,w)=-(\min_{w\in W} r(w,\theta))\\ r_S^-(\theta)&=\max_{w\in W} r_S^-(\theta,w)=-(\min_{w\in W} r_S(w,\theta)).\end{aligned}\tag{31}$$

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:*

$$\zeta_{gen}^{PD}(A)\le\left(\sqrt{4L\ell C_p^2}+\sqrt{4L\ell(C_p^w)^2}\right)\sqrt{\epsilon}+2\epsilon L.$$

Proof. Notice that

$$\begin{aligned}\zeta_{gen}^{PD}(A)&=E_SE_A[\Delta^{PD}(w_S^A,\theta_S^A)-\Delta_S^{PD}(w_S^A,\theta_S^A)]&(32)\\ &=E_SE_A[r(w_S^A)-r_S(w_S^A)]+E_SE_A[r^-(\theta_S^A)-r_S^-(\theta_S^A)].&(33)\end{aligned}$$

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

$$E_SE_A[r(w_S^A)-r_S(w_S^A)]\le\sqrt{4L\ell C_p^2}\sqrt{\epsilon}+\epsilon L$$

and

$$E_SE_A[r^-(\theta_S^A)-r_S^-(\theta_S^A)]\le\sqrt{4L\ell(C_p^w)^2}\sqrt{\epsilon}+\epsilon L.$$

Combining these two inequalities yields the desired result. □

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

In this section, we 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 14 ((Farnia and Ozdaglar, 2021))**.** *The stability of T steps of PPA can be bounded by $\epsilon \leq \mathcal{O}(T/n)$.*

Therefore, substituting the result of Lemma 14 in Theorem 7, we have the following bound for ζ_{gen} for T steps of PPA:

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

$$\zeta_{gen}^{PD}(T) \leq \mathcal{O}\left(\sqrt{T/n} + T/n\right).$$

J.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 15 ((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*

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

Combining Lemma 15 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:*

$$E_S[\Delta^{PD}(w_S^t, \theta_S^t)] \leq \mathcal{O}\left(1/T + \sqrt{T/n} + T/n\right).$$