Min-Max Optimization without Gradients: Convergence and Applications to Black-Box Evasion and Poisoning Attacks

Sijia Liu [*1] **Songtao Lu** [*2] **Xiangyi Chen** [*3] **Yao Feng** [*4] **Kaidi Xu** [*5] **Abdullah Al-Dujaili** [*6] **Mingyi Hong** [3] **Una-May O'Reilly** [7 1]

Abstract

In this paper, we study the problem of constrained min-max optimization in a black-box setting, where the desired optimizer cannot access the gradients of the objective function but may query its values. We present a principled optimization framework, integrating a zeroth-order (ZO) gradient estimator with an alternating projected stochastic gradient descent-ascent method, where the former only requires a small number of function queries and the later needs just one-step descent/ascent update. We show that the proposed framework, referred to as *ZO-Min-Max*, has a sublinear convergence rate under mild conditions and scales gracefully with problem size. We also explore a promising connection between black-box min-max optimization and black-box evasion and poisoning attacks in adversarial machine learning (ML). Our empirical evaluations on these use cases demonstrate the effectiveness of our approach and its scalability to dimensions that prohibit using recent black-box solvers.

1. Introduction

Min-max optimization problems have been studied for multiple decades (Wald, 1945), and the majority of the proposed methods assume access to first-order (FO) information, i.e. gradients, to find or approximate robust solutions (Nesterov, 2007; Gidel et al., 2017; Hamedani et al., 2018; Qian et al., 2019; Rafique et al., 2018; Sanjabi et al., 2018b; Lu et al., 2019; Nouiehed et al., 2019; Lu et al., 2020; Jin et al., 2019). Different from *standard* optimization, min-max optimization tackles a composition of an inner maximization problem and an outer minimization problem. It can be used in many real-world applications, which are faced with various forms of uncertainty or adversary. For instance, when training a ML model on user-provided data, malicious users can carry out a data poisoning attack: providing false data with the aim of corrupting the learned model (Steinhardt et al., 2017; Tran et al., 2018; Jagielski et al., 2018). At inference time, malicious users can evade detection of multiple models in the form of adversarial example attacks (Goodfellow et al., 2014; Liu et al., 2016; 2018a). Our study is particularly motivated by the design of *data poisoning* and *evasion* adversarial attacks from *black-box* machine learning (ML) or deep learning (DL) systems, whose internal configuration and operating mechanism are unknown to adversaries. We propose *zeroth-order (gradient-free)* min-max optimization methods, where gradients are neither symbolically nor numerically available, or they are tedious to compute.

Recently, *zeroth-order (ZO) optimization* has attracted increasing attention in solving ML/DL problems, where FO gradients (or stochastic gradients) are approximated based only on the function values (Liu et al., 2020). For example, ZO optimization serves as a powerful and practical tool for generation of adversarial examples to evaluate the adversarial robustness of black-box ML/DL models (Chen et al., 2017; Ilyas et al., 2018; Tu et al., 2018; Ilyas et al., 2019; Chen et al., 2018; Li et al., 2019). ZO optimization can also help to solve automated ML problems, where the gradients with respect to ML pipeline configuration parameters are intractable (Aggarwal et al., 2019; Wang & Wu, 2019). Furthermore, ZO optimization provides computationally-efficient alternatives of high-order optimization methods for solving complex ML/DL tasks, e.g., robust training by leveraging input gradient or curvature regularization (Finlay & Oberman, 2019; Moosavi-Dezfooli et al., 2019), network control and management (Chen & Giannakis, 2018; Liu et al., 2018b), and data processing in high dimension (Liu et al., 2018b; Golovin et al., 2019). Other recent applications include generating model-agnostic contrastive explanations (Dhurandhar et al., 2019) and escaping saddle points (Flokas et al., 2019).

[*]Equal contribution [1]MIT-IBM Watson AI Lab, IBM Research [2]Thomas J. Watson Research Center, IBM Research [3]University of Minnesota, Twin Cities [4]Tsinghua University [5]Northeastern University [6]Algorithmic Systems Group, Analog Devices [7]CSAIL, MIT. Correspondence to: Sijia Liu <Sijia.Liu@ibm.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Current studies (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2015; Duchi et al., 2015; Ghadimi et al., 2016; Shamir, 2017; Balasubramanian & Ghadimi, 2018; Liu et al., 2018c; 2019) suggested that ZO methods for solving *single-objective* optimization problems typically agree with the iteration complexity of FO methods but encounter a slow-down factor up to a small-degree polynomial of the problem dimensionality. To the best of our knowledge, it was an open question whether any convergence rate analysis can be established for black-box min-max (bi-level) optimization. In this paper, we develop a *provable* and *scalable* black-box min-max stochastic optimization method by integrating a *query-efficient* randomized ZO gradient estimator with a *computation-efficient* alternating gradient descent-ascent framework. Here the former requires a small number of function queries, and the latter needs just one-step descent/ascent update.

Contribution. We summarize our contributions as follows. ***(i)*** We identify a class of black-box attack problems which turn out to be min-max black-box optimization problems. ***(ii)*** We propose a scalable and principled framework (ZO-Min-Max) for solving constrained min-max saddle point problems under both one-sided and two-sided black-box objective functions. Here the one-sided setting refers to the scenario where only the outer minimization problem is black-box. ***(iii)*** We provide a novel convergence analysis characterizing the number of objective function evaluations required to attain locally robust solution to black-box min-max problems (structured by nonconvex outer minimization and strongly concave inner maximization). Our analysis handles stochasticity in both objective function and ZO gradient estimator, and shows that ZO-Min-Max yields $\mathcal{O}(1/T + 1/b + d/q)$ convergence rate, where T is number of iterations, b is mini-batch size, q is number of random direction vectors used in ZO gradient estimation, and d is number of optimization variables. ***(iv)*** We demonstrate the effectiveness of our proposal in practical data poisoning and evasion attack generation problems.[1]

2. Related Work

FO min-max optimization. Gradient-based methods have been applied with celebrated success to solve min-max problems such as robust learning (Qian et al., 2019), generative adversarial networks (GANs) (Sanjabi et al., 2018a; Lu et al., 2019), adversarial training (Al-Dujaili et al., 2018b; Madry et al., 2018), and robust adversarial attack generation (Wang et al., 2019b). Some of FO methods are motivated by theoretical justifications based on Danskin's theorem (Danskin, 1966), which implies that the negative of the gradient of the outer minimization problem at inner maximizer is a descent direction (Madry et al., 2018). Convergence analysis of other FO min-max methods has been studied under different problem settings, e.g., (Lu et al., 2020; Qian et al., 2019; Rafique et al., 2018; Sanjabi et al., 2018b; Nouiehed et al., 2019). It was shown in (Lu et al., 2020) that a deterministic FO min-max algorithm has $\mathcal{O}(1/T)$ convergence rate. In (Qian et al., 2019; Rafique et al., 2018), stochastic FO min-max methods have also been proposed, which yield the convergence rate in the order of $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T^{1/4})$, respectively. However, these works were restricted to unconstrained optimization at the minimization side. In (Sanjabi et al., 2018b), noncovnex-concave min-max problems were studied, but the proposed analysis requires solving the maximization problem only up to some small error. In (Nouiehed et al., 2019), the $\mathcal{O}(1/T)$ convergence rate was proved for nonconvex-nonconcave min-max problems under Polyak-Łojasiewicz conditions. Different from the aforementioned FO settings, ZO min-max stochastic optimization suffers randomness from both stochastic sampling in objective function and ZO gradient estimation, and this randomness would be coupled in alternating gradient descent-descent steps and thus make it more challenging in convergence analysis.

Gradient-free min-max optimization. In the black-box setup, coevolutionary algorithms were used extensively to solve min-max problems (Herrmann, 1999; Schmiedlechner et al., 2018). However, they may oscillate and never converge to a solution due to pathological behaviors such as *focusing* and *relativism* (Watson & Pollack, 2001). Fixes to these issues have been proposed and analyzed—e.g., asymmetric fitness (Jensen, 2003; Branke & Rosenbusch, 2008). In (Al-Dujaili et al., 2018c), the authors employed an evolution strategy as an unbiased approximate for the descent direction of the outer minimization problem and showed empirical gains over coevlutionary techniques, albeit without any theoretical guarantees. Min-max black-box problems can also be addressed by resorting to direct search and model-based descent and trust region methods (Audet & Hare, 2017; Larson et al., 2019; Rios & Sahinidis, 2013). However, these methods lack convergence rate analysis and are difficult to scale to high-dimensional problems. For example, the off-the-shelf model-based solver COBYLA only supports problems with 2^{16} variables at maximum in SciPy Python library (Jones et al., 2001), which is even smaller than the size of a single ImageNet image.

The recent work (Bogunovic et al., 2018) proposed a robust Bayesian optimization (BO) algorithm and established a theoretical lower bound on the required number of the min-max objective evaluations to find a near-optimal point. However, BO approaches are often tailored to low-dimensional problems and its computational complexity prohibits scalable application. From a game theory perspective, the min-max

[1]Source code is available at https://github.com/KaidiXu/ZO-minmax

solution for some problems correspond to the Nash equilibrium between the outer minimizer and the inner maximizer, and hence black-box Nash equilibria solvers can be used (Picheny et al., 2019; Al-Dujaili et al., 2018a). This setup, however, does not always hold in general. Our work contrasts with the above lines of work in designing and analyzing black-box min-max techniques that are *both* scalable and theoretically grounded.

3. Problem Setup

In this section, we define the black-box min-max problem and briefly motivate its applications. By *min-max*, we mean that the problem is a composition of inner maximization and outer minimization of the objective function f. By *black-box*, we mean that the *objective function* f is only accessible via functional evaluations. Mathematically, we have

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \quad f(\mathbf{x},\mathbf{y}) \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are optimization variables, f is a differentiable objective function, and $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ are compact convex sets. For ease of notation, let $d_x = d_y = d$. In (1), the objective function f could represent either a deterministic loss or stochastic loss $f(\mathbf{x},\mathbf{y}) = \mathbb{E}_{\boldsymbol{\xi}\sim p}\left[f(\mathbf{x},\mathbf{y};\boldsymbol{\xi})\right]$, where $\boldsymbol{\xi}$ is a random variable following the distribution p. In this paper, we cover the stochastic variant in (1).

We focus on two *black-box* scenarios:

(a) *One-sided black-box*: f is a white box w.r.t. $\mathbf{y}$ but a black box w.r.t. $\mathbf{x}$.

(b) *Two-sided black-box*: f is a black box w.r.t. both $\mathbf{x}$ and $\mathbf{y}$.

Motivation of setup (a) and (b). The formulation of the *one-sided black-box* min-max problem can be used to design the *black-box ensemble evasion attack*, where the attacker generates adversarial examples (i.e., crafted examples with slight perturbations for misclassification at the *testing* phase) and optimizes its worst-case performance against an *ensemble* of black-box classifiers and/or example classes. The formulation of *two-sided black-box* min-max problem represents another type of attack at the *training* phase, known as *poisoning attack*, where the attacker deliberately influences the training data (by injecting poisoned samples) to manipulate the results of a black-box predictive model. Although problems of designing ensemble evasion attack (Liu et al., 2016; 2018a; Wang et al., 2019b) and data poisoning attack (Jagielski et al., 2018; Wang et al., 2019a) have been studied in the literature, most of them assumed that the adversary has the *full* knowledge of the target ML model, leading to an impractical *white-box* attack setting. By contrast, we provide a solution to min-max attack generation under *black-box* ML models. We refer readers to Section 6 for further discussion and demonstration of our framework on these problems.

4. ZO-Min-Max: A Framework for Black-Box Min-Max Optimization

Our interest is in a scalable and theoretically principled framework for solving black-box min-max problems of the form (1). To this end, we first introduce a randomized gradient estimator that only requires a few number of pointwise function evaluations. Based on that, we then propose a ZO alternating projected gradient method to solve (1) under both one-sided and two-sided black-box setups.

Randomized gradient estimator. In the ZO setting, we adopt a randomized gradient estimator to estimate the gradient of a function with the *generic* form $h(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}}[h(\mathbf{x};\boldsymbol{\xi})]$ (Gao et al., 2014; Berahas et al., 2019),

$$\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) = \frac{1}{bq}\sum_{j\in\mathcal{I}}\sum_{i=1}^{q} \frac{d[h(\mathbf{x}+\mu\mathbf{u}_i;\boldsymbol{\xi}_j) - h(\mathbf{x};\boldsymbol{\xi}_j)]}{\mu}\mathbf{u}_i, \tag{2}$$

where d is number of variables, $\mathcal{I}$ denotes the mini-batch set of b *i.i.d.* stochastic samples $\{\boldsymbol{\xi}_j\}_{j=1}^{b}$, $\{\mathbf{u}_i\}_{i=1}^{q}$ are q *i.i.d.* random direction vectors drawn uniformly from the unit sphere, and $\mu > 0$ is a smoothing parameter. We note that the ZO gradient estimator (2) involves randomness from both stochastic sampling w.r.t. $\mathbf{u}_i$ as well as the random direction sampling w.r.t. $\boldsymbol{\xi}_j$. It is known from (Gao et al., 2014, Lemma 2) that $\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x})$ provides an unbiased estimate of the gradient of the smoothing function of h rather than the true gradient of h. Here the smoothing function of h is defined by $h_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{v}}[h(\mathbf{x}+\mu\mathbf{v})]$, where $\mathbf{v}$ follows the uniform distribution over the unit Euclidean ball. Besides the bias, we provide an upper bound on the variance of (2) in Lemma 1.

Lemma 1. *Suppose that for all $\boldsymbol{\xi}$, $h(\mathbf{x};\boldsymbol{\xi})$ has L_h Lipschitz continuous gradients and the gradient of $h(\mathbf{x};\boldsymbol{\xi})$ is upper bounded as $\|\nabla_{\mathbf{x}} h(\mathbf{x};\boldsymbol{\xi})\|_2^2 \le \eta^2$ at $\mathbf{x}\in\mathbb{R}^d$. Then $\mathbb{E}\left[\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x})\right] = \nabla_{\mathbf{x}} h_\mu(\mathbf{x})$, and*

$$\mathbb{E}\left[\|\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x})\|_2^2\right] \le \sigma^2(L_h,\mu,b,q,d), \tag{3}$$

where the expectation is taken over all randomness, and $\sigma^2(L_h,\mu,b,q,d) = \frac{2\eta^2}{b} + \frac{4d\eta^2+\mu^2L_h^2d^2}{q}$.

Proof: See Appendix A.2. □

In Lemma 1, if we choose $\mu \le 1/\sqrt{d}$, then the variance bound is given by $\mathcal{O}(1/b + d/q)$. In our problem setting (1), the ZO gradients $\widehat{\nabla}_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$ and $\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$ follow the generic form of (2) by fixing $\mathbf{y}$ and letting $h(\cdot) := f(\cdot,\mathbf{y})$ or by fixing $\mathbf{x}$ and letting $h(\cdot) := f(\mathbf{x},\cdot)$, respectively.

Algorithmic framework. To solve problem (1), we alternatingly perform ZO projected gradient descent/ascent

Algorithm 1 ZO-Min-Max to solve problem (1)

Input: given $\mathbf{x}^{(0)}$ and $\mathbf{y}^{(0)}$, learning rates α and β, the number of random directions q, and the possible mini-batch size b for stochastic optimization
for $t = 1, 2, \ldots, T$ **do**
  $\mathbf{x}$-step: perform ZO-PGD (4)
  $\mathbf{y}$-step:
  if $f(\mathbf{x}^{(t)}, \mathbf{y})$ is black box w.r.t. $\mathbf{y}$ **then**
    perform ZO-PGA (5)
  else
    perform PGA using $\nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)})$ like (5)
  end if
end for

method for updating $\mathbf{x}$ and $\mathbf{y}$. Specifically, the ZO projected gradient descent (ZO-PGD) is conducted over $\mathbf{x}$

$$\mathbf{x}^{(t)} = \mathrm{proj}_{\mathcal{X}} \left(\mathbf{x}^{(t-1)} - \alpha \widehat{\nabla}_{\mathbf{x}} f \left(\mathbf{x}^{(t-1)}, \mathbf{y}^{(t-1)} \right) \right), \quad (4)$$

where t is the iteration index, $\widehat{\nabla}_{\mathbf{x}} f$ denotes the ZO gradient estimate of f w.r.t. $\mathbf{x}$, $\alpha > 0$ is the learning rate at the $\mathbf{x}$-minimization step, and $\mathrm{proj}_{\mathcal{X}}(\mathbf{a})$ signifies the projection of $\mathbf{a}$ onto $\mathcal{X}$, given by the solution to the problem $\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{a}\|_2^2$. For one-sided ZO min-max optimization, besides (4), we perform FO projected gradient ascent (PGA) over $\mathbf{y}$. And for two-sided ZO min-max optimization, our update on $\mathbf{y}$ obeys ZO-PGA

$$\mathbf{y}^{(t)} = \mathrm{proj}_{\mathcal{Y}} \left(\mathbf{y}^{(t-1)} + \beta \widehat{\nabla}_{\mathbf{y}} f \left(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)} \right) \right), \quad (5)$$

where $\beta > 0$ is the learning rate at the $\mathbf{y}$-maximization step, and $\widehat{\nabla}_{\mathbf{y}} f$ denotes the ZO gradient estimate of f w.r.t. $\mathbf{y}$. The proposed method is named as *ZO-Min-Max*; see the pseudocode in Algorithm 1.

Why estimates gradient rather than distribution of function values? Besides ZO optimization using random gradient estimates, the black-box min-max problem (1) can also be solved using the Bayesian optimization (BO) approach, e.g., (Bogunovic et al., 2018; Al-Dujaili et al., 2018a). The core idea of BO is to approximate the objective function as a Gaussian process (GP) learnt from the history of function values at queried points. Based on GP, the solution to problem (1) is then updated by maximizing a certain reward function, known as acquisition function. The advantage of BO is its mild requirements on the setting of black-box problems, e.g., without the need of differentiability. However, BO usually does not scale beyond low-dimensional problems since learning the accurate GP model and solving the acquisition problem takes intensive computation cost per iteration. By contrast, our proposed method is more efficient, and mimics the first-order method by just using the random gradient estimate (2) as the descent/ascent direction. In Figure A1, we compare ZO-Min-Max with the BO based STABLEOPT algorithm proposed by (Bogunovic et al., 2018) through a toy example shown in Appendix B and a poisoning attack generation example in Sec. 6.2. We can see that ZO-Min-Max not only achieves more accurate solution but also requires less computation time.

Why is difficult to analyze the convergence of ZO-Min-Max? The convergence analysis of ZO-Min-Max is more challenging than the case of FO min-max algorithms. The stochasticity of the gradient estimator makes the convergence analysis sufficiently different from the FO deterministic case (Lu et al., 2020; Qian et al., 2019), since the errors in minimization and maximization are coupled as the algorithm proceeds.

Moreover, the conventioanl analysis of ZO optimization for single-objective problems cannot directly be applied to ZO-Min-Max. Even at the one-sided black-box setting, ZO-Min-Max conducts alternating optimization using one-step ZO-PGD and PGA with respect to $\mathbf{x}$ and $\mathbf{y}$ respectively. This is different from a reduced ZO optimization problem with respect to $\mathbf{x}$ only by solving problem $\min_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x})$, where $h(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. Here acquiring the solution to the inner maximization problem could be non-trivial and computationally intensive. Furthermore, the alternating algorithmic structure leads to opposite optimization directions (minimization vs maximization) over variables $\mathbf{x}$ and $\mathbf{y}$ respectively. Even applying ZO optimization only to one side, it needs to quantify the effect of ZO gradient estimation on the descent over both $\mathbf{x}$ and $\mathbf{y}$. We provide a detailed convergence analysis of ZO-Min-Max in Sec. 5.

5. Convergence Rate Analysis

We first elaborate on assumptions and notations used in analyzing the convergence of ZO-Min-Max (Algorithm 1).

A1: In (1), $f(\mathbf{x}, \mathbf{y})$ is continuously differentiable, and is strongly concave w.r.t. $\mathbf{y}$ with parameter $\gamma > 0$, namely, given $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}, \mathbf{y}_1) \leq f(\mathbf{x}, \mathbf{y}_2) + \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_2)^T (\mathbf{y}_1 - \mathbf{y}_2) - \frac{\gamma}{2} \|\mathbf{y}_1 - \mathbf{y}_2\|^2$ for all points $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$. And f is lower bounded by a finite number f^* and has bounded gradients $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \boldsymbol{\xi})\| \leq \eta^2$ and $\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \boldsymbol{\xi})\| \leq \eta^2$ for stochastic optimization with $\boldsymbol{\xi} \sim p$. Here $\|\cdot\|$ denotes the ℓ_2 norm. The constraint sets $\mathcal{X}, \mathcal{Y}$ are convex and bounded with diameter R.

A2: $f(\mathbf{x}, \mathbf{y})$ has Lipschitz continuous gradients, i.e., there exists $L_x, L_y > 0$ such that $\|\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y})\| \leq L_x \|\mathbf{x}_1 - \mathbf{x}_2\|$ for $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, and $\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y})\| \leq L_y \|\mathbf{x}_1 - \mathbf{x}_2\|$ and $\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_2)\| \leq L_y \|\mathbf{y}_1 - \mathbf{y}_2\|$ for $\forall \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$.

We remark that **A1** and **A2** are required for analyzing the convergence of ZO-Min-Max. They were used even for the analysis of first-order optimization methods with single

rather than bi-level objective functions (Chen et al., 2019; Ward et al., 2019). In **A1**, the strongly concavity of $f(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{y}$ holds for applications such as robust learning over multiple domains (Qian et al., 2019), and robust adversarial attack generation shown in Section 6. In **A2**, the assumption of smoothness (namely, Lipschitz continuous gradient) is required to quantify the descent of the alternating projected stochastic gradient descent-ascent method. For clarity, we summarize the problem and algorithmic parameters used in our convergence analysis in Table A1 of Appendix A.1.

We measure the convergence of ZO-Min-Max by the proximal gradient (Lu et al., 2020; Ghadimi et al., 2016; Lin et al., 2020),

$$\mathcal{G}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} (1/\alpha)\left(\mathbf{x} - \text{proj}_{\mathcal{X}}(\mathbf{x} - \alpha \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}))\right) \\ (1/\beta)\left(\mathbf{y} - \text{proj}_{\mathcal{Y}}(\mathbf{y} + \beta \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))\right) \end{bmatrix}, \quad (6)$$

where $(\mathbf{x}, \mathbf{y})$ is a first-order stationary point of (1) iff $\|\mathcal{G}(\mathbf{x}, \mathbf{y})\| = 0$.

Descent property in $\mathbf{x}$-minimization. In what follows, we delve into our convergence analysis. Since ZO-Min-Max (Algorithm 1) calls for ZO-PGD for solving both one-sided and two-sided black-box optimization problems, we first show the descent property at the $\mathbf{x}$-minimization step of ZO-Min-Max in Lemma 2.

Lemma 2. *(Descent lemma in minimization) Under **A1**-**A2**, let $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ be a sequence generated by ZO-Min-Max. When $f(\mathbf{x}, \mathbf{y})$ is black-box w.r.t. $\mathbf{x}$, then we have following descent property w.r.t. $\mathbf{x}$:*

$$\mathbb{E}[f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] \leq \mathbb{E}[f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})] - \left(\frac{1}{\alpha} - \frac{L_x}{2}\right) \mathbb{E}\|\Delta_{\mathbf{x}}^{(t+1)}\|^2 + \alpha \sigma_x^2 + L_x \mu^2 \quad (7)$$

where $\Delta_{\mathbf{x}}^{(t)} := \mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}$, and $\sigma_x^2 := \sigma^2(L_x, \mu, b, q, d)$ defined in (3).

Proof: See Appendix A.3.1. □

It is clear from Lemma 2 that updating $\mathbf{x}$ leads to the reduced objective value when choosing a small learning rate α. However, ZO gradient estimation brings in additional errors in terms of $\alpha\sigma_x^2$ and $L_x\mu^2$, where the former is induced by the variance of gradient estimates in (3) and the latter is originated from bounding the distance between f and its smoothing version; see (24) in Appendix A.3.

Convergence rate of ZO-Min-Max by performing PGA. We next investigate the convergence of ZO-Min-Max when PGA is used at the $\mathbf{y}$-maximization step (Line 8 of Algorithm 1) for solving one-sided black-box optimization problems.

Lemma 3. *(Descent lemma in maximization) Under **A1**-**A2**, let $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ be a sequence generated by Algorithm 1 and define the potential function as*

$$\mathcal{P}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)}) = \mathbb{E}[f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})] + \frac{4 + 4\beta^2 L_y^2 - 7\beta\gamma}{2\beta^2\gamma} \mathbb{E}\|\Delta_{\mathbf{y}}^{(t)}\|^2, \quad (8)$$

where $\Delta_{\mathbf{y}}^{(t)} := \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)}$. When $f(\mathbf{x}, \mathbf{y})$ is black-box w.r.t. $\mathbf{x}$ and white-box w.r.t. $\mathbf{y}$, then we have the following descent property w.r.t. $\mathbf{y}$:

$$\mathcal{P}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \Delta_{\mathbf{y}}^{(t+1)}) \leq \mathcal{P}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)}) - \left(\frac{1}{2\beta} - \frac{2L_y^2}{\gamma}\right) \mathbb{E}\|\Delta_{\mathbf{y}}^{(t+1)}\|^2 + \left(\frac{2}{\gamma^2\beta} + \frac{\beta}{2}\right) L_x^2 \mathbb{E}\|\Delta_{\mathbf{x}}^{(t+1)}\|^2. \quad (9)$$

Proof: See Appendix A.3.2. □

It is shown from (9) that when β is small enough, then the term $(1/(2\beta) - 2L_y^2/\gamma)\mathbb{E}\|\Delta_{\mathbf{y}}^{(t+1)}\|^2$ will give some descent of the potential function after PGA, while the last term in (9) will give some ascent to the potential function. However, such a quantity will be compensated by the descent of the objective function in the minimization step shown by Lemma 2. Combining Lemma 2 and Lemma 3, we obtain the convergence rate of ZO-Min-Max in Theorem 1.

Theorem 1. *Suppose that **A1**-**A2** hold, the sequence $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ over T iterations is generated by Algorithm 1 in which learning rates satisfy $\beta < 1/(4L_y^2)$ and $\alpha \leq \min\{1/L_x, 1/(L_x/2 + 2L_x^2/(\gamma^2\beta) + \beta L_x^2/2)\}$. When $f(\mathbf{x}, \mathbf{y})$ is black-box w.r.t. $\mathbf{x}$ and white-box w.r.t. $\mathbf{y}$, the convergence rate of ZO-Min-Max under a uniformly and randomly picked $(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$ from $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^T$ is given by*

$$\mathbb{E}\|\mathcal{G}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})\|^2 \leq \frac{c}{\zeta} \frac{(\mathcal{P}_1 - f^* - \nu R^2)}{T} + \frac{c\alpha\sigma_x^2}{\zeta} + \frac{cL_x\mu^2}{\zeta} \quad (10)$$

*where ζ is a constant independent on the parameters μ, b, q, d and T, $\mathcal{P}_t := \mathcal{P}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)})$ given by (8), $c = \max\{L_x + 3/\alpha, 3/\beta\}$, $\nu = \min\{4 + 4\beta^2 L_y^2 - 7\beta\gamma, 0\}/(2\beta^2\gamma)$, σ_x^2 is variance bound of ZO gradient estimate given in (7), and f^*, R, γ, L_x and L_y have been defined in **A1**-**A2**.*

Proof: See Appendix A.3.3. □

To better interpret Theorem 1, we begin by clarifying the parameters involved in our convergence rate (10). First, the parameter ζ in the denominator of the derived convergence error is non-trivially lower bounded given appropriate learning rates α and β (as will be evident in **Remark 1**). Second, the parameter c is inversely proportional to α and β. Thus, to guarantee the constant effect of the ratio c/ξ, it is better not to set these learning rates too small; see a specification in **Remark 1-2**. Third, the parameter ν is non-negative and appears in terms of $-\nu R^2$, thus, it will

not make convergence rate worse. Fourth, $\mathcal{P}_1$ is the initial value of the potential function (8). By setting an appropriate learning rate β (e.g., following **Remark 2**), $\mathcal{P}_1$ is then upper bounded by a constant determined by the initial value of the objective function, the distance of the first two updates, Lipschitz constant L_y and strongly concave parameter γ. We next provide **Remarks 1-3** on Theorem 1.

Remark 1. Recall that $\zeta = \min\{c_1, c_2\}$ (Appendix B.2.3), where $c_1 = 1/(2\beta) - 2L_y^2/\gamma$ and $c_2 = \frac{1}{\alpha} - (\frac{L_x}{2} + \frac{2L_x^2}{\gamma^2\beta} + \frac{\beta L_x^2}{2})$. Given the fact that L_x and L_y are Lipschitz constants and γ is the strongly concavity constant, a proper lower bound of ζ thus relies on the choice of the learning rates α and β. By setting $\beta \leq \frac{\gamma}{8L_y^2}$ and $\alpha \leq 1/(L_x + \frac{4L_x^2}{\gamma^2\beta} + \beta L_x^2)$, it is easy to verify that $c_1 \geq \frac{2L_y^2}{\gamma}$ and $c_2 \geq \frac{L_x}{2} + \frac{2L_x^2}{\gamma^2\beta} + \frac{\beta L_x^2}{2} \geq \frac{L_x}{2} + \frac{2L_x^2}{\gamma}$. Thus, we obtain that $\zeta \geq \min\{\frac{2L_y^2}{\gamma}, \frac{2L_x^2}{\gamma} + \frac{L_x}{2}\}$. This justifies that ζ has a non-trivial lower bound, which will not make the convergence error bound (10) vacuous (although the bound has not been optimized over α and β).

Remark 2. It is not wise to set learning rates α and β to extremely small values since c is *inversely* proportional to α and β. We could choose $\beta = \frac{\gamma}{8L_y^2}$ and $\alpha = 1/(L_x + \frac{4L_x^2}{\gamma^2\beta} + \beta L_x^2)$ in Remark 1 to guarantee the constant effect of c/ζ.

Remark 3. By setting $\mu \leq \min\{1/\sqrt{d}, 1/\sqrt{T}\}$, we obtain $\sigma_x^2 = \mathcal{O}(1/b + d/q)$ from Lemma 1, and Theorem 1 implies that ZO-Min-Max yields $\mathcal{O}(1/T + 1/b + d/q)$ convergence rate for one-sided black-box optimization. Compared to the FO rate $\mathcal{O}(1/T)$ (Lu et al., 2020; Sanjabi et al., 2018a), ZO-Min-Max converges only to a neighborhood of stationary points with $\mathcal{O}(1/T)$ rate, where the size of the neighborhood is determined by the mini-batch size b, the problem size d, and the number of random direction vectors q used in ZO gradient estimation. It is worth mentioning that the stationary gap also exists in the ZO projected stochastic gradient descent even for solving single-objective minimization problems (Ghadimi et al., 2016). The rate is worse than ZO optimization methods for solving simpler unconstrained non-stochastic problems (Nesterov & Spokoiny, 2015).

As shown in **Remark 3**, a large mini-batch size b or number of random direction vectors q can improve the iteration complexity of ZO-Min-Max. However, this requires $O(bq)$ times more function queries per iteration from (2). It implies the tradeoff between iteration complexity and function query complexity in ZO optimization.

Convergence rate of ZO-Min-Max by performing ZO-PGA. The previous convergence analysis of ZO-Min-Max is served as a basis for analyzing the more general two-sided black-box case, where ZO-PGA is used at the $\mathbf{y}$-maximization step. In Lemma 4, we examine the descent property in maximization by using ZO gradient estimation.

Lemma 4. *(Descent lemma in maximization) Under **A1-A2**, let $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ be a sequence generated by Algorithm 1 and define the potential function as*

$$\mathcal{P}'(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)}) = \mathbb{E}[f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})] + \frac{4 + 4(3L_y^2 + 2)\beta^2 - 7\beta\gamma}{\beta^2\gamma}\mathbb{E}\|\Delta_{\mathbf{y}}^{(t)}\|^2. \quad (11)$$

When function $f(\mathbf{x}, \mathbf{y})$ is black-box w.r.t. both $\mathbf{x}$ and $\mathbf{y}$, we have the following descent w.r.t. $\mathbf{y}$:

$$\mathcal{P}'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \Delta_{\mathbf{y}}^{(t+1)}) \leq \mathcal{P}'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)}) - \left(\frac{1}{2\beta} - \frac{6L_y^2 + 4}{\gamma}\right)\mathbb{E}\|\Delta_{\mathbf{y}}^{(t+1)}\|^2 + \left(\frac{6L_x^2}{\gamma^2\beta} + \frac{3\beta L_x^2}{2}\right)\mathbb{E}\|\Delta_{\mathbf{x}}^{(t+1)}\|^2 + \frac{7\beta^2\gamma^2 + 28\beta\gamma + 12}{\beta\gamma^2}\sigma_y^2 + \frac{\beta\gamma + 4}{4\beta^2\gamma}\mu^2 d^2 L_y^2, \quad (12)$$

where $\sigma_y^2 := \sigma^2(L_y, \mu, b, q, d)$ given in (3).

Proof: See Appendix A.4.1. □

Lemma 4 is analogous to Lemma 3 by taking into account the effect of ZO gradient estimate $\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ on the potential function (11). Such an effect is characterized by the terms related to σ_y^2 and $\mu^2 d^2 L_y^2$ in (12).

Theorem 2. *Suppose that **A1-A2** hold, the sequence $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ over T iterations is generated by Algorithm 1 in which learning rates satisfy $\beta < \gamma/(4(3L_y^2 + 2))$ and $\alpha \leq \min\{L_x, 1/(L_x/2 + 6L_x^2/(\gamma^2\beta) + 3\beta L_x^2/2)\}$. When $f(\mathbf{x}, \mathbf{y})$ is black-box w.r.t. both $\mathbf{x}$ and $\mathbf{y}$, the convergence rate of ZO-Min-Max under a uniformly and randomly picked $(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$ from $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^T$ is given by*

$$\mathbb{E}\|\mathcal{G}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})\|^2 \leq \frac{c}{\zeta'}\frac{\mathcal{P}_1' - f^* - \nu' R^2}{T} + \frac{c\alpha}{\zeta'}\sigma_x^2 + \left(\frac{cb_1}{\zeta'} + d^2 L_y^2\right)\mu^2 + \left(\frac{cb_2}{\zeta'} + 2\right)\sigma_y^2,$$

*where ζ' is a constant independent on the parameters μ, b, q, d and T, $\mathcal{P}_t' := \mathcal{P}'(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)})$ in (11), c has been defined in (10), $\nu' = \frac{\min\{4 + 4(3L_y^2+2)\beta^2 - 7\beta\gamma, 0\}}{\beta^2\gamma}$, $b_1 = L_x + \frac{d^2 L_y^2(4+\beta\gamma)}{4\beta^2\gamma}$ and $b_2 = \frac{7\beta^2\gamma^2 + 28\beta\gamma + 12}{\beta\gamma^2}$, σ_x^2 and σ_y^2 have been introduced in (7) and (12), and f^*, R, γ, L_x and L_y have been defined in **A1-A2**.*

Proof: See Appendix A.4.2. □

Following the similar argument in Remark 1 of Theorem 1, one can choose proper learning rates α and β to obtain valid lower bound on ζ'. However, different from Theorem 1, the convergence error shown by Theorem 2 involves an additional error term related to σ_y^2 and has worse dimension-dependence on the term related to μ^2. The latter yields a

more restricted choice of the smoothing parameter μ: we obtain $\mathcal{O}(1/T + 1/b + d/q)$ convergence rate when $\mu \leq 1/(d\sqrt{T})$.

6. Applications & Experiment Results

In what follows, we evaluate the empirical performance of ZO-Min-Max on two applications of adversarial exploration: a) design of black-box ensemble evasion attack against deep neural networks (DNNs), and b) design of black-box poisoning attack against a logistic regression model.

6.1. Ensemble attack via universal perturbation

A black-box min-max problem formulation. We consider the scenario in which the attacker generates a *universal* adversarial perturbation over multiple input images against an *ensemble* of multiple classifiers (Liu et al., 2016; 2018a). Considering I classes of images (the group of images within a class i is denoted by Ω_i) and J network models, the goal of the adversary is to find the *universal perturbation* $\mathbf{x}$ additive to I classes of images against J models. The proposed black-box attack formulation mimics the white-box attack formulation (Wang et al., 2019b)

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \ \underset{\mathbf{w}\in\mathcal{W}}{\text{maximize}} \quad f(\mathbf{x},\mathbf{w}) := \sum_{j=1}^{J}\sum_{i=1}^{I}\left[w_{ij}F_{ij}\left(\mathbf{x};\Omega_i\right)\right] - \lambda\|\mathbf{w}-\mathbf{1}/(IJ)\|_2^2, \quad (13)$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{w} \in \mathbb{R}^{IJ}$ are optimization variables, w_{ij} denotes the (i,j)th entry of $\mathbf{w}$ corresponding to the importance weight of attacking the set of images at class i under the model j, $\mathcal{X}$ denotes the perturbation constraint, e.g., $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq \epsilon, \forall \mathbf{z} \in \cup_i \Omega_i\}$, and $\mathcal{W}$ is the probabilistic simplex $\mathcal{W} = \{\mathbf{w} \mid \mathbf{1}^T\mathbf{w} = 1, \mathbf{w} \geq 0\}$. In problem (13), $F_{ij}\left(\mathbf{x};\Omega_i\right)$ is the attack loss for attacking images in Ω_i under model j, and $\lambda > 0$ is a regularization parameter to strike a balance between the worse-case attack loss and the average loss (Wang et al., 2019b). We note that $\{F_{ij}\}$ are black-box functions w.r.t. $\mathbf{x}$ since the adversary has no access to the internal configuration s of DNN models to be attacked. Accordingly, the input gradient cannot be computed by back-propagation. This implies that problem (13) belongs to the *one-sided black-box* optimization problem (white-box objective w.r.t. $\mathbf{w}$). We refer readers to Appendix C for more details on (13).

Implementation details. We consider $J = 2$ DNN-based classifiers, Inception-V3 (Szegedy et al., 2016) and ResNet-50 (He et al., 2016), and $I = 2$ image classes, each of which contains 20 images randomly selected from ImageNet (Deng et al., 2009). In (13), we choose $\lambda = 5$. In Algorithm 1, we set the learning rates by $\alpha = 0.05$ and $\beta = 0.01$. Also, we use the full batch of image samples and set $q = 10$ in gradient estimation (2), where we set $\mu = 5 \times 10^{-3}$. The function query complexity is thus q $(= 10)$ times more than the number of iterations.

Baseline methods for comparison. In experiments, we compare ZO-Min-Max with (a) *FO-Min-Max*, (b) *ZO-PGD* (Ghadimi et al., 2016), and (c) *ZO-Finite-Sum*, where the method (a) is the FO counterpart of Algorithm 1, the method (b) performs single-objective ZO minimization under the equivalent form of (13), $\min_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x})$ where $h(\mathbf{x}) = \max_{\mathbf{w}\in\mathcal{W}} f(\mathbf{x},\mathbf{w})$, and the method (c) performs ZO-PGD to minimize the finite-sum (average) loss rather than the worst-case (min-max) loss. We remark that the baseline method (b) calls for the solution to the inner maximization problem $\max_{\mathbf{w}\in\mathcal{W}} f(\mathbf{x},\mathbf{w})$, which is elaborated on Appendix C. It is also worth mentioning that although ZO-Finite-Sum tackles an objective function different from (13), it is motivated by the previous work on designing the adversarial attack against model ensembles (Liu et al., 2018a), and we can fairly compare ZO-Min-Max with ZO-Finite-Sum in terms of the attack performance of obtained universal adversarial perturbations.

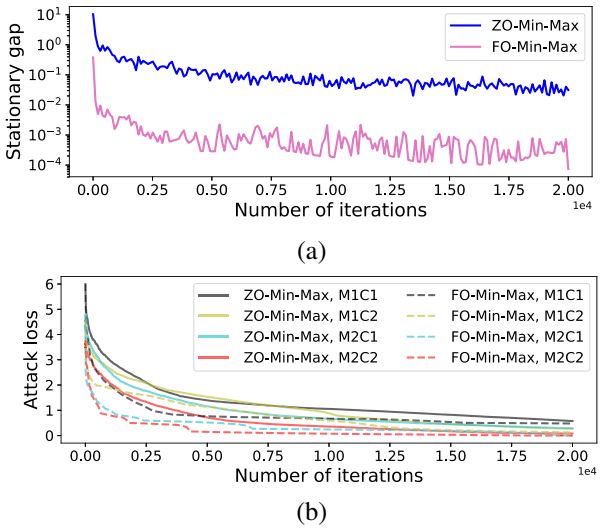


Figure 1: ZO-Min-Max vs. FO-Min-Max in attack generation. (a) Stationary gap; (b) attack loss at each model-class pair.

Results. In Figure 1, we demonstrate the empirical convergence of ZO-Min-Max, in terms of the stationary gap $\|\mathcal{G}(\mathbf{x},\mathbf{y})\|_2$ given in (6) and the attack loss corresponding to each model-class pair MjCi. Here M and C represents network model and image class, respectively. For comparison, we also present the convergence of FO-Min-Max. Figure 1-(a) shows that the stationary gap decreases as the iteration increases, and converges to an iteration-independent bias compared with FO-Min-Max. In Figure 1-(b), we see that ZO-Min-Max yields slightly worse attack performance (in terms of higher attack loss at each model-class pair) than FO-Min-Max. However, it does not need to have access to

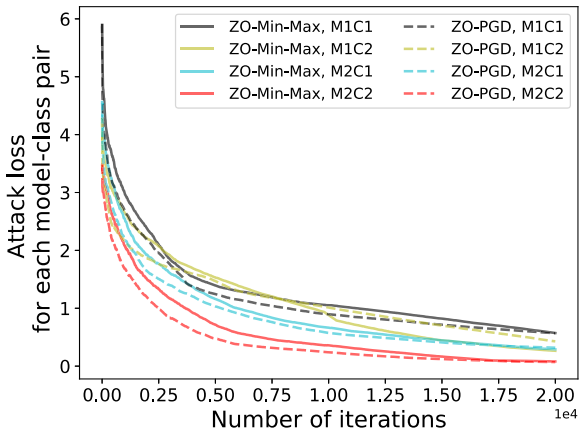


Figure 2: Comparison of attack losses achieved by ZO-Min-Max and ZO-PGD.

the configuration of the victim neural network models.

In Figure 2, we compare the attack loss of using ZO-Min-Max with that of using ZO-PGD, which solves problem (13) by calling for an internal maximization oracle in $\mathbf{w}$ (see Appendix C). As we can see, ZO-Min-Max converges slower than ZO-PGD at early iterations. That is because the former only performs one-step PGA to update $\mathbf{w}$, while the latter solves the $\mathbf{w}$-maximization problem analytically. However, as the number of iterations increases, ZO-Min-Max achieves almost the same performance as ZO-PGD.

In Appendix C, we have provided additional comparisons on ZO-Min-Max versus ZO-Finite-Sum, and versus per-image PGD attack (Madry et al., 2018).

6.2. Black-box data poisoning attack

Data poisoning against logistic regression. Let $\mathcal{D} = \{\mathbf{z}_i, t_i\}_{i=1}^n$ denote the training dataset, among which $n' \ll n$ samples are corrupted by a perturbation vector $\mathbf{x}$, leading to poisoned training data $\mathbf{z}_i + \mathbf{x}$ towards breaking the training process and thus the prediction accuracy. The poisoning attack problem is then formulated as

$$\underset{\|\mathbf{x}\|_\infty \leq \epsilon}{\text{maximize}} \ \underset{\boldsymbol{\theta}}{\text{minimize}} \ g(\mathbf{x}, \boldsymbol{\theta}) := \{F_{\text{tr}}(\mathbf{x}, \boldsymbol{\theta}; \mathcal{D}_0) + \lambda \|\boldsymbol{\theta}\|_2^2\}, \quad (14)$$

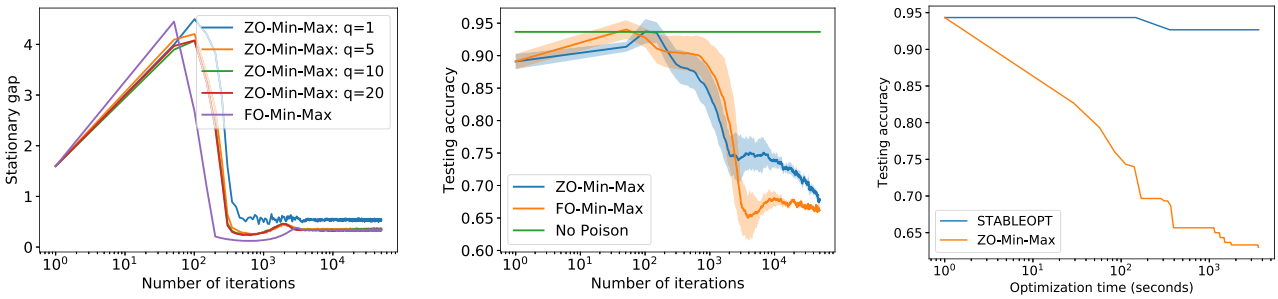
where $\mathbf{x}$ and $\boldsymbol{\theta}$ are optimization variables, $F_{\text{tr}}(\mathbf{x}, \boldsymbol{\theta}; \mathcal{D}_0)$ denotes the training loss over model parameters $\boldsymbol{\theta}$ at the presence of data poison $\mathbf{x}$, and $\lambda > 0$ is a regularization parameter. Note that problem (14) can be written in the form of (1), $\min_{\mathbf{x}} \max_{\boldsymbol{\theta}} -g(\mathbf{x}, \boldsymbol{\theta})$. Clearly, if F_{tr} is a convex loss, e.g., logistic regression or linear regression (Jagielski et al., 2018)), then $-g$ is strongly concave in $\boldsymbol{\theta}$. In (14), we assume that the adversary knows the form of the classification loss (cross entropy), however, the expression of the classifier (logistic regression model) is *not* known. Thus, $g(\mathbf{x}, \boldsymbol{\theta})$ is a *two-sided black-box* function in both $\mathbf{x}$ and $\boldsymbol{\theta}$ from the adversary's perspective.

Implementation & Baseline. In problem (14), we set the poisoning ratio $n'/n = 15\%$ and $\lambda = 10^{-3}$ for problem (14), where the sensitivity of λ is studied in Figure A4. More details on the specification of problem (14) are provided in Appendix D. In Algorithm 1, unless specified otherwise we choose $b = 100$, $q = 5$, $\alpha = 0.02$, $\beta = 0.05$, and $T = 50000$. We report the empirical results *averaged* over 10 independent trials with random initialization. We compare our method with FO-Min-Max and the state-of-the-art BO-based method STABLEOPT (Bogunovic et al., 2018).

In Figure 3, we present the convergence of ZO-Min-Max to generate data poisoning attack and validate the resulting attack performance in terms of testing accuracy of the logistic regression model trained on the poisoned dataset. Figure 3-(a) shows the stationary gap of ZO-Min-Max under different number of random direction vectors in gradient estimation (2). As we can see, a moderate choice of q (e.g., $q \geq 5$ in our example) is sufficient to achieve near-optimal solution compared with FO-Min-Max. However, it suffers from a convergence bias, consistent with Theorem 2. Figure 3-(b) demonstrates the testing accuracy (against iterations) of the model learnt from poisoned training data, where the poisoning attack is generated by ZO-Min-Max (black-box attack) and FO-Min-Max (white-box attack). As we can see, ZO-Min-Max yields promising attacking performance comparable to FO-Min-Max. We can also see that by contrast with the testing accuracy of the clean model (94% without poison), the poisoning attack eventually reduces the testing accuracy (below 70%). Furthermore In Figure 3-(c), we compare ZO-Min-Max with STABLEOPT (Bogunovic et al., 2018) in terms of testing accuracy versus computation time, where the lower the accuracy is, the stronger the generated attack is. We observe that STABLEOPT has a poorer scalability while our method reaches a data poisoning attack that induces much lower testing accuracy within 500 seconds. In Appendix D, we provide additional results on the model learnt under different data poisoning ratios.

7. Conclusion

This paper addresses black-box robust optimization problems given a finite number of function evaluations. In particular, we present ZO-Min-Max: a framework of alternating, randomized gradient estimation based ZO optimization algorithm to find a first-order stationary solution to the black-box min-max problem. Under mild assumptions, ZO-Min-Max enjoys a sub-linear convergence rate. It scales to dimensions that are infeasible for recent robust solvers based on Bayesian optimization. Furthermore, we experimentally demonstrate the potential application of the framework on generating black-box evasion and data poisoning attacks.



(a) (b) (c)

Figure 3: Empirical performance of ZO-Min-Max in design of poisoning attack: a) stationary gap versus iterations b) testing accuracy versus iterations (the shaded region represents variance of 10 random trials), and c) comparison between ZO-Min-Max and STABLEOPT on testing accuracy versus optimization time.

Acknowledgements

This work was supported by the MIT-IBM Watson AI Lab research grant. M. Hong and X. Chen were supported by NSF under the grant CIF-1910385 and in part by an AFOSR grant 19RT0424, and an ARO grant W911NF-19-1-0247.

References

Aggarwal, C., Bouneffouf, D., Samulowitz, H., Buesser, B., Hoang, T., Khurana, U., Liu, S., Pedapati, T., Ram, P., Rawat, A., et al. How can ai automate end-to-end data science? *arXiv preprint arXiv:1910.14436*, 2019.

Al-Dujaili, A., Hemberg, E., and O'Reilly, U.-M. Approximating nash equilibria for black-box games: A bayesian optimization approach. *arXiv preprint arXiv:1804.10586*, 2018a.

Al-Dujaili, A., Huang, A., Hemberg, E., and O'Reilly, U.-M. Adversarial deep learning for robust detection of binary encoded malware. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 76–82. IEEE, 2018b.

Al-Dujaili, A., Srikant, S., Hemberg, E., and O'Reilly, U.-M. On the application of danskin's theorem to derivative-free minimax optimization. *arXiv preprint arXiv:1805.06322*, 2018c.

Audet, C. and Hare, W. *Derivative-free and blackbox optimization*. Springer, 2017.

Balasubramanian, K. and Ghadimi, S. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Advances in Neural Information Processing Systems*, pp. 3455–3464, 2018.

Berahas, A. S., Cao, L., Choromanski, K., and Scheinberg, K. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *arXiv preprint arXiv:1905.01332*, 2019.

Bogunovic, I., Scarlett, J., Jegelka, S., and Cevher, V. Adversarially robust optimization with gaussian processes. In *Proc. of Advances in Neural Information Processing Systems*, pp. 5765–5775, 2018.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Branke, J. and Rosenbusch, J. New approaches to coevolutionary worst-case optimization. In *International Conference on Parallel Problem Solving from Nature*, pp. 144–153. Springer, 2008.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57. IEEE, 2017.

Chen, J., Yi, J., and Gu, Q. A frank-wolfe framework for efficient and effective adversarial attacks. *arXiv preprint arXiv:1811.10828*, 2018.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26. ACM, 2017.

Chen, T. and Giannakis, G. B. Bandit convex optimization for scalable and dynamic IoT management. *IEEE Internet of Things Journal*, 2018.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. *International Conference on Learning Representations*, 2019.

Danskin, J. M. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.

Dhurandhar, A., Pedapati, T., Balakrishnan, A., Chen, P.-Y., Shanmugam, K., and Puri, R. Model agnostic contrastive explanations for structured data. *arXiv preprint arXiv:1906.00117*, 2019.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Finlay, C. and Oberman, A. M. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

Flokas, L., Vlatakis-Gkaragkounis, E.-V., and Piliouras, G. Efficiently avoiding saddle points with zero order methods: No gradients required. *arXiv preprint arXiv:1910.13021*, 2019.

Gao, X., Jiang, B., and Zhang, S. On the information-adaptive variants of the ADMM: an iteration complexity perspective. *Optimization Online*, 12, 2014.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Gidel, G., Jebara, T., and Lacoste-Julien, S. Frank-Wolfe Algorithms for Saddle Point Problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 362–371. PMLR, 20–22 Apr 2017.

Golovin, D., Karro, J., Kochanski, G., Lee, C., Song, X., et al. Gradientless descent: High-dimensional zeroth-order optimization. *arXiv preprint arXiv:1911.06317*, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hamedani, E. Y., Jalilzadeh, A., Aybat, N. S., and Shanbhag, U. V. Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Herrmann, J. W. A genetic algorithm for minimax optimization problems. In *CEC*, volume 2, pp. 1099–1103. IEEE, 1999.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018.

Ilyas, A., Engstrom, L., and Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkMiWhR5K7.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., and Li, B. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 19–35. IEEE, 2018.

Jensen, M. T. A new look at solving minimax problems with coevolutionary genetic algorithms. In *Metaheuristics: computer decision-making*, pp. 369–384. Springer, 2003.

Jin, C., Netrapalli, P., and Jordan, M. I. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Jones, E., Oliphant, T., Peterson, P., et al. SciPy: Open source scientific tools for Python, 2001. URL http://www.scipy.org/.

Larson, J., Menickelly, M., and Wild, S. M. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

Li, Y., Li, L., Wang, L., Zhang, T., and Gong, B. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *arXiv preprint arXiv:1905.00441*, 2019.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020.

Liu, J., Zhang, W., and Yu, N. Caad 2018: Iterative ensemble adversarial attack. *arXiv preprint arXiv:1811.03456*, 2018a.

Liu, S., Chen, J., Chen, P.-Y., and Hero, A. O. Zeroth-order online admm: Convergence analysis and applications. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pp. 288–297, April 2018b.

Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Proc. of Advances in Neural Information Processing Systems*, 2018c.

Liu, S., Chen, P.-Y., Chen, X., and Hong, M. signSGD via zeroth-order oracle. In *Proc. of International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJe-DsC5Fm.

Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero, A., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning. *IEEE Signal Processing Magazine*, 2020.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Lu, S., Singh, R., Chen, X., Chen, Y., and Hong, M. Alternating gradient descent ascent for nonconvex min-max problems in robust learning and gans. In *Proc. of Asilomar Conference on Signals, Systems, and Computers*, pp. 680–684, 2019.

Lu, S., Tsaknakis, I., and Hong, M. Block alternating optimization for non-convex min-max problems: Algorithms and applications in signal processing and communications. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4754–4758, May 2019.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 2(17):527–566, 2015.

Nouiehed, M., Sanjabi, M., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

Parikh, N., Boyd, S., et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Picheny, V., Binois, M., and Habbal, A. A bayesian optimization approach to find nash equilibria. *Journal of Global Optimization*, 73(1):171–192, 2019.

Qian, Q., Zhu, S., Tang, J., Jin, R., Sun, B., and Li, H. Robust optimization over multiple domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4739–4746, 2019.

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

Rios, L. M. and Sahinidis, N. V. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3): 1247–1293, 2013.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training gans with regularized optimal transport. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, pp. 7091–7101, 2018a.

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

Schmiedlechner, T., Al-Dujaili, A., Hemberg, E., and O'Reilly, U.-M. Towards distributed coevolutionary gans. *arXiv preprint arXiv:1807.08194*, 2018.

Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*, pp. 3517–3529, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, pp. 8000–8010, 2018.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. *arXiv preprint arXiv:1805.11770*, 2018.

Wald, A. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pp. 265–280, 1945.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. *Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks*, pp. 0, 2019a.

Wang, C. and Wu, Q. Flo: Fast and lightweight hyperparameter optimization for automl. *arXiv preprint arXiv:1911.04706*, 2019.

Wang, J., Zhang, T., Liu, S., Chen, P.-Y., Xu, J., Fardad, M., and Li, B. Towards a unified min-max framework for adversarial exploration and robustness, 2019b.

Ward, R., Wu, X., and Bottou, L. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6677–6686, 2019.

Watson, R. A. and Pollack, J. B. Coevolutionary dynamics in a minimal substrate. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, pp. 702–709. Morgan Kaufmann Publishers Inc., 2001.