Zeroth-Order Methods for Convex-Concave Minmax Problems: Applications to Decision-Dependent Risk Minimization

Chinmay Maheshwari
University of California Berkeley

Chih-Yuan Chiu
University of California Berkeley

Eric Mazumdar
California Institute of Technology

Shankar Sastry
University of California, Berkeley

Lillian Ratliff
University of Washington

Abstract

Min-max optimization is emerging as a key framework for analyzing problems of robustness to strategically and adversarially generated data. We propose the random reshuffling-based gradient-free Optimistic Gradient Descent-Ascent algorithm for solving convex-concave min-max problems with finite sum structure. We prove that the algorithm enjoys the same convergence rate as that of zeroth-order algorithms for convex minimization problems. We deploy the algorithm to solve the distributionally robust strategic classification problem, where gradient information is not readily available, by reformulating the latter into a finite-dimensional convex concave min-max problem. Through illustrative simulations, we observe that our proposed approach learns models that are simultaneously robust against adversarial distribution shifts and strategic decisions from the data sources, and outperforms existing methods from the strategic classification literature.

1 INTRODUCTION

The deployment of learning algorithms in real-world scenarios necessitates versatile and robust algorithms that operate efficiently under mild information structures. Min-max optimization has been used as a tool ensure robustness in variety of domains e.g. robust optimization [Ben-Tal et al., 2009], robust control [Hast et al., 2013], to name a few. Recently, min-max optimization has emerged as a promising framework for framing problems of algorithmic robustness against adversaries [Goodfellow et al., 2014, Steinhardt et al., 2017, Madry et al., 2017], strategically generated data [Dong et al., 2018, Brown et al., 2020], and distributional shifts in dynamic environments [Yu et al., 2021].

Despite this, recent works in machine learning and robust optimization on designing and analyzing stochastic algorithms for min-max optimization problems have largely operated on a number of assumptions that preclude their application to a broad range of real-world problems e.g., access to first-order oracles that provide exact gradients [Yang et al., 2020, Nouiehed et al., 2019, Jin et al., 2020] or restrictive structural assumptions such as strong convexity [Liu et al., 2020, Wang et al., 2020, Sadiev et al., 2021]. Moreover, the developed theory is often not well-aligned with the practical implementation of these algorithms in real-world machine learning applications. For example, [Beznosikov et al., 2020] propose zeroth-order methods for convex-concave problems but the proposed algorithm may not be suitable for machine learning applications where the objective function is a sum of large numbers of component functions (depending on the size of dataset). Indeed, in order to compute the gradient estimate at any iteration Beznosikov et al requires perturbing **all** the functions which might not be suitable/possible for many applications. Furthermore, stochastic gradient methods are often used with random reshuffling (without replacement) in practice, yet their theoretical performance is usually characterized under the assumption of uniform sampling with replacement [Bottou, 2009, Jain et al., 2019].

In this work, we do away with these assumptions

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

and formulate a gradient-free (zeroth-order), random reshuffling-based algorithm with non-asymptotic convergence guarantees under mild structural assumptions on the underlying min-max objective. Our convergence guarantees are established by balancing the bias and variance of the zeroth-order gradient estimator [Bravo et al., 2018], using coupling-based arguments to analyze the correlations between iterates due to the random reshuffling procedure [Jain et al., 2019], and exploiting the recent connections between the Optimistic Gradient Descent Ascent (OGDA) and Proximal Gradient algorithms [Mokhtari et al., 2020b].

One of the primary problem areas in which such an algorithm becomes necessary is in learning from strategically generated or decision-dependent data, a classical problem in operations research (see, e.g., [Hellemo et al., 2018] and references therein). This problem has garnered a lot of attention of late in the machine learning community under the name "performative prediction" [Perdomo et al., 2020, Miller et al., 2021, Brown et al., 2020] due to the growing recognition that learning algorithms are increasingly dealing with data from strategic agents. In such problems, assuming access to the response map of strategic agents is often too restrictive, and the introduction of agent's strategic responses into a convex loss function can often result in non-convex objectives.

As an example of such a decision-dependent problem, consider a scenario in which a ride-sharing platform seeks to devise an adaptive pricing strategy which is responsive to changes in supply and demand. The platform observes the current supply and demand in the environment and adjusts the price to increase the supply of drivers (and potentially decrease the demand) as needed. Drivers, however, have the ability to adjust their availability, and can strategically create dips in supply to trigger price increases. Such gaming has been observed in real ride-share markets (see, e.g., [Hamilton, 2019, Youn, 2019]) and results in negative externalities like higher prices for passengers. Importantly, in this situation, the platform does not observe precisely the decision making process of the drivers, only their strategically generated availability, and must learn to optimize through these agents' responses. This lack of precise knowledge regarding the data generation process, and the reactive nature of the data, motivate the use of game theoretic abstractions for the decision problem, as well as algorithms for finding solutions in the absence of full information.

Previous work analyzing this problem studies this phenomenon through the lens of risk minimization in which the data distribution is decision-dependent, and seeks out settings in which the decision maker can optimize the decision-dependent risk [Miller et al., 2021]. These works, however, do not account for model misspecification in their analysis. In particular, if the data generation model is incorrect, the performance of the optimal solution returned by their training methods may potentially degrade rapidly, something we explore in our experiments.

We show that the decision-dependent learning (performative prediction) problem can be *robustified* by taking a distributional robustness perspective on the original problem. Moreover, we show that, under mild assumptions, the distributionally robust decision-dependent learning problem can be transformed to a min-max problem and hence our zeroth-order random reshuffling algorithm can be applied. The gradient-free nature of our algorithm is important for applications where data is generated by strategic users that one must query; in these scenarios, the decision-maker is unlikely to access the best response map (data generation mechanism) of the strategic users, and hence will lack access to precise gradients.

Contributions. In this paper, we analyze the class of convex-concave min-max problems given by

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

where $\mathcal{X}\subset\mathbb{R}^{d_x}$, $\mathcal{Y}\subset\mathbb{R}^{d_y}$, and $L:\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\to\mathbb{R}$ has the finite-sum structure given by $L:=\frac{1}{n}\sum_{i=1}^{n}L_i$, where $L_1,\dots,L_n:\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\to\mathbb{R}$ denote n individual loss functions. This formulation is ubiquitous in machine learning applications, where the overall loss objective is often the average of the loss function evaluated over each data point in a dataset.

The contributions of this paper can be summarized as follows.

I) We propose an efficient zeroth-order random reshuffling-based OGDA algorithm for a convex-concave min-max optimization problem, *without* assuming any other structure on the curvature of the min-max loss (e.g., strong convexity or strong concavity). We provide (to our knowledge) the *first* non-asymptotic analysis of OGDA algorithm *with* random reshuffling and zeroth-order gradient information.

II) As an important application, we formulate the *Wasserstein distributionally robust learning with decision-dependent data* problem as a constrained finite-dimensional, smooth convex-concave min-max problem of the form (1). In particular, we consider the setting of learning from strategically generated data, where the goal is to fit a generalized linear model, and where an ambiguity set is used to capture model misspecification regarding the data generation process. This setting encapsulates a distributionally robust version of the recently introduced problem of strategic classification [Hardt et al., 2016]. We show that this problem, under mild assumptions on data generation model and the ambiguity set, can be transformed into a convex-concave min-max problem to which our algorithm applies.

III) We complement the theoretical contributions of this paper by presenting illustrative numerical examples.

2 RELATED WORK

Our work draws upon the existing literature on zeroth-order methods, random reshuffling-based methods, decision-dependent learning, and distributionally robust optimization.

Zero-Order Methods for Min-Max Optimization. Zeroth-order methods provide a computationally efficient method for applications in which first-order or higher-order information is inaccessible or impractical to compute, e.g., when generating adversarial examples to test the robustness of black-box machine learning models [Liu et al., 2020, Chen et al., 2017, Ilyas et al., 2018, Tu et al., 2019]. Recently, Liu et al. (2020) and others [Gao et al., 2018] provided the first non-asymptotic convergence bounds for zeroth-order algorithms, based on analysis methods for gradient-free methods in convex optimization [Nesterov and Spokoiny, 2017]. However, these works assume that the min-max objective is either strongly concave in the maximizing variable [Liu et al., 2020, Wang et al., 2020] or strongly convex [Gidel et al., 2017] in the minimizing variable, an assumption that fails to hold in many applications [Dong et al., 2018, Yu et al., 2021]. In contrast, the zeroth-order algorithm presented in this work provides non-asymptotic guarantees under the less restrictive assumption that the objective function is convex-concave.

Random reshuffling of data sets. In single-variable optimization problems, first-order stochastic gradient descent algorithms are empirically observed to converge faster when random reshuffling (RR, or sampling without replacement) is deployed, compared to sampling with replacement [Recht and Ré, 2011, Bottou, 2009]. Although considerably more difficult to analyze theoretically, gradient-based RR methods have recently been shown to enjoy faster convergence when the underlying objective function is convex [Shamir, 2016, Jain et al., 2019, HaoChen and Sra, 2019, Safran and Shamir, 2019]. Recently, these theoretical results have been extended to first-order methods for convex-concave min-max optimization problems [Yu et al., 2021].

Distributionally Robust Optimization. *Distributionally Robust Optimization (DRO)* seeks to find solutions to optimization problems (e.g., supervised learning tasks) robust against changes in the data distribution between training and test time [Madry et al., 2017, Yu et al., 2021]. These distributional differences may arise due to imbalanced data, sample selection bias, or adversarial perturbations or deletions [Candela et al., 2009, Madry et al., 2017], and are often modeled as min-max optimization problems, in which the classifier and an adversarial noise component are respectively modeled as the minimizer and maximizer of a common min-max loss objective [Bagnell, 2005, Bertsimas et al., 2010, Rahimian and Mehrotra, 2019]. In particular, the noise is assumed to generate the worst possible loss corresponding to a bounded training data distribution shift, with the bound given by either the f-divergence or Wasserstein distance. [Yu et al., 2021, Ben-Tal et al., 2013, Namkoong and Duchi, 2016, Hu et al., 2018, Shafieezadeh-Abadeh et al., 2015]. While these works consider adversarial noise in generated data, largely in a worst-case context, it has yet to capture strategically generated data wherein a data source generates data via a best response mapping.

Strategic Classification and Performative Prediction. *Strategic classification* [Hardt et al., 2016, Dong et al., 2018, Sessa et al., 2020] and *performative prediction* [Perdomo et al., 2020, Miller et al., 2021, Drusvyatskiy and Xiao, 2020] concern supervised learning problems in which the training data distribution shifts in response to the deployed classifier or predictor more generally. This setting naturally arises in machine learning applications in which the selection of the deployed classifier either directly changes the training data (e.g., decisions based on credit scores, such as loan approvals, themselves change credit scores), or prompts the data source to artificially alter their attributes (e.g. withdrawals during bank runs spur worried clients to make more withdrawals) [Perdomo et al., 2020, Miller et al., 2021]. Here, the learner accesses only perturbed features representing the strategic agents' best responses to a deployed classifier, and not the true underlying features [Dong et al., 2018]. This is a recently introduced formulation to machine learning; the results in this body of literature (to our knowledge) have not introduced the concept of robustness to model misspecification or the data generation process, in the same manner as we capture in this work.

3 PRELIMINARIES

Recall that in this paper, we consider the class of convex-concave min-max problems given by:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \; L(x,y), \tag{2}$$

where $\mathcal{X}\subset\mathbb{R}^{d_x}$, $\mathcal{Y}\subset\mathbb{R}^{d_y}$, and $L := \frac{1}{n}\sum_{i=1}^{n} L_i$, where $L_1,\ldots,L_n : \mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\to\mathbb{R}$ denote n individual loss functions. For convenience, we denote $d := d_x + d_y$.

Assumption 3.1. *The following statements hold:*

(i) The sets $\mathcal{X}\subset\mathbb{R}^{d_x}$ and $\mathcal{Y}\subset\mathbb{R}^{d_y}$ are convex and compact.

(ii) The functions $L_1,\ldots,L_n : \mathbb{R}^d\to\mathbb{R}$ are convex in $x\in\mathbb{R}^{d_x}$ for each $y\in\mathbb{R}^{d_y}$, concave in $y\in\mathbb{R}^{d_y}$ for each $x\in\mathbb{R}^{d_x}$, and G-Lipschitz and ℓ-smooth in $(x,y)\in\mathbb{R}^d$ (which implies that $L:\mathbb{R}^d\to\mathbb{R}$, by definition, also possesses the same properties).

For ease of exposition, we denote $u := (x,y)$, $M_L := \sup_{u\in\mathcal{X}\times\mathcal{Y}} |L(u)|$, $D := \sup_{u,u'\in\mathcal{X}\times\mathcal{Y}} \|u-u'\|_2$, and define the operators $F, F_i : \mathbb{R}^d\to\mathbb{R}^d$, for each $i\in[n]$, by:

$$F(u) := \begin{bmatrix}\nabla_x L(u)\\ -\nabla_y L(u)\end{bmatrix}, \quad F_i(u) := \begin{bmatrix}\nabla_x L_i(u)\\ -\nabla_y L_i(u)\end{bmatrix}. \tag{3}$$

Observe that under Assumption 3.1, $M_L, D<\infty$, and F and each F_i are monotone[1]. Finally, we define the *gap function* $\Delta:\mathbb{R}^d\to[0,\infty)$ associated with the loss L by

$$\Delta(x,y) := L(x,y^\star) - L(x^\star,y) \geq 0, \tag{4}$$

where $u^\star := (x^\star,y^\star)\in\mathcal{X}\times\mathcal{Y}$ denotes the min-max saddle point of the overall loss $L(x,y)$, and $(x,y)\in\mathcal{X}\times\mathcal{Y}$ denotes any feasible point. This gap function allows us to measure the convergence rate of our proposed algorithm. To this end, we define the ϵ-optimal saddle-point of (2) as follows.

Definition 3.1 (**ϵ-optimal saddle point solution**). *A feasible point $(x,y)\in\mathcal{X}\times\mathcal{Y}$ is said to be an ϵ-optimal saddle-point solution of (2) if*

$$\Delta(x,y) = L(x,y^\star) - L(x^\star,y) \leq \epsilon.$$

4 ALGORITHMS & ANALYSIS

In this section we introduce a gradient-free version of the well-studied Optimistic Gradient Descent Ascent (OGDA) algorithm, and provide non-asymptotic rates showing that it can efficiently find the saddle point in constrained convex-concave problems.

4.1 Zeroth-Order Gradient Estimates

In our zeroth-order, random reshuffling-based variant of the OGDA algorithms, we use the one-shot randomized gradient estimator [Spall, 1997, Flaxman et al., 2005]. In particular, given the current iterate $u\in\mathbb{R}^d$ and a *query radius* $R>0$, we sample a vector v uniformly from unit sphere $\mathcal{S}^{d-1}$ (i.e. $v\sim\mathsf{Unif}(\mathcal{S}^{d-1})$), and define the zeroth-order estimator $\hat{F}(u;R,v)\in\mathbb{R}^d$ of the min-max loss $L(u)$ to be:

$$\hat{F}(u;R,v) := \frac{d}{R} L(u+Rv)v \tag{5}$$

Properties of this zeroth-order estimator are stated in Proposition A.4 in supplementary material.

4.2 Optimistic Gradient Descent Ascent with Random Reshuffling (OGDA-RR)

In this subsection, we formulate our main algorithm, Optimistic Gradient Descent Ascent with Random Reshuffling (OGDA-RR). In each *epoch* $t\in\{0,1,\cdots,T-1\}$, the algorithm generates a uniformly random permutation $\sigma^t := (\sigma_1^t,\cdots,\sigma_n^t)$ of $[n] := \{1,\cdots,n\}$ independently of any other randomness. This is what is referred as random reshuffling (or sampling without replacement) where within every epoch we do not re-sample and this naturally gives rise to correlations between different iterations within an epoch. Furthermore, the algorithm fixes a query radius $R^t>0$ and search direction $v_i^t\in\mathbb{R}^d$ in every epoch t. Note that query radii only depends on epoch indices t, and not on sample indices $\{\sigma_i^t\}_{i=1}^n$. For each $i\in[n], t\in[T-1]$, we compute the OGDA-RR update as follows:

$$\begin{aligned} u_{i+1}^t = \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}} \Big(& u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t;R^t,v_i^t) \\ & - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t;R^t,v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;R^t,v_{i-1}^t)\Big), \end{aligned} \tag{6}$$

where the terms $\hat{F}_{\sigma_i^t}$ and $\hat{F}_{\sigma_{i-1}^t}$ are the zeroth-order estimators of gradients $F_{\sigma_i^t}$ and $F_{\sigma_{i-1}^t}$ (defined in (3)).

After repeating this process for T epochs, the algorithm returns the step-size weighted average of the iterates, $\tilde{u}^T := \frac{1}{n\cdot\sum_{t=0}^{T-1}\eta^t}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t u_i^t$. The following theorem states that if we run Algorithm 1 long enough then $\tilde{u}^T$ will be close to the saddle point.

Theorem 4.1. *Let $L(u)$ denote the objective function in the constrained min-max optimization problem given by (1), and let $u^\star = (x^\star,y^\star)\in\mathcal{X}\times\mathcal{Y}$ denote any saddle point of $L(u)$. Fix $\epsilon>0$. Suppose Assumption 3.1 holds, and the number of epochs T, step sizes sequence*

[1] A function $F:\mathbb{R}^d\to\mathbb{R}^d$ is called *monotone* if $\langle F(x)-F(y), x-y\rangle \geq 0$ for all $x,y\in\mathbb{R}^d$.

Algorithm 1: OGDA-RR Algorithm

```
Input: stepsizes $\eta^t, R^t$, data points
    $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}, u_0^{(0)}$, time horizon duration $T$;
for $t = 0, 1, \cdots, T-1$ do
    $\sigma^t = (\sigma_1^t, \cdots, \sigma_n^t) \leftarrow$ a random permutation of
      set $[n]$;
    for $i = 0, \ldots, n-1$ do
        Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
        $u_{i+1}^t \leftarrow (6)$
    end
    $u_0^{(t+1)} \leftarrow u_n^t$
    $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
end
Output: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$.
```

$\{\eta^t\}_{t=0}^{T-1}$, and query radius sequence $\{R^t\}_{t=0}^{T-1}$ satisfy:

$$\eta^t := \eta^0 \cdot (t+1)^{-3/4+\chi}, \qquad \forall t \in \{0, 1, \cdots, T-1\},$$
$$R^t := R^0 \cdot (t+1)^{-1/4}, \qquad \forall t \in \{0, 1, \cdots, T-1\},$$
$$T > \frac{1}{\epsilon^4}\left(\frac{3}{16n}D + \frac{5}{4}C \cdot \max\left\{R^0, \eta^0, \eta^0 R^0, \frac{\eta^0}{R^0}, \frac{\eta^0}{(R^0)^2}\right\}\left(1 + \frac{1}{\chi}\right)\right)^{\frac{4}{1-4\chi}},$$

for some initial step size $\eta^0 \in \left(0, \frac{1}{2\ell}\right)$, initial query radius $R^0 > 0$, parameter $\chi \in (0, 1/4)$, and constant $C = \mathcal{O}(nd^2 D)$. Then the iterates $\{u_i^t\}$ generated by the OGDA-RR Algorithm (Alg. 1) satisfy:

$$\mathbb{E}\left[\Delta(\tilde{u}^T)\right] < \epsilon.$$

There are three main components to the proof of Theorem 4.1: First, we bound the bias introduced due to random reshuffling (or sampling without replacement) by Wasserstein distance between two appropriate distributions that characterize the correlations introduced between iterates because of random reshuffling. Second, we bound the aforementioned Wasserstein distance by constructing an appropriate coupling between iterates generated with and without random reshuffling [Jain et al., 2019]. The coupled iterates thus obtained are then bounded by exploiting the recent connections between OGDA method and proximal point methods [Mokhtari et al., 2020a], which is one of the main contributions of our proof technique. Third, we balance the bias and variance introduced due to zeroth-order gradient estimator by suitably choosing the step size sequence $\{\eta^t\}$ and the perturbation radius sequence $\{R^t\}$. The details of the proof of Theorem 4.1 are deferred to Appendix A.2.

Remark. Note that one can obtain better convergence rates if we use a multi-point zeroth-order estimator as opposed to the single-point zeroth-order estimator (5). For instance, if we use the following two-point gradient estimator:

$$\hat{F}(u; R, v) = \frac{d}{2R}(L(u + Rv) - L(u - Rv))v$$

then it follows easily from our analysis that the epochs required to obtain an $\epsilon-$optimal saddle point decreases from $\tilde{\mathcal{O}}\left(\frac{n^4 d^8}{\epsilon^4}\right)$ to $\tilde{\mathcal{O}}\left(\frac{n^2 d^4}{\epsilon^2}\right)$. But we restrict our presentation to single point estimators, as the application presented in Sec. 5 demands that we query the objective function as minimally as possible. It is an interesting future research direction to study the OGDA-RR algorithm with more advanced zeroth-order methods.

Remark. The analysis of OGDA algorithm with random reshuffling and exact gradient information is an immediate feature of our proof technique. For such algorithms, the number of epochs required to obtain an $\epsilon-$optimal saddle point is $\tilde{\mathcal{O}}\left(\frac{n^2}{\epsilon^2}\right)$. Note that there is no dependence on d with exact gradient based methods.

Remark. Note that the OGDA-RR algorithm is computationally more efficient than Alg. 2 in Yu et. al. (2021), *if* one replaces the gradient estimates with true gradient values. This is because that algorithm requires $\mathcal{O}(\log(n))$ inner loop iterations to approximate a proximal point update. Here, we overcome extra computations by exploiting the recent perspective that the OGDA update is a perturbed proximal point update [Mokhtari et al., 2020b].

5 APPLICATIONS TO DECISION-DEPENDENT DRO

In this section we discuss a novel convex-concave min-max reformulation of a class of decision-dependent distributional robust risk minimization problems, which reflects the need for learning classifiers that are simultaneously robust to strategic data sources and adversarial model-specification. In particular, we present a distributionally robust formulation of strategic classification [Dong et al., 2018] with generalized linear loss, a semi-infinite optimization problem that can be reformulated to a finite-dimensional convex-concave min-max problem.

Strategic classification is an emerging paradigm in machine learning which attempts to “close the loop"— i.e., account for data (user) reaction at training time—while designing classifiers to be deployed in strategic environments in the real world, where deploying *naïve* classifier (designed ignoring the distribution shift) can be catastrophic. Modeling the exact behavior of such strategic interactions is very complex, since the decision-maker (learner) does not have access to the strategic users' preferences and hence lacks access to their best response function. To overcome this difficulty, we use a natural model for these strategic behaviors that has been exploited in Dong et. al.(2018), and then impose robustness conditions (in the form of an ambiguity set on the decision-dependent data distribution) to capture model misspecification. To facilitate the discussion, we provide a primer on decision-dependent DRO in the next subsection.

5.1 Primer on decision-dependent distributionally robust optimization

Consider a *generalized linear problem*, where the goal is to estimate the parameter $\theta \in \Theta$, which is assumed to be a compact set, by solving the following convex optimization program:

$$\inf_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}} \left[\phi \left(\langle \bar{x}, \theta \rangle \right) - \bar{y} \langle \bar{x}, \theta \rangle \right]$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a smooth convex function and the tuple $(\bar{x}, \bar{y}) \in \mathbb{R}^d \times \{-1, +1\}$ is sampled from an unknown distribution $\mathcal{D}$, often approximated by the empirical distribution of a set of observed data. The generalized linear model encompasses a wide range of machine learning formulations [McCullagh and Nelder, 2019].

A distributionally robust generalized linear problem, on the other hand, minimizes the worst case expectation over an uncertainty set $\mathcal{P}$ in the space of probability measures. This setup can be envisioned as a game between a learning algorithm and an adversary. Based on parameters chosen by the learning algorithm, the adversary then picks a probability measure from the uncertainty set which maximizes the risk for that choice of parameter:

$$\inf_{\theta} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\phi \left(\langle \bar{x}, \theta \rangle \right) - \bar{y} \langle \bar{x}, \theta \rangle \right],$$

where $(\bar{x}, \bar{y}) \sim \mathbb{P} \in \mathcal{P}$. Typically $\mathcal{P}$ is chosen as a Wasserstein ball around the empirical distribution $\tilde{\mathcal{D}}_n$ of a set of n observed data points, $\{(\tilde{x}_i, \tilde{y}_i) \in \mathbb{R}^d \times \{-1, 1\}\}_{i=1}^n$, sampled independently from the data distribution $\mathcal{D}$. Then, for any $\delta > 0$ the uncertainty set $\mathcal{P}$ is given by $\mathbb{B}_\delta(\tilde{\mathcal{D}}_n) = \{\mathbb{P} : \mathcal{W}(\mathbb{P}, \tilde{\mathcal{D}}_n) \leq \delta\}$.

A critique of the above problem formulation is that the underlying data distribution $\mathcal{D}$ is considered fixed, while in many strategic settings underlying data distribution will depend on the classifier parameter θ. *Decision-dependent supervised learning* aims to tackle such distribution shifts. When specialized to the generalized linear model, the problem formulation becomes:

$$\inf_{\theta} \mathbb{E}_{\mathcal{D}(\theta)} \left[\phi \left(\langle \bar{x}, \theta \rangle \right) - \bar{y} \langle \bar{x}, \theta \rangle \right],$$

where $(\bar{x}, \bar{y}) \sim \mathcal{D}(\theta)$. In this work, we take a step forward and work with the *distributionally robust decision-dependent generalized linear model*, defined as:

$$\inf_{\theta} \sup_{\mathbb{P} \in \mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}} \left[\phi \left(\langle \bar{x}, \theta \rangle \right) - \bar{y} \langle \bar{x}, \theta \rangle \right], \tag{7}$$

where $(\bar{x}, \bar{y}) \sim \mathbb{P} \in \mathcal{P}(\theta)$ and $\mathcal{P}(\theta) = \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta))$. Here, the dependence of $\mathbb{P}$ on the choice of classifier θ is captured by its inclusion in $\mathcal{P}(\theta) = \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta))$. To describe decision-dependent distribution shifts $\tilde{\mathcal{D}}_n(\theta)$, we restrict our focus to the setting of strategic classification. The following subsection formalizes our setting.

5.2 Model for strategic response

Below, we denote the data points sampled from *true distribution* by $(\tilde{x}_i, \tilde{y}_i) \sim \mathcal{D}$ where $\mathcal{D}$ is a unknown, underlying distribution. For ease of presentation, we associate each data point index i with an agent. For each agent $i \in [n]$, let $u_i(x; \theta, \tilde{x}_i, \tilde{y}_i) \in \mathbb{R}$ denote its utility function that a strategic agent seeks to maximize. In other words, when a classifier parametrized by $\theta \in \mathbb{R}^d$ is deployed, the agent $i \in [n]$ responds by reporting $b_i(\theta, \tilde{x}_i, \tilde{y}_i)$, defined as:

$$b_i(\theta, \tilde{x}_i, \tilde{y}_i) \in \arg\max_{x} u_i(x; \theta, \tilde{x}_i, \tilde{y}_i).$$

Note that we allow different agent to have different utility function.

We now impose the following assumptions on the utility functions; these are crucial for ensuring guaranteed convergence of our proposed algorithms.

Assumption 5.1. *For each agent $i \in [n]$, define $u_i(x; \theta, \tilde{x}_i, \tilde{y}_i) := \frac{1 - \tilde{y}_i}{2} \langle x, \theta \rangle - g_i(x - \tilde{x}_i)$, where $g_i : \mathbb{R}^d \to \mathbb{R}$ satisfies:*

(i) $g_i(x) > 0$ *for all* $x \neq 0$*;*
(ii) g_i *is convex on* $\mathbb{R}^d$*;*
(iii) g_i *is positive homogeneous*[2] *of degree* $p > 1$*;*
(iv) *Its convex conjugate* $g_i^*(\theta) := \sup_{x \in \mathbb{R}^d} \langle x, \theta \rangle - g_i(x)$ *is* G_i*-Lipschitz and* $\bar{G}_i$*-smooth on* Θ*.*

As is pointed out in Dong et. al. (2018), a large class of functions $g(\cdot)$ satisfy the requirements posited in Assumption 5.1. For example, for any arbitrary norm and any $p > 1$ the function $g(x) = \frac{1}{p} \|x\|^p$ is a candidate. Note that these assumptions are not very restrictive and capture a large variety of practical scenarios [Dong et al., 2018]. A natural consequence of the above modeling paradigm is that $b_i(\theta, \tilde{x}_i, +1) = \tilde{x}_i$. To wit, the agents act strategically only if their true label is -1. This is a reasonable setting for many real

[2] A function $f : \mathbb{R}^d \to \mathbb{R}$ is *positive homogenous of degree* r if for any scalar $\alpha > 0$ and $x \in \mathbb{R}^d$ we have $f(\alpha x) = \alpha^r f(x)$

world applications [Dong et al., 2018]. We now present a technical lemma which will be helpful in subsequent presentation.

Lemma 5.2 (Dong et. al. (2018))**.** *Under Assumption 5.1, for each agent $i \in [n]$, the set of best responses $\arg\max_x u_i(x;\theta,\tilde{x}_i,\tilde{y}_i)$ is finite and bounded. The function $\theta \mapsto \langle b_i(\theta,\tilde{x}_i,\tilde{y}_i),\theta\rangle$ is convex. To wit, for any $i \in [n]$: $\langle b_i(\theta,\tilde{x}_i,\tilde{y}_i),\theta\rangle = \langle \tilde{x}_i,\theta\rangle + \frac{1-\tilde{y}_i}{2} q g_i^*(\theta)$ where $\frac{1}{p}+\frac{1}{q}=1$*

Against the preceding backdrop, we now present the convex-concave min-max reformulation of the Wasserstein Distributionally Robust Strategic Classification (WDRSC) problem.

5.3 Reformulation of the WDRSC Problem

The WDRSC problem formulation contains two main components—the *strategic component* that accounts for a distribution shift $\mathcal{D}(\theta)$ in response to the choice of classifier θ, and the *adversarial component* that accounts for the uncertainty set $\mathcal{P}(\theta)$. As per the modeling assumptions described in Section 5.2, we have $(\tilde{x}_i,\tilde{y}_i) \sim \mathcal{D}$ and $(b_i(\theta,\tilde{x}_i,\tilde{y}_i),\tilde{y}_i) \sim \mathcal{D}(\theta)$ for all $i \in [n]$. We now impose certain restrictions on the adversarial component that would enable us to reformulate the WDRSC problem as a convex-concave min-max optimization problem. Crudely speaking, we allow adversarial modifications on *features* for all data points, but adversarial modifications on *labels* only when the true label is $+1$.

For the distributionally robust strategic classification problem, we consider a specific form of uncertainty set $\mathcal{P}(\theta)$ that allows us to reformulate the infinite-dimensional optimization problem as a finite-dimensional convex-concave min-max problem. As described above, in our formulation, the features of a given data point i can be perturbed strategically if $\tilde{y}_i = -1$, but not if $\tilde{y}_i = +1$. On top of the strategic perturbations we also consider the adversarial perturbations to the data points. Specifically, we also assume that the adversary can perturb both the features and label of a data point i if $\tilde{y}_i = 1$, but can only perturb the features and not the label if $\tilde{y}_i = -1$. A rigorous exposition of this restriction is deferred to Appendix B.1. Under these assumptions, we now present a convex-concave min-max reformulation of the WDRSC problem.

Theorem 5.3. *Let the strategic behavior of the agents be governed in accordance with Assumption 5.1. Suppose ϕ is convex and β-smooth. In addition, suppose $\mathbb{R} \ni x \mapsto \phi(x) + x \in \mathbb{R}$ is non-decreasing. Then the WDRSC problem* (7) *can be reformulated into the following convex-concave min-max problem:*

$$\begin{aligned}
\min_{(\theta,\alpha)} \max_{\gamma\in\mathbb{R}^n} \Bigg\{ & \alpha(\delta-\kappa) + \frac{1}{n}\sum_i \frac{1+\tilde{y}_i}{2}\left(\phi\left(\langle b_i(\theta),\theta\rangle\right)\right) \\
& + \gamma_i\left(\langle b_i(\theta),\theta\rangle - \alpha\kappa\right) \\
& + \frac{1}{n}\sum_i \frac{1-\tilde{y}_i}{2}\left(\phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle\right) \Bigg\} \\
s.t. \|\theta\| \le \alpha/(\beta+1),\ & \|\gamma\|_\infty \le 1
\end{aligned} \tag{8}$$

where for any $i \in [n]$, we have concisely written $b_i(\theta,\tilde{x}_i,\tilde{y}_i)$ as $b_i(\theta)$.

The proof of Theorem 5.3 is presented in Appendix B.2.

Remark. The non-decreasing assumption on the map $\mathbb{R} \ni x \mapsto \phi(x) + x \in \mathbb{R}$ is not overly restrictive; in fact, it is satisfied by the logistic regression model in supervised learning (see Appendix C).

Remark. Note that we can convert the smooth convex-concave minmax problem (8) into a non-smooth convex minimization problem by explictly taking maximization over γ. But we refrain from doing as it has been observed [Yu et al., 2021] that solving the smooth minimax optimization problem is faster than solving the non-smooth problem. In fact, we have presented an experimental study in Appendix C which corroborates this observation.

Throughout the rest of this paper, we denote the min-max objective in (8) by $L(\alpha,\theta,\gamma)$.

6 EMPIRICAL RESULTS

In this section we deploy zeroth-order OGDA algorithm with random reshuffling to solve the convex concave reformulation of WDRSC as presented in (8). We point out that in order to solve (8), the zeroth-order method should only be applied to estimate the gradient with respect to θ. This is because the gradient with respect to other variables, namely (α,γ), can be exactly computed. Specifically, to compute derivative with respect to θ the designer must know the best response function which is often not available and it can only be queried.

We now present some illustrations of the empirical performance of our proposed algorithm, as well as empirical justification for solving the WDRSC problem over existing prior approaches to strategic classification.

6.1 Experimental Setup

Our first set of empirical results uses synthetic data to illustrate the effectiveness of our algorithms. The datasets used in this section are constructed as follows: the ground truth classifier $\theta^\star$ and features $\tilde{x}_i$ are sampled as $\theta^\star \sim \mathcal{N}(0, I_d)$ and $\tilde{x}_i \sim$ i.i.d. $\mathcal{N}(0, I_d)$, for each $i \in [n]$, while the ground truth labels $\tilde{y}_i$ are given by $\tilde{y}_i = \mathsf{sign}(\langle \tilde{x}_i, \theta^\star \rangle + z_i)$ for each $i \in [n]$, where $z_i \sim$ i.i.d. $\mathcal{N}(0, 0.1 \cdot I_d)$. We use $n \in \{500, 1000\}$ with $d = 10$. The first five of the $d = 10$ features are chosen to be strategic. In all experiments, we take $\kappa = 0.5$ and $\delta = 0.4$. Each strategic agent $i \in [n]$ has a utility function given by:

$$u_i(x; \theta, \tilde{x}_i, \tilde{y}_i, \zeta_i) = \frac{1 - \tilde{y}_i}{2} \langle x, \theta \rangle - \frac{1}{2\zeta_i} \|x - \tilde{x}_i\|^2, \quad (9)$$

where ζ_i denote the perturbation "power" of agent i. For simplicity, we assume all agents are homogeneous, in the sense that $\zeta_i = \zeta > 0$ for all $i \in [n]$; in practice, one need not impose this assumption. Given this utility function, the best response of agents takes the form:

$$b_i(\theta, \tilde{x}_i, \tilde{y}_i; \zeta) = \begin{cases} \tilde{x}_i & \text{if } \tilde{y}_i = +1, \\ \tilde{x}_i + \zeta\theta & \text{if } \tilde{y}_i = -1 \end{cases} \quad (10)$$

where, in our simulations, we fix $\zeta = 0.05$. We reemphasize that our algorithm does not use the value of ζ in any of its computations. For purposes of illustration, we focus on the performance of the following algorithms:

(A-I) Zeroth-order optimistic-GDA *with* random reshuffling (see Algorithm 1),
(A-II) Zeroth-order optimistic-GDA *without* random reshuffling (see Appendix C),
(A-III) Zeroth-order stochastic-GDA *with* random reshuffling (see Appendix C),
(A-IV) Zeroth-order stochastic-GDA *without* random reshuffling (see Appendix C).

and we evaluate the proposed algorithms and model formulation on two criteria:

(i) *Suboptimality*: To measure suboptimality, we use the *gap function* $\Delta(\alpha, \theta, \gamma) = L(\alpha, \theta, \gamma^\star) - L(\alpha^\star, \theta^\star, \gamma)$ (Def. 4) where $(\alpha^\star, \theta^\star, \gamma^\star)$ is a solution of the min-max reformulation (8) of the WDRSC problem. If the objective $L(\cdot)$ is convex-concave, $\Delta(\cdot)$ is non-negative, and equals zero at (and only at) saddle points.

(ii) *Accuracy*: Given a data set $\{(\tilde{x}_i, \tilde{y}_i)\}_{i \in [n]}$, the accuracy of a classifier θ is measured as $\frac{1}{n} \sum_{i \in [n]} \tilde{y}_i \langle b_i(\theta, \tilde{x}_i, \tilde{y}_i; \zeta), \theta \rangle$. Under this criterion we compare the accuracy under different perturbations for different classifiers θ; To compute suboptimality, we first compute a true min-max saddle point $(\alpha^\star, \theta^\star, \gamma^\star)$ via a first order gradient based algorithm (namely, GDA). All experiments were run using Python 3.7 on a standard MacBook Pro laptop (2.6 GHz Intel Core i7 and 16 GB of RAM).

6.2 Results

Simulation results presented in Figure (1a)-(1b) show that our proposed algorithm (i.e. (A-I)) outperforms algorithms without reshuffling (i.e. (A-II) and (A-IV)). However, its performance resembles that of zeroth-order stochastic-GDA with random reshuffling. More experimental studies need to be conducted to more conclusively determine whether (A-I) outperforms (A-III), or vice versa. In fact , there has been no theoretical investigations even for the *first order* stochastic-GDA algorithm with random reshuffling; this is an interesting future direction to explore.

In Figure 1, we also compare the robustness of the classifier obtained by using Algorithm (A-I) with that obtained from prior work on solving probems of strategic classification trained with $\zeta = 0.05$ (referred as *LogReg SC* in Figure 1). As expected, due to the formulation, the performance of the classifier obtained via (A-I) degrades gracefully even when subject to large perturbations, while the performance of existing approaches to strategic classification degrades rapidly. Further numerical results on synthetically generated and real world datasets are given in Appendix C.

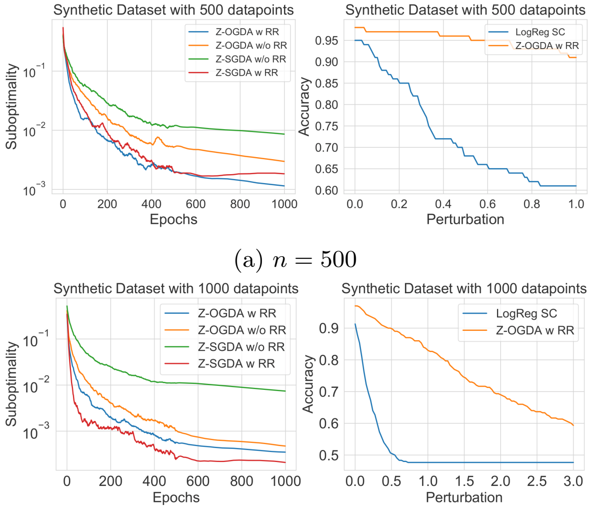
7 CONCLUSION AND FUTURE WORK

7.1 Summary

This paper presents the first non-asymptotic convergence rates for a gradient-free optimistic min-max optimization algorithm with random reshuffling. Our theoretical results, established for smooth convex-concave min-max objectives, do not require any additional, restrictive structural assumptions to hold. As a concrete application, we reformulate a distributionally robust strategic classification problem as a convex-concave min-max optimization problem that can be iteratively solved using our method. Empirical results on synthetic and real datasets demonstrate the efficiency and effectiveness of our algorithm, as well as its robustness against adversarial distributional shifts and strategic behavior of the data sources.

7.2 Current Limitations and Future Work

Potential directions for future work include establishing convergence results for the random reshuffling-based



(a) $n = 500$

(b) $n = 1000$

Figure 1: Experimental results for a synthetic dataset with $n = 500$ and $n = 1000$. (Left panes of (1a), (1b))) Suboptimality iterates generated by the four algorithms (A-I), (A-II), (A-III), (A-IV), respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right panes of (1a), (1b))) Comparison between decay in accuracy of strategic classification with logistic regression (trained with $\zeta = 0.05$) and Alg. (A-I) with change in perturbation.

Stochastic Gradient Descent Ascent (SGDA-RR) algorithm, as well as performing more extensive experimental studies to better understand the empirical performance of our algorithm. In addition, the assumptions posited on the uncertainty set in WDRSC problem formulation, in Section 5, could be relaxed.

Acknowledgements

Research supported by NSF under grant DMS 2013985 “THEORINet: Transferable, Hierarchical, Expressive, Optimal, Robust and Interpretable Networks”. Chinmay Maheshwari, Chih-Yuan Chiu and S. Shankar Sastry were also supported in part by U.S. Office of Naval Research MURI grant N00014-16-1- 2710.

References

[Bagnell, 2005] Bagnell, J. A. (2005). Robust supervised learning. In *AAAI*, volume 2, pages 714–719.

[Ben-Tal et al., 2009] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*. Princeton university press.

[Ben-Tal et al., 2013] Ben-Tal, A., Hertog, D. D., Waegenaere, A. D., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.*, 59:341–357.

[Bertsimas et al., 2010] Bertsimas, D., Brown, D., and Caramanis, C. (2010). Theory and applications of robust optimization. *SIAM Review*, 53.

[Beznosikov et al., 2020] Beznosikov, A., Sadiev, A., and Gasnikov, A. (2020). Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. *ArXiv*.

[Bottou, 2009] Bottou, L. (2009). Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the Symposium on Learning and Data Science, Paris*.

[Bravo et al., 2018] Bravo, M., Leslie, D., and Mertikopoulos, P. (2018). Bandit learning in concave N-person games. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5666–5676, Red Hook, NY, USA.

[Brown et al., 2020] Brown, G., Hod, S., and Kalemaj, I. (2020). Performative prediction in a stateful world. *arXiv preprint arXiv:2011.03885*.

[Candela et al., 2009] Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. The MIT Press.

[Chen et al., 2017] Chen, P., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. (2017). ZOO: Zeroth-Order Optimization-based Black-box Attacks to Deep Neural Networks Without Training Substitute Models. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*.

[Dong et al., 2018] Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. (2018). Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, page 55–70, New York, NY, USA. Association for Computing Machinery.

[Drusvyatskiy and Xiao, 2020] Drusvyatskiy, D. and Xiao, L. (2020). Stochastic optimization with decision-dependent distributions. *arXiv*.

[Flaxman et al., 2005] Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA. Society for Industrial and Applied Mathematics.

[Gao et al., 2018] Gao, X., Jiang, B., and Zhang, S. (2018). On the information-adaptive variants of the ADMM: An iteration complexity perspective. *Journal of Scientific Computing*, 76:327–363.

[Gidel et al., 2017] Gidel, G., Jebara, T., and Lacoste-Julien, S. (2017). Frank-wolfe algorithms for saddle point problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 362–371, Fort Lauderdale, FL, USA. PMLR.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27.

[Hamilton, 2019] Hamilton, I. (June 14, 2019). Uber drivers are reportedly colluding to trigger surge prices because they say the company is not paying them enough. *Business Insider*.

[HaoChen and Sra, 2019] HaoChen, J. Z. and Sra, S. (2019). Random shuffling beats SGD after finite epochs. In *ICML*.

[Hardt et al., 2016] Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 111–122, New York, NY, USA. Association for Computing Machinery.

[Hast et al., 2013] Hast, M., Åström, K. J., Bernhardsson, B., and Boyd, S. (2013). Pid design by convex-concave optimization. In *2013 European Control Conference (ECC)*, pages 4460–4465. IEEE.

[Hellemo et al., 2018] Hellemo, L., Barton, P. I., and Tomasgard, A. (2018). Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395.

[Hu et al., 2018] Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037. PMLR.

[Ilyas et al., 2018] Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. (2018). Black-box adversarial attacks with limited queries and information. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2137–2146. PMLR.

[Jain et al., 2019] Jain, P., Nagaraj, D. M., and Netrapalli, P. (2019). SGD without replacement: sharper rates for general smooth convex functions. In *ICML*.

[Jin et al., 2020] Jin, C., Netrapalli, P., and Jordan, M. I. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *ICML*.

[Lee, 2013] Lee, J. M. (2013). *Introduction to Smooth Manifolds*. Springer Science+Business Media New York.

[Liu et al., 2020] Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O'Reilly, U.-M. (2020). Min-max optimization without gradients: Convergence and applications to adversarial ML. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6282–6293. PMLR.

[Madry et al., 2017] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *ArXiv*.

[McCullagh and Nelder, 2019] McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*. Routledge.

[Miller et al., 2021] Miller, J., Perdomo, J., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. *arXiv preprint arXiv:2102.08570*.

[Mokhtari et al., 2020a] Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020a). A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *AISTATS*.

[Mokhtari et al., 2020b] Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020b). Convergence Rate of O(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM J. Optim.*, 30:3230–3251.

[Namkoong and Duchi, 2016] Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29.

[Nesterov, 2014] Nesterov, Y. (2014). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition.

[Nesterov and Spokoiny, 2017] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566.

[Nouiehed et al., 2019] Nouiehed, M., Sanjabi, M., Huang, T., Lee, J., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*.

[Perdomo et al., 2020] Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR.

[Rahimian and Mehrotra, 2019] Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. In *SIAM*.

[Recht and Ré, 2011] Recht, B. and Ré, C. (2011). Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5.

[Recht and Wright, 2021] Recht, B. and Wright, S. (2021). *Optimization for Data Analysis*. Cambridge University Press, 1 edition.

[Rudin, 1976] Rudin, W. (1976). *Principles Of Mathematical Analysis*. McGraw-Hill, Inc.

[Sadiev et al., 2021] Sadiev, A., Beznosikov, A., Dvurechensky, P., and Gasnikov, A. (2021). Zeroth-Order Algorithms for Smooth Saddle-Point Problems. *ArXiv*.

[Safran and Shamir, 2019] Safran, I. and Shamir, O. (2019). How good is SGD with random shuffling? *ArXiv*, abs/1908.00045.

[Sessa et al., 2020] Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. (2020). Learning to play sequential games versus unknown opponents. *ArXiv*, abs/2007.05271.

[Shafieezadeh-Abadeh et al., 2015] Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *NeurIPS*.

[Shamir, 2016] Shamir, O. (2016). Without-Replacement Sampling for Stochastic Gradient Methods. In *NIPS*.

[Spall, 1997] Spall, J. (1997). A one-measurement form of simultaneous perturbation stochastic approximation. *Autom.*, 33:109–112.

[Steinhardt et al., 2017] Steinhardt, J., Koh, P. W., and Liang, P. (2017). Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3520–3532, Red Hook, NY, USA.

[Tu et al., 2019] Tu, C.-C., Ting, P.-S., Chen, P., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. (2019). AutoZOOM: autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *AAAI*.

[Wang et al., 2020] Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. (2020). Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *ArXiv*, abs/2001.07819.

[Yang et al., 2020] Yang, J., Kiyavash, N., and He, N. (2020). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 1153–1165.

[Youn, 2019] Youn, S. (May 18, 2019). Uber, Lyft drivers coordinate to manipulate surge pricing at Virginia airport over pay concerns: Report. *ABC News*.

[Yu et al., 2021] Yu, Y., Lin, T., Mazumdar, E. V., and Jordan, M. I. (2021). Fast distributionally robust learning with variance reduced min-max optimization. *ArXiv*, abs/2104.13326.

Supplementary Material: Zeroth-Order Methods for Convex-Concave Minmax Problems: Applications to Decision-Dependent Risk Minimization

A RESULTS FOR THE PROOF OF THEOREM 4.1

A.1 Lemmas for Theorem 4.1

First, we list some fundamental facts regarding projections onto convex, compact subsets of an Euclidean space. Below, for any fixed convex, compact subset $\Omega \subset \mathbb{R}^d$, we denote the projection operator onto Ω by $\text{Proj}_\Omega(x) := \text{argmin}_{z\in\Omega} \|x - z\|_2$ for each $x \in \mathbb{R}^d$. Note that $\text{Proj}_\Omega(x)$ is well-defined (i.e., exists and is unique) for each $x \in \mathbb{R}^d$, if $\Omega \subset \mathbb{R}^d$ were convex and compact.

We begin by summarizing some fundamental properties of the projection operator $\text{Proj}_\Omega(\cdot)$.

Proposition A.1. *Let $\Omega \subset \mathbb{R}^d$ be compact and convex, and fix $x, y \in \mathbb{R}^d$ arbitrarily. Then:*

$$\begin{aligned}&\big\| \text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\big\|_2^2 \le \big(\text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\big)^\top (x - y),\\ &\| \text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\|_2 \le \|x - y\|_2.\end{aligned}$$

Proof. From [Nesterov, 2014], Lemma 3.1.4 (see also [Recht and Wright, 2021], Lemma 7.4), we have:

$$\begin{aligned}&\big(\text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\big)^\top \big(x - \text{Proj}_\Omega(x)\big) \ge 0,\\ &\big(\text{Proj}_\Omega(y) - \text{Proj}_\Omega(x)\big)^\top \big(y - \text{Proj}_\Omega(y)\big) \ge 0.\end{aligned}$$

Adding the two expressions and rearranging terms, we obtain:

$$\begin{aligned}&\big(\text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\big)^\top \big((x - y) - (\text{Proj}_\Omega(x) - \text{Proj}_\Omega(y))\big) \ge 0,\\ \Rightarrow &\| \text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\|_2^2 \le \big(\text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\big)^\top (x - y),\end{aligned}$$

as given in the first claim. The Cauchy Schwarz inequality then implies:

$$\begin{aligned}\| \text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\|_2^2 &\le \big(\text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\big)^\top (x - y)\\ &\le \| \text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\|_2 \cdot \|x - y\|_2.\end{aligned}$$

If $\text{Proj}_\Omega(x) = \text{Proj}_\Omega(y)$, then the second claim becomes $0 \le \|x - y\|_2$, which is clearly true. Otherwise, dividing both sides above by $\| \text{Proj}_\Omega(x) - \text{Proj}_\Omega(y)\|_2$ gives the second claim. □

Lemma A.2. *Let $\Omega \subset \mathbb{R}^d$ be a compact, convex subset of $\mathbb{R}^d$, and consider the update $z_{k+1} = \text{Proj}_\Omega(z_k - \eta F(z_{k+1}) + \gamma_k)$, where $z_k, z_{k+1}, \gamma_k \in \mathbb{R}^d$. Then, for each $z \in \Omega$:*

$$\begin{aligned}&\langle F(z_{k+1}), z_{k+1} - z\rangle\\ \le &\frac{1}{2\eta}\|z_k - z\|^2 - \frac{1}{2\eta}\|z_{k+1} - z\|^2 - \frac{1}{2\eta}\|z_{k+1} - z_k\|^2 + \frac{1}{\eta}\langle \gamma_k, z_{k+1} - z\rangle .\end{aligned}$$

Proof. Note that:

$$\begin{aligned}\|z_{k+1} - z\|^2 &= \|z_{k+1} - z_k + z_k - z\|^2\\ &= \|z_{k+1} - z_k\|^2 + \|z_k - z\|^2 + 2\langle z_{k+1} - z_k, z_k - z\rangle\\ &= \|z_{k+1} - z_k\|^2 + \|z_k - z\|^2 + 2\langle z_{k+1} - z_k, z_k - z_{k+1} + z_{k+1} - z\rangle\\ &= \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 + 2\langle z_{k+1} - z_k, z_{k+1} - z\rangle\end{aligned}$$

By definition of z_{k+1}, and optimality conditions for the projection operator:

$$\begin{aligned} &\langle z_{k+1} - z, z_{k+1} - z_k + \eta F(z_{k+1}) - \gamma_k \rangle \leq 0, \\ \Rightarrow &\langle z_{k+1} - z_k, z_{k+1} - z \rangle \leq \langle \gamma_k, z_{k+1} - z \rangle - \eta \cdot \langle F(z_{k+1}), z_{k+1} - z \rangle . \end{aligned}$$

Substituting back, we obtain:

$$\begin{aligned} \|z_{k+1} - z\|^2 &= \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 + 2 \langle z_{k+1} - z_k, z_{k+1} - z \rangle \\ &\leq \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 + 2 \langle \gamma_k, z_{k+1} - z \rangle - 2\eta \cdot \langle F(z_{k+1}), z_{k+1} - z \rangle . \end{aligned}$$

Rearranging and dividing by η gives the claim in the lemma. □

Next, we state the properties of the mean and variance of the zeroth-order gradient estimator defined in Section 4.1 ([Bravo et al., 2018], Lemma C.1). Below, we define the R-smoothed loss function $L^R : \mathbb{R}^d \to \mathbb{R}$ by $L^R(u) := \mathbb{E}_{\overline{v} \sim \mathsf{Unif}(B^d)}[L(u + R\overline{v})]$, where $\mathcal{S}^{d-1}$ denotes the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$, B^d denotes the d-dimensional unit open ball in $\mathbb{R}^d$, and $\mathsf{Unif}(\cdot)$ denotes the continuous uniform distribution over a set. Similarly, we define $L_i^R : \mathbb{R}^d \to \mathbb{R}$ by $L_i^R(u) := \mathbb{E}_{\overline{v} \sim \mathsf{Unif}(B^d)}[L_i(u + R\overline{v})]$, for each $i \in [n] := \{1, \cdots, n\}$. We further define $R \cdot \mathcal{S}^{d-1} := \{Rv : v \in \mathcal{S}^{d-1}\}$ and $R \cdot B^d := \{R\overline{v} : \overline{v} \in B^d\}$. Finally, we use $\text{vol}_d(\cdot)$ to denote the volume of a set in d dimensions.

Proposition A.3. *Let $\hat{F}(u; R, v) = \frac{d}{R} \cdot L(u + Rv)v$ and $F(u) = \nabla L(u)$. Then the following holds:*

$$\mathbb{E}_{v \sim \textit{Unif}(\mathcal{S}^{d-1})}\left[\hat{F}(u; R, v)\right] = \nabla L^R(u), \tag{11}$$

$$\|\nabla L^R(u) - F(u)\|_2 \leq \ell R, \tag{12}$$

$$\|\hat{F}(u; R, v)\|_2 \leq dG + \frac{dM_L}{R}, \tag{13}$$

$$\|\hat{F}(u; R, v) - F(u)\| \leq \min\left\{(d+1)G + \frac{dM_L}{R}, \ell R + 2dG + \frac{2dM_L}{R}\right\}. \tag{14}$$

Proof. First, to establish (11), observe that since $L^R(u) = \mathbb{E}_{v \sim \mathsf{Unif}(B^d)}[L(u + Rv)]$ and $\hat{F}(u; R, v) = \frac{d}{R} \cdot L(u + Rv)v$ for each $u \in \mathbb{R}^d$, $R > 0$, and $v \in \mathcal{S}^{d-1}$:

$$\begin{aligned} \nabla L^R(u) &= \nabla \mathbb{E}_{\overline{v} \sim \mathsf{Unif}(B^d)}\left[L(u + R\overline{v})\right] \\ &= \nabla \mathbb{E}_{\overline{v} \sim \mathsf{Unif}(R \cdot B^d)}\left[L(u + \overline{v})\right] \\ &= \frac{1}{\text{vol}_d(R \cdot B^d)} \cdot \nabla \left(\int_{R \cdot B^d} L(u + \overline{v}) \, d\overline{v} \right) \\ &= \frac{1}{\text{vol}_d(R \cdot B^d)} \cdot \int_{R \cdot \mathcal{S}^{d-1}} L(u + v) \cdot \frac{v}{\|v\|_2} \, dv, \qquad (15) \\ \mathbb{E}_{v \sim \mathsf{Unif}(\mathcal{S}^{d-1})}\left[\hat{F}(u; R, v)\right] &= \frac{d}{R} \cdot \mathbb{E}_{v \sim \mathsf{Unif}(\mathcal{S}^{d-1})}\left[L(u + Rv)v\right] \\ &= \frac{d}{R} \cdot \mathbb{E}_{v \sim \mathsf{Unif}(R \cdot \mathcal{S}^{d-1})}\left[L(u + v) \cdot \frac{v}{\|v\|_2}\right] \\ &= \frac{d}{R} \cdot \frac{1}{\text{vol}_{d-1}(R \cdot \mathcal{S}^{d-1})} \cdot \int_{R \cdot \mathcal{S}^{d-1}} L(u + v) \cdot \frac{v}{\|v\|_2} \, dv, \end{aligned}$$

where (15) follows because Stokes' Theorem (see, e.g., Lee, Theorem 16.11 [Lee, 2013]) implies that:

$$\nabla \int_{R \cdot B^d} L(u + \overline{v}) \, d\overline{v} = \int_{R \cdot \mathcal{S}^{d-1}} L(u + v) \cdot \frac{v}{\|v\|_2} \, dv.$$

The equality (11) now follows by observing that the surface-area-to-volume ratio of $R \cdot B^d$ is d/R.

Next, to establish (12), we note that:

$$\begin{align}
\|\nabla L^R(u) - F(u)\|_2 &= \left\|\nabla \mathbb{E}_{\overline{v}\sim \mathsf{Unif}(B^d)}\left[L^R(u) - L(u)\right]\right\|_2 \notag \\
&= \frac{1}{\mathrm{vol}_d(B^d)} \cdot \left\|\nabla\left(\int_{B^d}\left[L(u+R\overline{v}) - L(u)\right]d\overline{v}\right)\right\|_2 \notag \\
&\leq \frac{1}{\mathrm{vol}_d(B^d)} \cdot \left\|\int_{B^d}\left[F(u+R\overline{v}) - F(u)\right]d\overline{v}\right\|_2 \tag{16} \\
&\leq \frac{1}{\mathrm{vol}_d(B^d)} \cdot \int_{B^d}\left\|F(u+R\overline{v}) - F(u)\right\|_2 d\overline{v} \notag \\
&\leq \frac{1}{\mathrm{vol}_d(B^d)} \cdot \int_{B^d} \ell R \cdot \|\overline{v}\|_2\, d\overline{v} \notag \\
&\leq \ell R, \notag
\end{align}$$

where (16) follows by differentiating under the integral sign (see, e.g., Rudin, Theorem 9.42 [Rudin, 1976]), and the remaining inequalities follow from the fact that F is ℓ-Lipschitz.

Next, we establish (13) by using the triangle inequality and the M_L-boundedness of $L(\cdot)$ on $\mathcal{X} \times \mathcal{Y}$, and the G-Lipschitzness of $L(\cdot)$:

$$\begin{aligned}
|\hat{F}(u; R, v)| &= \frac{d}{R}|L(u+Rv)| \cdot \|v\|_2 \\
&\leq \frac{d}{R} \cdot \big(|L(u)| + |L(u+Rv) - L(u)|\big) \cdot 1 \\
&\leq \frac{d}{R} \cdot (M_L + RG).
\end{aligned}$$

We can then use (13) to establish (14) by observing that:

$$|\hat{F}(u; R, v) - F(u)| \leq |\hat{F}(u; R, v)| + |F(u)| \leq (d+1)G + \frac{dM_L}{R}.$$

and, from (13):

$$\begin{aligned}
&|\hat{F}(u; R, v) - F(u)| \\
\leq& \left|\hat{F}(u; R, v) - \mathbb{E}_v[\hat{F}(u; R, v)|u]\right| + \left|\mathbb{E}_v[\hat{F}(u; R, v)|u] - F(u)\right| \\
\leq& \left|\hat{F}(u; R, v) - \mathbb{E}_v[\hat{F}(u; R, v)|u]\right| + \left|\nabla L^R(u) - F(u)\right| \\
\leq& 2\left(dG + \frac{dM_L}{R}\right) + \ell R
\end{aligned}$$

This concludes the proof. □

Below, we present technical lemmas that allow us to analyze the convergence rate of the correlated iterates $\{u_i^t\}$ in our random reshuffling-based OGDA Algorithm (Alg. 1).

Let $\sigma^0, \cdots, \sigma^{t-1}$ denote the permutations drawn from epoch 0 to epoch $t-1$, and let $\{u_i^t(\sigma^t)\}_{1\leq i\leq n}$ and $\{u_i^t(\tilde{\sigma}^t)\}_{1\leq i\leq n}$ denote the iterates obtained at epoch t, when the permutations σ^t and $\tilde{\sigma}^t$ are used for the epoch t, respectively. Moreover, let $\mathcal{D}_{i,t}$ denote the distribution of $\{u_i^t(\sigma^t)\}_{1\leq i\leq n}$ under σ^t, and for $1 \leq r \leq n$ let $\mathcal{D}_{i,t}^{(r)}$ denote the distribution of $\{u_i^t(\sigma^t)\}_{1\leq i\leq n}$ with σ^t conditioned on the event $\{\sigma_{i-1}^t = r\}$.

We use the *p-Wasserstein distance between probability distributions on* $\mathbb{R}^d$, defined below, to characterize the distance between $\mathcal{D}_{i,t}$ and $\mathcal{D}_{i,t}^{(r)}$. This is used in the coupling-based techniques employed to establish non-asymptotic convergence results for our random reshuffling algorithm. Note the difference between the p-Wasserstein distance for probability distributions on $\mathbb{R}^d$, and the *Wasserstein distance on* $\mathcal{Z} := \mathbb{R}^d \times \{+1, -1\}$ *associated with a metric* $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$, defined in Appendix B.2 (Definition B.1).

Definition A.1 (**p-Wasserstein distance between distributions on $\mathbb{R}^d$**)**.** *Let μ, ν be probability distributions over $\mathbb{R}^d$ with finite p-th moments, for some $p \geq 1$, and let $\Pi(\mu, \nu)$ denote the set of all couplings (joint distributions) between μ and ν. The p-Wasserstein distance between μ and ν, denoted $\mathcal{W}_p(\mu, \nu)$, is defined by:*

$$\mathcal{W}_p(\mu, \nu) = \inf_{(X,X')\sim\pi\in\Pi(\mu,\nu)} \Big(\mathbb{E}_\pi\big[\|X - X'\|^p\big]\Big)^{1/p}.$$

The following proposition characterizes the 1-Wasserstein distance as a measure of the gap between Lipschitz functions of random variables.

Proposition A.4 (**Kantorovich Duality**)**.** *If μ, ν are probability distributions over $\mathbb{R}^d$ with finite second moments, then:*

$$\mathcal{W}_1(\mu, \nu) = \sup_{g\in\text{Lip}(1)} \mathbb{E}_{X\sim\mu}[g(X)] - \mathbb{E}_{Y\sim\nu}[g(Y)],$$

where $\text{Lip}(1) := \{g : \mathbb{R}^d \to \mathbb{R} : g$ *is 1-Lipschitz*$\}$.

Using [Yu et al., 2021, Lemma C.2], we now bound the difference between the unbiased gap $\mathbb{E}[\Delta(u_i^t)]$ and the biased gap $\mathbb{E}[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)]$ using the Wasserstein metric.

Lemma A.5. *Let $u^\star := (x^\star, y^\star) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} = \mathbb{R}^d$ denote a saddle point of the min-max optimization problem (2). Then, for each $t \in [T]$ and $i \in [n]$, the iterates $\{u_i^t\} = \{(x_i^t, y_i^t)\}$ of the OGDA-RR algorithm satisfy:*

$$\Big|\mathbb{E}[\Delta(u_{i+1}^t)] - \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\big]\Big| \leq \frac{G}{n}\sum_{r=1}^n \mathcal{W}_2\big(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^r\big)$$

Proof. Since σ^t and $\tilde{\sigma}^t$ are independently generated permutations of $[n]$, the iterates $\{u_i^t\}_{1\leq i\leq n} = \{u_i^t(\sigma^t)\}_{1\leq i\leq n}$ and $\{u_i^t(\tilde{\sigma}^t)\}_{1\leq i\leq n}$ are i.i.d. Thus, we have:

$$\mathbb{E}[\Delta(u_{i+1}^t)] = \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t(\tilde{\sigma}^t), y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t(\tilde{\sigma}^t))\big],$$

and thus:

$$\begin{aligned}
&\Big|\mathbb{E}[\Delta(u_{i+1}^t)] - \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\big]\Big| \\
=&\Big|\mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t(\tilde{\sigma}^t), y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t(\tilde{\sigma}^t))\big] - \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\big]\Big| \\
=&\Bigg|\frac{1}{n}\sum_{r=1}^n \mathbb{E}\big[L_r(x_{i+1}^t(\tilde{\sigma}^t), y^\star) - L_r(x^\star, y_{i+1}^t(\tilde{\sigma}^t))\big] \qquad\qquad (17)\\
&\qquad - \frac{1}{n}\sum_{r=1}^n \mathbb{E}\big[L_r(x_{i+1}^t, y^\star) - L_r(x^\star, y_{i+1}^t)\big|\sigma_i^t = r\big]\Bigg| \\
\leq&\frac{1}{n}\sum_{r=1}^n \Big|\mathbb{E}\big[L_r(x_{i+1}^t(\tilde{\sigma}^t), y^\star) - L_r(x^\star, y_{i+1}^t(\tilde{\sigma}^t))\big] - \mathbb{E}\big[L_r(x_{i+1}^t, y^\star) - L_r(x^\star, y_{i+1}^t)\big|\sigma_i^t = r\big]\Big| \\
\leq&\frac{1}{n}\sum_{r=1}^n \sup_{g\in\text{Lip}(G)} \Big(\mathbb{E}\big[g(x_{i+1}^t(\tilde{\sigma}^t), y_{i+1}^t(\tilde{\sigma}^t))\big] - \mathbb{E}\big[g(x_{i+1}^t, y_{i+1}^t)|\sigma_i^t = r\big]\Big) \qquad\qquad (18)\\
\leq&\frac{1}{n}\sum_{r=1}^n G\cdot\mathcal{W}_1(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^{(r)}) \qquad\qquad (19)\\
\leq&\frac{1}{n}\sum_{r=1}^n G\cdot\mathcal{W}_2(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^{(r)}), \qquad\qquad (20)
\end{aligned}$$

where (17) follows by properties of the conditional expectation on $\{\sigma_i^t = r\}$ and the fact that σ^t and $\tilde{\sigma}^t$ are independent, (18) follows from the fact that L is Lipschitz, (19) follows from Proposition A.4, and (20) follows from the fact that $\mathcal{W}_1(\mu, \nu) \leq \mathcal{W}_2(\mu, \nu)$ for any two probability distributions μ, ν. □

The next lemma bounds the difference in the iterates $\{u_i^t(\sigma^t)\}$ and $\{u_i^t(\tilde{\sigma}^t)\}$ (assuming, as before, that $\sigma^0, \cdots, \sigma^{t-1}$ were fixed and identical for both sequences.)

Lemma A.6. *Denote, with a slight abuse of notation, $u_i^t := u_i^t(\sigma^t)$ and $\tilde{u}_i^t := u_i^t(\tilde{\sigma}^t)$. Then:*

$$\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \le \left(6nd + 14n + 2 \cdot \sum_{i=1}^n \mathbf{1}\{\sigma_i^t \neq \tilde{\sigma}_i^t\}\right) G \cdot \eta^t + 6ndM_L \cdot \frac{\eta^t}{R^t}.$$

Proof. Our proof strategy is to bound the differences between zeroth-order and first-order OGDA updates, and between the OGDA and proximal point updates. To this end, we define:

$$\begin{aligned}
u_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t)\right),\\
\tilde{u}_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(\tilde{u}_i^t - \eta^t \hat{F}_{\tilde{\sigma}_i^t}(\tilde{u}_i^t; R^t, v_i^t) - \eta^t \hat{F}_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_i^t; R^t, v_i^t) + \eta^t \hat{F}_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_{i-1}^t; R^t, v_{i-1}^t)\right),\\
v_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t F_{\sigma_i^t}(u_i^t) - \eta^t F_{\sigma_{i-1}^t}(u_i^t) + \eta^t F_{\sigma_{i-1}^t}(u_{i-1}^t)\right),\\
\tilde{v}_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(\tilde{u}_i^t - \eta^t F_{\tilde{\sigma}_i^t}(\tilde{u}_i^t) - \eta^t F_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_i^t) + \eta^t F_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_{i-1}^t)\right),\\
w_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t F_{\sigma_i^t}(w_{i+1}^t)\right),\\
\tilde{w}_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(\tilde{u}_i^t - \eta^t F_{\tilde{\sigma}_i^t}(\tilde{w}_{i+1}^t)\right).
\end{aligned}$$

By the triangle inequality:

$$\begin{aligned}
\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \le \|u_{i+1}^t - v_{i+1}^t\|_2 + \|v_{i+1}^t - w_{i+1}^t\|_2 + \|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 \\
+ \|\tilde{w}_{i+1}^t - \tilde{v}_{i+1}^t\|_2 + \|\tilde{v}_{i+1}^t - \tilde{u}_{i+1}^t\|_2.
\end{aligned} \tag{21}$$

Observe that bounding the fourth term is equivalent to bounding the second term, and bounding the fifth term is equivalent to bounding the first term.

To bound the first term on the right hand side, we use Proposition A.3 to conclude that:

$$\begin{aligned}
\|u_{i+1}^t - v_{i+1}^t\|_2 &\le \eta^t \cdot \|\hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) - F_{\sigma_i^t}(u_i^t)\| + \eta^t \cdot \|\hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) - F_{\sigma_{i-1}^t}(u_i^t)\| \\
&\quad + \eta^t \cdot \|\hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t)\| \\
&\le 3(d+1)G\eta^t + 3dM_L \cdot \frac{\eta^t}{R^t}
\end{aligned} \tag{22}$$

For the second term, we use the G-Lipschitzness of L_r, for each $r \in [n]$ to conclude that:

$$\begin{aligned}
\|v_{i+1}^t - w_{i+1}^t\|_2 &\le \eta^t \cdot |F_{\sigma_i^t}(u_i^t)| + \eta^t \cdot |F_{\sigma_{i-1}^t}(u_i^t)| + \eta^t \cdot |F_{\sigma_{i-1}^t}(u_{i-1}^t)| + \eta^t \cdot |F_{\sigma_i^t}(w_{i+1}^t)| \\
&\le 4G \cdot \eta^t.
\end{aligned} \tag{23}$$

For the third term, we observe that if $\sigma_i^t \neq \tilde{\sigma}_i^t$, then:

$$\begin{aligned}
\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 &\le \|u_i^t - \tilde{u}_i^t\|_2 + \eta^t \cdot \|F_{\sigma_i^t}(w_{i+1}^t) - F_{\tilde{\sigma}_i^t}(\tilde{w}_{i+1}^t)\|_2 \\
&\le \|u_i^t - \tilde{u}_i^t\|_2 + 2G \cdot \eta^t.
\end{aligned} \tag{24}$$

On the other hand, if $\sigma_i^t = \tilde{\sigma}_i^t$, then:

$$\begin{aligned}
w_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t F_{\sigma_i^t}(w_{i+1}^t)\right),\\
\tilde{w}_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(\tilde{u}_i^t - \eta^t F_{\sigma_i^t}(\tilde{w}_{i+1}^t)\right),
\end{aligned}$$

so we have:

$$\begin{aligned}
&\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2^2 \\
\le\, &(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top \left((u_i^t - \eta \cdot F_{\sigma_i^t}(w_{i+1}^t)) - (\tilde{u}_i^t - \eta \cdot F_{\sigma_i^t}(\tilde{w}_{i+1}^t))\right) && (25)\\
=\, &(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top (u_i^t - \tilde{u}_i^t) - \eta (w_{i+1}^t - \tilde{w}_{i+1}^t)^\top \left(F_{\sigma_i^t}(w_{i+1}^t)) - F_{\sigma_i^t}(\tilde{w}_{i+1}^t))\right) \\
\le\, &(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top (u_i^t - \tilde{u}_i^t) && (26)\\
\le\, &\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 \cdot \|u_i^t - \tilde{u}_i^t\|_2, && (27)
\end{aligned}$$

so $\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 \leq \|u_i^t - \tilde{u}_i^t\|_2$. Here, (25) follows from the definitions of w_{i+1}^t and $\tilde{w}_{i+1}^t$, as well as Proposition A.1, while (26) holds because the monotonicity of F_i, for each $i \in [n]$, implies that $(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top \big(F_{\sigma_i^t}(w_{i+1}^t) - F_{\sigma_i^t}(\tilde{w}_{i+1}^t)\big) \geq 0$. Putting together (22), (23), (24), (27), we have:

$$\begin{aligned}\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \leq \|u_i^t - \tilde{u}_i^t\|_2 + (6d+14)G \cdot \eta^t + 6dM_L \cdot \frac{\eta^t}{R^t} \\ + 2G \cdot \mathbf{1}\{\sigma_i^t \neq \tilde{\sigma}_i^t\} \cdot \eta^t,\end{aligned}$$

where the indicator $\mathbf{1}(A)$ returns 1 if the given event A occurs, and 0 otherwise.

Since $u_0^t = \tilde{u}_0^t$, we can iteratively apply the above inequality to obtain that, for any and epoch t and $i \in [n]$:

$$\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \leq (6d+14)nG \cdot \eta^t + 6ndM_L \cdot \frac{\eta^t}{R^t} + 2\eta_t G \cdot \sum_{i=1}^n \mathbf{1}\{\sigma_i^t \neq \tilde{\sigma}_i^t\},$$

□

Remark. In the theorems and lemmas below, we will be concerned with the case where σ^t and $\tilde{\sigma}^t$ have the following specific relationship. Let $\mathcal{R}_n$ denote the set of all random permutations over the set $[n]$. For each $l, m \in [n]$, let $S_{l,m} : \mathcal{R}_n \to \mathcal{R}_n$ denote the map that swaps, for each input permutation σ, the l-th and m-th entries. For each $r, i \in [n]$, define the map $\omega_{r,i} : \mathcal{R}_n \to \mathcal{R}_n$ as follows:

$$\omega_{r,i}(\sigma) = \begin{cases} \sigma, & \text{if } \sigma_{i-1} = r, \\ S_{i-1,j}(\sigma), & \text{if } \sigma_j = r \text{ and } j \neq i-1. \end{cases}.$$

Intuitively, $\omega_{r,i}$ performs a single swap such that the $(i-1)$-th position of the permutation is r. Clearly, if σ^t is a random permutation (i.e., selected from a uniform distribution over $\mathcal{R}_n$), then $\omega_{r,i}(\sigma^s)$ has the same distribution as $\sigma^t|(\sigma_{i-1}^t = r)$. Based on this construction, we have $u_i(\sigma^t) \sim \mathcal{D}_{i,t}$ and $u_i(\omega_{r,i}(\sigma^t)) \sim \mathcal{D}_{i,t}^{(r)}$. This gives a coupling between $\mathcal{D}_{s,t}$ and $\mathcal{D}_{s,t}^{(r)}$. Since σ^t and $\tilde{\sigma}^t$ differ by at most two entries, by iteratively applying Lemma A.6, we have:

$$\begin{aligned}\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 &\leq n\left((6d+14)G \cdot \eta^t + 6dM_L \cdot \frac{\eta^t}{R^t}\right) + 4G \cdot \eta^t \\ &= (6nd + 14n + 4)G \cdot \eta^t + 6ndM_L \cdot \frac{\eta^t}{R^t},\end{aligned}$$

as claimed.

Lemma A.7. *If $\eta^t \leq 1/(2\ell)$ for each $t \in \{0, 1, \cdots, T-1\}$, the iterates $\{u_i^t\} = \{(x_i^t, y_i^t)\}$ of the OGDA-RR algorithm satisfy, for each $u \in \mathcal{X} \times \mathcal{Y}$:*

$$\begin{aligned}&2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u\Big\rangle\Big] \\ \leq &\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big] \\ &+ 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \\ &- 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big] \\ &+ 6C_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right),\end{aligned}$$

where $C_1 := d^2 \max\big\{6G\ell D, 18G^2 + 6M_L\ell D, 30M_LG, 12M_L^2\big\}$ is a constant independent of the sequences $\{\eta^t\}$ and $\{R^t\}$.

Proof. The iterates of the OGDA-RR algorithm are given by:

$$
\begin{aligned}
u_{i+1}^t &= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) \\
&\qquad\qquad - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t)\Big) \\
&= \mathrm{Proj}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t F_{\sigma_i^t}(u_{i+1}^t) + \eta^t\big(\gamma_i^t + E_{i,1}^t + E_{i,2}^t + E_{i,3}^t\big)\Big), \qquad (28)
\end{aligned}
$$

where we have defined:

$$
\begin{aligned}
\gamma_i^t &:= F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_i^t) + F_{\sigma_{i-1}^t}(u_{i-1}^t), \\
E_{i,1}^t &:= F_{\sigma_i^t}(u_i^t) - \hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t), \\
E_{i,2}^t &:= F_{\sigma_{i-1}^t}(u_i^t) - \hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t), \\
E_{i,3}^t &:= F_{\sigma_{i-1}^t}(u_{i-1}^t) - \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t).
\end{aligned}
$$

First, by applying Lemma A.2 we have:

$$
\begin{aligned}
&2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u\Big\rangle\Big] \qquad (29) \\
\leq &\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] \\
&\quad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle \gamma_i^t, u_{i+1}^t - u\Big\rangle\Big] + \sum_{k=1}^{3} 2\eta^t \cdot \mathbb{E}\Big[\Big\langle E_{i,k}^t, u_{i+1}^t - u\Big\rangle\Big].
\end{aligned}
$$

Below, we proceed to bound the inner product terms on the right-hand-side of (29). First, we bound $\big\langle \gamma_i^t, u_{i+1}^t - u\big\rangle$:

$$
\begin{aligned}
\Big\langle \gamma_i^t, u_{i+1}^t - u\Big\rangle &= \Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle \\
&\qquad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_{i+1}^t - u\Big\rangle \\
&= \Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle \\
&\qquad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle \\
&\qquad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_{i+1}^t - u_i^t\Big\rangle \\
&\leq \Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle \qquad (30) \\
&\qquad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle \\
&\qquad + \frac{1}{2}\ell \cdot \|u_i^t - u_{i-1}^t\|_2^2 + \frac{1}{2}\ell \cdot \|u_{i+1}^t - u_i^t\|_2^2.
\end{aligned}
$$

Note that the final inequality follows by applying Young's inequality, and noting that F is ℓ-Lipschitz. Next, we bound $\langle E_{i,1}^t, u_{i+1}^t - u\rangle$:

$$
\begin{aligned}
&\mathbb{E}\big[\langle E_{i,1}^t, u_{i+1}^t - u\rangle\big] \\
=&\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_{i+1}^t - u\Big\rangle\Big] \\
=&\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \nabla L_{\sigma_i^t}^{R^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \\
&\quad + \mathbb{E}\Big[\Big\langle \mathbb{E}\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_{i+1}^t - u\Big\rangle\Big] \\
=&\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \nabla L_{\sigma_i^t}^{R^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \qquad (31) \\
&\quad + \mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_i^t - u\Big\rangle\Big] \\
&\quad + \mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_{i+1}^t - u_i^t\Big\rangle\Big],
\end{aligned}
$$

where the first equality above follows by applying Proposition A.3, (11), and we have used the shorthand $\mathbb{E}_v := \mathbb{E}_{v\sim \mathsf{Unif}(\mathcal{S}^{d-1})}$. (Recall that $L^R(u) := \mathbb{E}_{v\sim \mathsf{Unif}(\mathcal{S}^{d-1})}\big[L(u+Rv)\big]$) Next, we upper bound each of the three quantities in (31). First, by Proposition A.3, (12), we have:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \nabla L_{\sigma_i^t}^{R^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \\
\leq &\mathbb{E}\Big[\|F_{\sigma_i^t}(u_i^t) - \nabla L_{\sigma_i^t}^{R^t}(u_i^t)\|_2 \cdot \|u_{i+1}^t - u\|_2\Big] \\
\leq &\ell D \cdot R^t,
\end{aligned} \tag{32}$$

with $C_1 > 0$ as given in Lemma A.7. Meanwhile, the law of iterated expectations can be used to bound the second quantity:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v)|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_i^t - u\Big\rangle\Big] \\
= &\mathbb{E}\Big[\mathbb{E}_v\Big[\Big\langle \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_i^t - u\Big\rangle\Big|u_i^t\Big]\Big] - \mathbb{E}\Big[\Big\langle \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_i^t - u\Big\rangle\Big] \\
= &0,
\end{aligned} \tag{33}$$

and we can upper-bound the third quantity as shown below. By using the compactness of $\mathcal{X} \times \mathcal{Y}$ and the continuity of L, we have:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v)|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_{i+1}^t - u_i^t\Big\rangle\Big] \\
\leq &\Big(\big\|\mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v)|u_i^t\big]\big\|_2 + \|\hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t)\|\Big) \cdot \|u_{i+1}^t - u_i^t\|_2 \\
\leq &2 \cdot \frac{d}{R^t} \cdot \sup_{\substack{u \in \mathcal{X}\times\mathcal{Y} \\ v\sim \mathsf{Unif}(\mathcal{S}^{d-1})}} |L(u_i^t + R^t v)| \cdot \|u_{i+1}^t - u_i^t\|_2, \\
\leq &2 \cdot \frac{d}{R^t} \cdot (M_L + R^t G) \cdot \|u_{i+1}^t - u_i^t\|_2,
\end{aligned} \tag{34}$$

and using (32) and the bound for each $\|\hat{F}_{\sigma_i^t}\|_2$ given in (34), we have:

$$\begin{aligned}
&\|u_{i+1}^t - u_i^t\|_2 \\
\leq &\eta^t \cdot \|\hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) + \hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) - \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t)\| \\
\leq &\eta^t \cdot \|F_{\sigma_i^t}(u_i^t) + F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t)\|_2 \\
&\quad + \eta^t d \cdot \|\hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) - F_{\sigma_i^t}(u_i^t)\|_2 \\
&\quad + \eta^t d \cdot \|\hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) - F_{\sigma_{i-1}^t}(u_i^t)\|_2 \\
&\quad + \eta^t d \cdot \|\hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t)\|_2 \\
\leq &3G\eta^t + 3\eta^t d \cdot \left(2(M_L + GR^t) \cdot \frac{1}{R^t} + \ell D \cdot R^t\right).
\end{aligned} \tag{35}$$

Substituting (35) back into (34), we have:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; R^t, v)|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, R^t, v_i^t), u_{i+1}^t - u_i^t\Big\rangle\Big] \\
\leq &d^2 \ell D 6\eta^t G \cdot R^t + 6d^2\eta^t(3G^2 + M_L \ell D) + 30 d^2 \eta^t M_L G \cdot \frac{1}{R^t} + 12 d^2 \eta^t M_L^2 \cdot \left(\frac{1}{R^t}\right)^2 \\
\leq &C_1 \cdot \left(\eta^t R^t + \eta^t + \frac{\eta^t}{R^t} + \frac{\eta^t}{(R^t)^2}\right),
\end{aligned} \tag{36}$$

where $C_1 := d^2 \cdot \max\big\{6G\ell D, 18G^2 + 6M_L \ell D, 30 M_L G, 12 M_L^2\big\}$ is a constant independent of the sequences $\{\eta^t\}$ and $\{R^t\}$. The quantities $\mathbb{E}\big[\langle E_{i,2}^t, u_{i+1}^t - u\rangle\big]$ and $\mathbb{E}\big[\langle E_{i,3}^t, u_{i+1}^t - u\rangle\big]$ can be similarly bounded. Substituting

(32), (33), (36) back into (31), and substituting (31) and (30) into (29), we find that:

$$\begin{aligned}
&2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u \Big\rangle\Big] \\
=&\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] \\
&\qquad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle \gamma_i^t, u_{i+1}^t - u \Big\rangle\Big] + 2\eta^t \cdot \sum_{k=1}^{3} \mathbb{E}\Big[\Big\langle E_{i,k}^t, u_{i+1}^t - u \Big\rangle\Big] \\
\leq&\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] \\
&\qquad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u \Big\rangle\Big] \\
&\qquad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u \Big\rangle\Big] \\
&\qquad + \eta^t \ell \cdot \mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big] + \eta^t \ell \cdot \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] \\
&\qquad + 6C_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right),
\end{aligned}$$

In particular, since by assumption $\eta^t \leq 1/(2\ell)$ for each $t \in \{0, 1, \cdots, T-1\}$, then:

$$\begin{aligned}
&2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u \Big\rangle\Big] \\
\leq&\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big] \\
&\qquad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u \Big\rangle\Big] \\
&\qquad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u \Big\rangle\Big] \\
&\qquad + 6C_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right),
\end{aligned}$$

□

Finally, to bound the step size terms above, we require the following lemma, which follows from standard calculus arguments.

Lemma A.8.

$$\begin{aligned}
\sum_{t=1}^{T} t^{-\beta} &\geq \frac{1}{1-\beta} T^{1-\beta}, \qquad && \forall\, \beta < 1, \\
\sum_{t=1}^{T} t^{-(1+\beta)} &\leq \frac{1}{\beta} + 1, \qquad && \forall\, \beta > 0.
\end{aligned}$$

A.2 Proof of Theorem 4.1

Proof. (**Proof of Theorem 4.1**) By applying Lemma A.7 (note that $\eta^t \leq \eta^0 \leq \frac{1}{2\ell}$, for each $t \in \{0, 1, \cdots, T-1\}$) and using convex-concave nature of L_r (refer Proposition 1 in [Mokhtari et al., 2020b]), for each $r \in \{1, \cdots n\}$, we have:

$$\begin{aligned}
&2\eta^t \cdot \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\big] \\
\leq& 2\eta^t \cdot \mathbb{E}\big[\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u^\star\rangle\big] \\
\leq& \mathbb{E}\big[\|u_i^t - u^\star\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u^\star\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big] \\
&\quad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u^\star\Big\rangle\Big] \\
&\quad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u^\star\Big\rangle\Big] \\
&\quad + 6C_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right).
\end{aligned} \tag{37}$$

Meanwhile, Lemma A.5, Proposition A.4 (Kantorovich Duality), and Lemma A.6 imply that:

$$\begin{aligned}
\Big|\mathbb{E}[\Delta(u_{i+1}^t)] - \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\big]\Big| &\leq \frac{G}{n}\sum_{r=1}^n \mathcal{W}_2\big(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^r\big) \\
&\leq \frac{G}{n}\sum_{r=1}^n \sqrt{\mathbb{E}\big[\big\|u_{i+1}^t(\sigma^t) - u_{i+1}^t(\tilde{\sigma}^t)\big\|_2^2\big]} \\
&\leq G \cdot \left((6nd + 14n + 4)G \cdot \eta^t + 6ndM_L \cdot \frac{\eta^t}{R^t}\right).
\end{aligned}$$

Substituting back into (37), we have:

$$\begin{aligned}
&2\eta^t \cdot \mathbb{E}\big[\Delta(u_i^t)\big] \\
\leq& 2\eta^t \cdot \mathbb{E}\big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\big] \\
&\quad + G \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{R^t}\right) \\
\leq& \mathbb{E}\big[\|u_i^t - u^\star\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u^\star\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big] \\
&\quad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \\
&\quad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big] \\
&\quad + 6C_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right) \\
&\quad + G \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{R^t}\right).
\end{aligned} \tag{38}$$

We can now sum the above telescoping terms across the t-th epoch, as shown below:

$$\begin{aligned}
&2 \cdot \sum_{i=1}^n \eta^t \cdot \mathbb{E}\big[\Delta(u_i^t)\big] \\
\leq& \mathbb{E}\big[\|u_1^t - u^\star\|_2^2\big] - \mathbb{E}\big[\|u_1^{t+1} - u^\star\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_1^t - u_0^t\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_1^{t+1} - u_0^{t+1}\|_2^2\big] \\
&\quad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_0^t}(u_1^t) - F_{\sigma_0^t}(u_0^t), u_1^t - u^\star\Big\rangle\Big] \\
&\quad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_0^{t+1}}(u_1^{t+1}) - F_{\sigma_0^{t+1}}(u_0^{t+1}), u_1^{t+1} - u^\star\Big\rangle\Big] \\
&\quad + 6nC_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right) \\
&\quad + nG \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{R^t}\right).
\end{aligned}$$

Meanwhile, we have for each $t = 0, 1, \cdots, T-1$, $i \in [n]$:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u^\star \Big\rangle\Big] \\
\leq &\mathbb{E}\big[\|F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t)\| \cdot \|u_{i+1}^t - u^\star\|\big] \\
= &\ell \cdot \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|\big] \cdot D \\
\leq &\ell D \cdot \mathbb{E}\Big[\| - \eta^t \hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t)\|\Big] \\
\leq &3\ell D \cdot \eta^t \cdot \left(dG + \frac{dM_L}{R^t}\right) \\
= &3\ell DdG \cdot \eta^t + 3\ell DdM_L \cdot \frac{\eta^t}{R^t},
\end{aligned}$$

where the final inequality follows from Proposition A.3, (13). We can upper bound $\mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big]$ in a similar fashion. Substituting back into (38), we have:

$$\begin{aligned}
&2 \cdot \sum_{i=1}^n \eta^t \cdot \mathbb{E}\big[\Delta(u_i^t)\big] \\
\leq &\mathbb{E}\big[\|u_1^t - u^\star\|_2^2\big] - \mathbb{E}\big[\|u_1^{t+1} - u^\star\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_1^t - u_0^t\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_1^{t+1} - u_0^{t+1}\|_2^2\big] \\
&\quad + 6nC_1 \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right) \\
&\quad + nG \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{R^t}\right) \\
&\quad + 6\ell DdG \cdot (\eta^t)^2 + 6\ell DdM_L \cdot \frac{(\eta^t)^2}{R^t} \\
\leq &\mathbb{E}\big[\|u_1^t - u^\star\|_2^2\big] - \mathbb{E}\big[\|u_1^{t+1} - u^\star\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_1^t - u_0^t\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_1^{t+1} - u_0^{t+1}\|_2^2\big] \qquad (39) \\
&\quad + 2C \cdot \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right),
\end{aligned}$$

where $C := \max\{3nC_1, (6nd + 14n + 4)nG, 6ndM_L, 3\ell DdG, 3\ell DdM_L\}$.

Finally, summing the above telescoping terms over $i \in [n]$ and $t \in \{0, 1, \cdots, T-1\}$, and removing non-positive terms, we obtain:

$$\begin{aligned}
&\frac{\sum_{t=0}^{T-1}\sum_{i=1}^n \eta^t \cdot \mathbb{E}\big[\Delta(u_i^t)\big]}{\sum_{t=0}^{T-1}\sum_{i=1}^n \eta^t} \\
\leq &\frac{1}{2 \cdot \sum_{t=0}^{T-1}\sum_{i=1}^n \eta^t}\left(\|u_0^0 - u^\star\|_2 - \mathbb{E}\big[\|u_n^{T-1} - u^\star\|_2\big] + \frac{1}{2}\|u_1^0 - u_0^0\|_2 - \frac{1}{2}\mathbb{E}\big[\|u_n^{T-1} - u_{n-1}^{T-1}\|_2\big]\right) \\
&\quad + C \cdot \frac{1}{\sum_{t=0}^{T-1}\sum_{i=1}^n \eta^t} \cdot \sum_{t=0}^{T-1}\left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right) \\
\leq &\frac{1}{\sum_{t=0}^{T-1}\eta^t} \cdot \frac{3D}{4n} + C \cdot \frac{1}{n\sum_{t=0}^{T-1}\eta^t} \cdot \sum_{t=0}^{T-1}\left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2}\right), \qquad (40)
\end{aligned}$$

By definition, $\eta^t = \eta^0 \cdot (t+1)^{-3/4-\chi}$ and $R^t = R^0 \cdot (t+1)^{-1/4}$, so by Lemma A.8, we have:

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta^t &= \eta^0 \cdot \sum_{t=1}^{T} t^{-3/4-\chi} \geq 4\eta^0 \cdot T^{1/4-\chi}, \\
\sum_{t=0}^{T-1} \eta^t R^t &= \eta^0 R^0 \cdot \sum_{t=1}^{T} t^{-(1+\chi)} \leq \eta^0 R^0 \cdot \left(1 + \frac{1}{\chi}\right), \\
\sum_{t=0}^{T-1} (\eta^t)^2 &= (\eta^0)^2 \cdot \sum_{t=1}^{T} t^{-3/2-2\chi} \leq (\eta^0)^2 \cdot \left(1 + \frac{1}{\frac{1}{2} + 2\chi}\right) \leq 3 \cdot (\eta^0)^2, \\
\sum_{t=0}^{T-1} (\eta^t)^2 R^t &= (\eta^0)^2 R^0 \cdot \sum_{t=1}^{T} t^{-7/4-2\chi} \leq (\eta^0)^2 R^0 \cdot \left(1 + \frac{1}{\frac{3}{4} + 2\chi}\right) \leq \frac{7}{4} \cdot (\eta^0)^2 \epsilon^0, \\
\sum_{t=0}^{T-1} \frac{(\eta^t)^2}{R^t} &= \frac{(\eta^0)^2}{R^0} \cdot \sum_{t=1}^{T} t^{-5/4-2\chi} \leq \frac{(\eta^0)^2}{R^0} \cdot \left(1 + \frac{1}{\frac{1}{4} + 2\chi}\right) \leq 5 \cdot \frac{(\eta^0)^2}{\epsilon^0}, \\
\sum_{t=0}^{T-1} \frac{(\eta^t)^2}{(R^t)^2} &= \frac{(\eta^0)^2}{(R^0)^2} \cdot \sum_{t=1}^{T} t^{-1-2\chi} \leq \frac{(\eta^0)^2}{(R^0)^2} \cdot \left(1 + \frac{1}{2\chi}\right).
\end{aligned}$$

Substituting back into (40) and using the convexity of the gap function $\Delta(\cdot)$, we have:

$$\begin{aligned}
&\mathbb{E}\big[\Delta(u^T)\big] \\
\leq\, &\frac{\sum_{t=0}^{T-1} \sum_{i=1}^{n} \eta^t \cdot \mathbb{E}\big[\Delta(u_i^t)\big]}{\sum_{t=0}^{T-1} \sum_{i=1}^{n} \eta^t} \\
\leq\, &\frac{1}{\sum_{t=0}^{T-1} \eta^t} \cdot \frac{3}{4n} D + C \cdot \frac{1}{\sum_{t=0}^{T-1} \eta^t} \cdot \sum_{t=0}^{T-1} \left(\eta^t R^t + (\eta^t)^2 R^t + (\eta^t)^2 + \frac{(\eta^t)^2}{R^t} + \frac{(\eta^t)^2}{(R^t)^2} \right) \\
\leq\, &\left(\frac{3}{16n} D + \frac{47}{4n} \cdot C \max\left\{ R^0, \eta^0, \eta^0 R^0, \frac{\eta^0}{R^0}, \frac{\eta^0}{(R^0)^2} \right\} \left(1 + \frac{1}{\chi}\right) \right) T^{-1/4+\chi} \\
\leq\, &R.
\end{aligned}$$

where the final inequality follows by definition of T. □

B WASSERSTEIN DISTRIBUTIONALLY ROBUST STRATEGIC CLASSIFICATION

B.1 Model of Adversary

In this subsection, we formally define our model for the adversary, and the uncertainty set of distributions for the resulting strategically and adversarially perturbed data. For better exposition, in this section we summarize the various distributions used in the main article in Table 1 below.

Table 1: Table of notations

Notation	Explanation
$\mathcal{D}$	Unknown underlying distribution
$\mathcal{D}(\theta)$	Unknown underlying distribution strategically perturbed by θ
$\tilde{\mathcal{D}}_n(\theta)$	Empirical distribution of strategically perturbed data
$\mathbb{P}$	An element of uncertainty set $\mathcal{P}(\theta)$
$\mathbb{P}^i_\theta$	Conditional distribution of adversarially generated data given i^{th} data point

The WDRSL problem formulation contains two main components—the *strategic component* that accounts for the distribution shift $\mathcal{D}(\theta)$ in response to the choice of classifier θ, and the *adversarial component*, which accounts for the uncertainty set $\mathcal{P}(\theta)$. As per the modeling assumptions put forth in Section 5.2, we have $(\tilde{x}_i, \tilde{y}_i) \sim \mathcal{D}$ and $(b_i(\theta, \tilde{x}_i, \tilde{y}_i), \tilde{y}_i) \sim \mathcal{D}(\theta)$ for all $i \in [n]$. For the sake of brevity, we shall use $b_i(\theta)$ in place of $b_i(\theta, \tilde{x}_i, \tilde{y}_i)$ for all $i \in [n]$.

As per the standard formulation of distributionally robust optimization, we restrict $\mathcal{P}(\theta)$ to be a Wasserstein neighborhood of $\tilde{\mathcal{D}}_n(\theta)$ (the empirical distribution of strategic responses $\{(b_i(\theta), \tilde{y}_i)\}_{i=1}^n$), i.e., we set $\mathcal{P}(\theta) \subset \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta))$ for some $\delta > 0$. However, to ensure that the min-max problem reformulated from the WDRSC problem is convex-concave, we further require the adversary to modify the *label* of an data point i in the empirical distribution only when the true label $\tilde{y}_i$ is $+1$, although they are still always allowed to modify the *feature* $b_i(\theta)$. As a consequence, this imposes some restrictions on the conditional distribution $\mathbb{P}^i_\theta$ of (dx, y), as generated by the adversary, given a data point i in the empirical distribution. In particular:

$$\mathbb{P}^i_\theta(dx, +1 | b_i(\theta), -1) = 0, \qquad \forall\, i \in [n].$$

By definition of conditional distributions, we obtain that any distribution $\mathbb{P}$ can be expressed as the average of the conditional distribution $\mathbb{P}^i_\theta$. That is,

$$\mathbb{P}(dx, y) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}^i_\theta(dx, y | b_i(\theta), \tilde{y}_i).$$

Below, we formally state the restriction described above.

Assumption B.1. *We assume that* $\mathbb{P} \in \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta))$ *and* $\mathbb{P}^i_\theta(dx, +1 | b_i(\theta), -1) = 0$ *for all* $i \in [n]$. *As a direct result, the uncertainty set* $\mathcal{P}(\theta)$ *is characterized as:*

$$\mathcal{P}(\theta) = \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta)) \cap \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{P}^i_\theta(dx, y | b_i(\theta), \tilde{y}_i) \middle| \mathbb{P}^i_\theta(dx, +1 | b_i(\theta), -1) = 0, \forall\, i \in [n] \right\}. \tag{41}$$

In the following subsection, we reformulate the WDRSC problem with a generalized linear model and with the uncertainty set defined in (41).

B.2 Proof of Theorem 5.3

The proof takes inspirations from [Shafieezadeh-Abadeh et al., 2015, Theorem 1]. First, we define the *Wasserstein distance between distributions on $\mathcal{Z}$ with cost function c*; note that this is different from the *p-Wasserstein distance between probability distributions on $\mathbb{R}^d$* defined in Appendix A.1.

Definition B.1. *(**Wasserstein distance between distributions on $\mathcal{Z}$ with cost Function** c) Let μ, ν be probability distributions over $\mathcal{Z} := \mathbb{R}^d \times \{+1, -1\}$ with finite second moments, and let $\Pi(\mu, \nu)$ denote the set of all couplings (joint distributions) between μ and ν. Given a metric $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ on $\mathcal{Z}$, we define:*

$$\mathcal{W}_c(\mu, \nu) = \inf_{(Z,Z')\sim\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi\big[c(Z, Z')\big].$$

In Theorem 7 and in our proof below, we use the cost function $c(z, z') := \|x - x'\|_2^2 + \kappa \cdot |y - y'|$, with a fixed constant $\kappa > 0$, for each $z := (x, y) \in \mathcal{Z}$ and $z' := (x', y') \in \mathcal{Z}$.

Proof. (**Proof of Theorem 5.3**) Fix a $\theta \in \Theta$. Note that $b_i(\theta, \tilde{x}_i, +1) = \tilde{x}_i$. For any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, let $\ell((x, y), \theta) := \phi(\langle x, \theta\rangle) - y\,\langle x, \theta\rangle$. We first analyze the inner supremum term, i.e.

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x, \theta\rangle) - y\langle x, \theta\rangle] \\
&= \sup_{\mathbb{P}\in\mathcal{P}(\theta)} \int_{\mathcal{Z}} \ell(z, \theta)\mathbb{P}(z)dz \\
&= \begin{cases} \displaystyle\sup_{\pi_\theta\in\Pi(\mathbb{P},\tilde{\mathcal{D}}_n(\theta))} & \int_{\mathcal{Z}} \ell(z, \theta)\pi_\theta(dz, \mathcal{Z}) \\ \text{s.t.} & \int_{\mathcal{Z}\times\mathcal{Z}} \|z - \tilde{z}\|\pi_\theta(dz, d\tilde{z}) \le \delta \end{cases}
\end{aligned}$$

Here, $\Pi(\mathbb{P}, \tilde{\mathcal{D}}_n(\theta))$ denotes the set of all joint distributions that couple $\mathbb{P} \in \mathcal{P}(\theta)$ and $\tilde{\mathcal{D}}_n(\theta)$. Since the marginal distribution $\tilde{\mathcal{D}}_n(\theta)$ of $\tilde{z}$ is discrete, such couplings π_θ are completely determined by the conditional distribution $\mathbb{P}^i_\theta$ of z given $\tilde{z}_i = (\tilde{x}_i(\theta), \tilde{y}_i)$ for each $i \in \{1, \dots, n\}$. That is:

$$\pi_\theta(dz, d\tilde{z}) = \frac{1}{n}\sum_{i\in[n]} \vartheta_{(b_i(\theta),\tilde{y}_i)}(d\tilde{z})\mathbb{P}^i_\theta(dz)$$

where for any $(x, y) \in \mathcal{Z}$, $\vartheta_{(x,y)}$ is a *Dirac delta* distribution with its support at point (x, y).

We introduce some notations. Let $\mathcal{I}_{+1} = \{i \in [n] : \tilde{y}_i = +1\}$ and $\mathcal{I}_{-1} = \{i \in [n] : \tilde{y}_i = -1\}$. Let's introduce two distributions μ^i_θ and ν^i_θ such that

$$\mathbb{P}^i_\theta = \begin{cases} \mu^i_\theta & \text{if } i \in \mathcal{I}_{+1} \\ \nu^i_\theta & \text{if } i \in \mathcal{I}_{-1} \end{cases}$$

Due to the constraint (41), we have $\nu^i_\theta(dx, +1) = 0$ at every x. This implies:

$$\pi_\theta(dz, d\tilde{z}) = \frac{1}{n}\left(\sum_{i\in\mathcal{I}_{+1}} \vartheta_{(b_i(\theta),1)}(d\tilde{z})\mu^i_\theta(dz) + \sum_{i\in\mathcal{I}_{-1}} \vartheta_{(b_i(\theta),-1)}(d\tilde{z})\nu^i_\theta(dz)\right)$$

With a slight abuse of notation, we denote $\mu^i_{\theta,+1}(dx) = \mu^i_\theta(dx, +1)$, $\mu^i_{\theta,-1}(dx) = \mu^i_\theta(dx, -1)$ and $\nu^i_\theta(dx) = \nu^i_\theta(dx, -1)$. The optimization problem of concern then simplifies to:

$$\begin{aligned}
\sup_{\mu^i_{\theta,\pm1},\nu^i_\theta}\quad & \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}} \int_{\mathbb{R}^d} \ell((x, +1), \theta)\mu^i_{\theta,+1}(dx) + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}} \int_{\mathbb{R}^d} \ell((x, -1), \theta)\mu^i_{\theta,-1}(dx) \\
&\quad + \frac{1}{n}\sum_{i\in\mathcal{I}_{-1}} \int_{\mathbb{R}^d} \ell((x, -1), \theta)\nu^i_\theta(dx) \\
\text{s.t.}\quad & \frac{1}{n}\sum_{i:\tilde{y}_i=+1} \int_{\mathbb{R}^d} \|(x, +1) - (b_i(\theta), \tilde{y}_i)\|\mu^i_{\theta,+1}(dx) \\
&\quad + \frac{1}{n}\sum_{i:\tilde{y}_i=+1} \int_{\mathbb{R}^d} \|(x, -1) - (b_i(\theta), \tilde{y}_i)\|\mu^i_{\theta,-1}(dx) \\
& \int_{\mathbb{R}^d} \mu^i_{\theta,+1}(dx) + \int_{\mathbb{R}^d} \mu^i_{\theta,-1}(dx) = 1, \quad \forall \quad i \in \mathcal{I}_{+1} \\
& \int_{\mathbb{R}^d} \nu^i_\theta(dx) = 1, \quad \forall \quad i \in \mathcal{I}_{-1}
\end{aligned}$$

First, we rewrite the inequality constraint above as follows. Recall that:

$$\begin{aligned}\frac{2\kappa}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\mu^i_{\theta,-1}(dx)&+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,+1}(dx)\\&+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,-1}(dx)+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{-1}}\|x-b_i(\theta)\|\nu^i_{\theta}(dx)\leq\delta.\end{aligned}$$

Hence,

$$\begin{aligned}\sup_{\mu^i_{\theta,\pm1},\nu^i_\theta}\quad&\frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}\int_{\mathbb{R}^d}\ell((x,+1),\theta)\mu^i_{\theta,+1}(dx)+\frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}\int_{\mathbb{R}^d}\ell((x,-1),\theta)\mu^i_{\theta,-1}(dx)\\&\qquad+\frac{1}{n}\sum_{\tilde{y}_i=-1}\int_{\mathbb{R}^d}\ell((x,-1),\theta)\nu^i_\theta(dx)\\\text{s.t. }&\frac{2\kappa}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\mu^i_{\theta,-1}(dx)+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,+1}(dx)\\&\qquad+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,-1}(dx)+\frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{-1}}\|x-b_i(\theta)\|\nu^i_\theta(dx)\leq\delta\\&\int_{\mathbb{R}^d}\mu^i_{\theta,+1}(dx)+\int_{\mathbb{R}^d}\mu^i_{\theta,-1}(dx)=1,\quad\forall\quad i\in\mathcal{I}_{+1}\\&\int_{\mathbb{R}^d}\nu^i_\theta(dx)=1,\quad\forall\quad i\in\mathcal{I}_{-1}\end{aligned}$$

Now, we can use duality to reformulate the infinite-dimensional optimization problem into a finite-dimensional problem:

$$\begin{aligned}&\sup_{\mathbb{P}\in\mathcal{P}(\theta)}\ \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle)-y\langle x,\theta\rangle]\\&=\begin{cases}\inf_{\alpha,s_i} & \alpha\delta+\frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}s_i+\frac{1}{n}\sum_{i\in\mathcal{I}_{-1}}t_i\\ \text{s.t.} & \sup_x\ell((x,+1),\theta)-\alpha\cdot\frac{1+\tilde{y}_i}{2}\|x-b_i(\theta)\|\leq s_i\quad\forall\ i\in\mathcal{I}_{+1}\\ & \sup_x\ell((x,-1),\theta)-\alpha\cdot\frac{1+\tilde{y}_i}{2}\|x-b_i(\theta)\|-\alpha\kappa(1+\tilde{y}_i)\leq s_i\quad\forall\ i\in\mathcal{I}_{+1}\\ & \sup_x\ell((x,-1),\theta)-\alpha\cdot\frac{1-\tilde{y}_i}{2}\|x-b_i(\theta)\|\leq t_i\quad\forall\ i\in\mathcal{I}_{-1}\\ & \alpha\geq0\end{cases}\end{aligned}$$

which is equivalent to:

$$\begin{aligned}&\sup_{\mathbb{P}\in\mathcal{P}(\theta)}\ \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle)-y\langle x,\theta\rangle]\\&=\begin{cases}\inf_{\alpha,s_i} & \alpha\delta+\frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}s_i+\frac{1}{n}\sum_{i\in\mathcal{I}_{-1}}t_i\\ \text{s.t.} & \sup_x\ell((x,+1),\theta)-\alpha\|x-b_i(\theta)\|\leq s_i\quad\forall\ i\in\mathcal{I}_{+1}\\ & \sup_x\ell((x,-1),\theta)-\alpha\|x-b_i(\theta)\|-2\alpha\kappa\leq s_i\quad\forall\ i\in\mathcal{I}_{+1}\\ & \sup_x\ell((x,-1),\theta)-\alpha\|x-b_i(\theta)\|\leq t_i\quad\forall\ i\in\mathcal{I}_{-1}\\ & \alpha\geq0\end{cases}\end{aligned}$$

We now invoke [Yu et al., 2021, Lemma A.1], which claims that for any $\tilde{y}\in\{+1,-1\}$ and $\tilde{x}\in\mathbb{R}^d$:

$$\sup_x\ell((x,\tilde{y}),\theta)-\alpha\|x-\tilde{x}\|=\begin{cases}\ell((\tilde{x},\tilde{y}),\theta) & \text{if }\|\theta\|\leq\alpha/(L+1)\\-\infty & \text{otherwise.}\end{cases}$$

We now have:

$$\begin{aligned}&\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle) - y\langle x,\theta\rangle]\\ &= \begin{cases} \inf_{\alpha,s_i} & \alpha\delta + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}} s_i + \frac{1}{n}\sum_{i\in\mathcal{I}_{-1}} t_i\\ \text{s.t.} & \ell((b_i(\theta),+1),\theta) \le s_i \quad \forall\ i\in\mathcal{I}_{+1}\\ & \ell((b_i(\theta),-1),\theta) - 2\alpha\kappa \le s_i \quad \forall\ i\in\mathcal{I}_{+1}\\ & \ell((b_i(\theta),-1),\theta) \le t_i \quad \forall\ i\in\mathcal{I}_{-1}\\ & \alpha \ge 0\\ & \|\theta\| \le \alpha/(L+1)\end{cases}\end{aligned}$$

In the above presented optimization problem we can conclude that:

$$\begin{aligned} t_i &= \phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle && \forall i\in\mathcal{I}_{-1}\\ s_i &= \max\{\ell((b_i(\theta),+1),\theta), \ell((b_i(\theta),-1),\theta) - 2\alpha\kappa\} && \forall i\in\mathcal{I}_{+1}.\end{aligned}$$

To further simplify the s_i expression, note that:

$$\begin{aligned} s_i &= \max\{\phi(\langle b_i(\theta),\theta\rangle) - \langle b_i(\theta),\theta\rangle, \phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle - 2\alpha\kappa\}\\ &= \phi(\langle b_i(\theta),\theta\rangle) - \langle b_i(\theta),\theta\rangle + \max\{0, 2\langle b_i(\theta),\theta\rangle - 2\alpha\kappa\}\\ &= \phi(\langle b_i(\theta),\theta\rangle) - \alpha\kappa + \max_{\gamma_i:|\gamma_i|\le 1} \gamma_i\left(\langle b_i(\theta),\theta\rangle - \alpha\kappa\right),\end{aligned}$$

so the overall objective can be written as:

$$\begin{aligned}&\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle) - y\langle x,\theta\rangle]\\ &= \begin{cases} \inf_\alpha \max_{\gamma:\|\gamma\|_\infty\le 1} & \alpha(\delta-\kappa) + \frac{1}{n}\sum_i \frac{1+\tilde{y}_i}{2}\left(\phi(\langle b_i(\theta),\theta\rangle) + \gamma_i\left(\langle b_i(\theta),\theta\rangle - \alpha\kappa\right)\right)\\ & \qquad + \frac{1}{n}\sum_i \frac{1-\tilde{y}_i}{2}\left(\phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle\right)\\ \text{s.t.} & \|\theta\| \le \alpha/(L+1)\end{cases}\end{aligned}$$

We claim that the minimax objective above is convex is θ. There are mainly two cases to analyze:

1. **Case I ($i\in\mathcal{I}_{+1}$)**: We have $b_i(\theta) = \tilde{x}_i$ as per the strategic classification model. Therefore $\langle b_i(\theta),\theta\rangle$ is a linear function. For every γ,α, we claim that the mapping $\theta\mapsto\phi(\langle b_i(\theta),\theta\rangle) + \gamma_i(\langle b_i(\theta),\theta\rangle - \alpha\kappa)$ is convex. Indeed, the assumption that ϕ is convex and the observation that $\langle b_i(\theta),\theta\rangle$ is affine in θ ensures the convexity.
2. **Case II ($i\in\mathcal{I}_{-1}$)**: We know from Lemma 5.2 that $\langle b_i(\theta),\theta\rangle$ is convex in θ. Moreover, the convexity of ϕ and the assumption that $z\mapsto\phi(z)+z$ is non-decreasing ensures that $\phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle$ is convex for every i.

This concludes the proof.

□

C DETAILS ON THE EXPERIMENTAL STUDY AND ADDITIONAL RESULTS

Code used to reproduce the results in the main paper is available at `https://drive.google.com/drive/folders/1spuB3R6vEU2AqaXxAxeeXo9z5QMVdtdl?usp=sharing`

C.1 Algorithms

In our experiments, we compare the OGDA-RR algorithm (Alg. 1) with three other zeroth-order algorithms—Optimistic Gradient Descent Ascent with Sampling with Replacement (OGDA-WR), Stochastic Gradient Descent Ascent with Random Reshuffling (SGDA-RR), and Stochastic Gradient Descent Ascent with Sampling with Replacement (SGDA-WR)—characterized by the update equations (43), (44), (45), respectively. For convenience, we have reproduced (6), the update equation for the OGDA-RR algorithm (Algorithm 1), as (42) below:

$$u_{i+1}^t = \text{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t; R^t, v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t)\right), \tag{42}$$

$$u_{i+1}^t = \text{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t \hat{F}_{j_i^t}(u_i^t; R^t, v_i^t) - \eta^t \hat{F}_{j_{i-1}^t}(u_i^t; R^t, v_i^t) + \eta^t \hat{F}_{j_{i-1}^t}(u_{i-1}^t; R^t, v_{i-1}^t)\right), \tag{43}$$

$$u_{i+1}^t = \text{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t; R^t, v_i^t)\right), \tag{44}$$

$$u_{i+1}^t = \text{Proj}_{\mathcal{X}\times\mathcal{Y}}\left(u_i^t - \eta^t \hat{F}_{j_i^t}(u_i^t; R^t, v_i^t)\right), \tag{45}$$

where the indices σ_i^t and j_i^t are as defined in Algorithms 2, 3, and 4.

Algorithm 2: OGDA-WR Algorithm

```
Input: stepsizes $\eta^t, R^t$, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}, u_0^{(0)}$, time horizon duration $T$;
for $t = 0, 1, \cdots, T-1$ do
    for $i = 0, \ldots, n-1$ do
        Sample $j_i^t \sim \mathsf{Unif}(\{1, \cdots, n\})$
        Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
        $u_{i+1}^t \leftarrow (43)$
    end
    $u_0^{(t+1)} \leftarrow u_n^t$
    $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
end
Output: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$.
```

Algorithm 3: SGDA-RR Algorithm

```
Input: stepsizes $\eta^t, R^t$, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}, u_0^{(0)}$, time horizon duration $T$;
for $t = 0, 1, \cdots, T-1$ do
    for $i = 0, \ldots, n-1$ do
        Sample $j_i^t \sim \mathsf{Unif}(\{1, \cdots, n\})$
        Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
        $u_{i+1}^t \leftarrow (44)$
    end
    $u_0^{(t+1)} \leftarrow u_n^t$
    $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
end
Output: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$.
```

Algorithm 4: SGDA-WR Algorithm

```
Input: stepsizes $\eta^t, R^t$, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}, u_0^{(0)}$, time horizon duration $T$;
for $t = 0, 1, \cdots, T-1$ do
    $\sigma^t = (\sigma_1^t, \cdots, \sigma_n^t) \leftarrow$ a random permutation of set $[n]$;
    for $i = 0, \ldots, n-1$ do
        Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
        $u_{i+1}^t \leftarrow (45)$
    end
    $u_0^{(t+1)} \leftarrow u_n^t$
    $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
end
Output: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$.
```

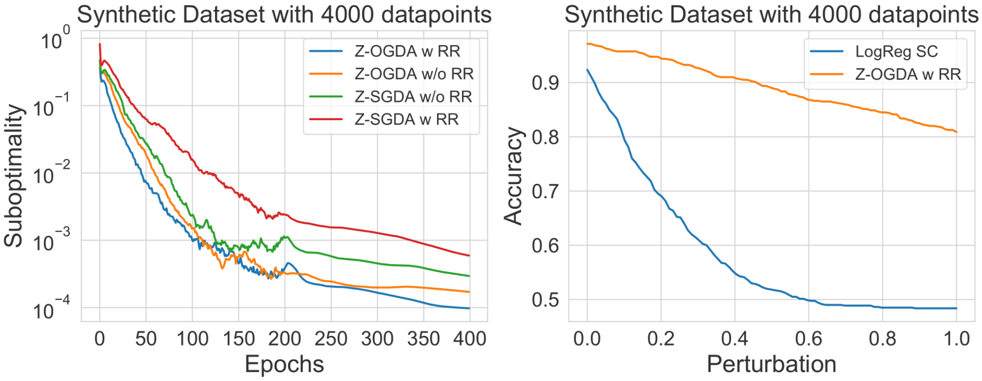


Figure 2: Experimental results for a synthetic dataset with $n = 4000$. (Left pane)) Suboptimality iterates generated by the four algorithms (A-I), (A-II), (A-III), (A-IV), respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right pane) Comparison between decay in accuracy of strategic classification with logistic regression (trained with $\zeta = 0.05$) and Alg. (A-I) with changes in perturbation.

C.2 Additional Experimental Results

In this section, we present more experimental findings, on both synthetic and real-world datasets, that reinforces the utility of the proposed algorithm. In all experimental results throughout this subsection, we take $\delta = 0.4$, $\kappa = 0.5$ and $\zeta = 0.05$.

C.2.1 Experimental Study On Synthetic Datasets

Figure 2 compares the performance of (A-I)-(A-IV) on a synthetic dataset (whose generating process is the same as that described in Section 6), with 4000 training points and 800 test points. Our proposed algorithm performs better empirically compared to most of its counterparts. Moreover, the proposed classifier, (A-I), is significantly more robust than a classifier obtained without considering adversarial perturbations. Note, however, that we cannot make any conclusive claims yet, because of the inherent randomness in these algorithms. Indeed, even if we fix the initialization, then there are two sources of randomness—the construction of the zeroth-order gradient estimator, and the sampling process that generates the data points.

To illustrate the variability in these algorithms' performance, we run each algorithm repeatedly on a data set with 500 synthetically generated data points, using the same initialization, and present confidence interval plots with ±2 standard deviations for the resulting performance (Figure 3). On average, our proposed algorithm (A-I) outperforms the other algorithms (A-II)-(A-IV). It is also interesting to point out that the performance of algorithms with random reshuffling is generally higher, and fluctuate less, compared to the performance of algorithms without random reshuffling.

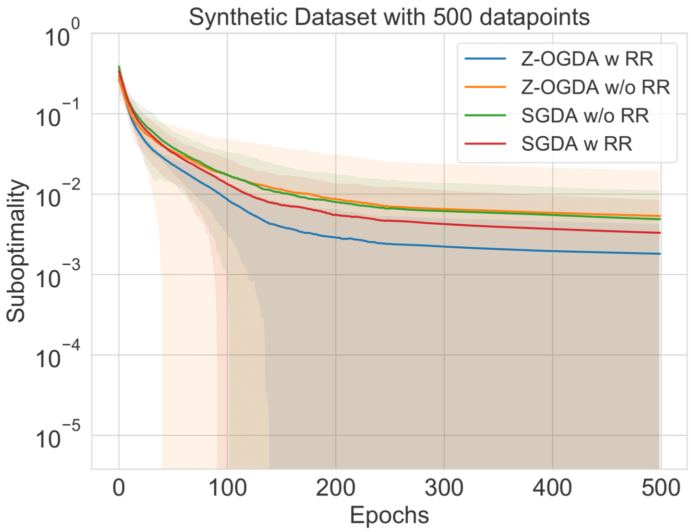


Figure 3: Experimental results for a synthetic dataset with $n = 500$. Suboptimality iterates generated by the four algorithms (A-I), (A-II), (A-III), and (A-IV) are respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, and *Z-SGDA w/o RR*.

We now illustrate the performance of our algorithm on two real-world data sets—the "GiveMeSomeCredit" dataset [3], and the "Porto Bank" data set[4].

C.2.2 Experimental Study on Credit Dataset

In modern times, banks use machine learning to determine whether or not to finance a customer. This process can be encoded into a classification framework, by using features such as age, debt ratio, monthly income to classify a customer as either likely or unlikely to default. However, those algorithms generally do not account for strategic or adversarial behavior on the part of the agents.

To illustrate the effect of our algorithm on datasets of practical significance, we deploy our algorithms on the "GiveMeSomeCredit"(GMSC) dataset, while assuming that the underlying features are subject to strategic or adversarial perturbations. We use a subset of the dataset of size 2000 with balanced labels. In Figure 4, we compare the empirical performance of our algorithm (A-I) with that of (A-II)-(A-IV). The left pane shows that (A-I) performs well, and the right pane illustrates that our classifier is significantly more robust to adversarial perturbations in data, compared to the strategic classification-based logistic regression algorithm developed recently in the literature [Dong et al., 2018].

C.2.3 Experimental Study on Porto-Bank Dataset

Next, we present empirical results obtained by applying our algorithm to the "Porto-Bank" dataset, which describes marketing campaigns of term deposits at Portuguese financial institutions. The classification task in this scenario aims to predict whether a customer with given features (eg. age, job, marital status etc.) would enroll for term deposits.

In Figure 5, we present the performance of our proposed algorithm (A-I) on the Porto-Bank dataset. For ease of illustration, we consider a subset of the dataset with 2000 training data points, 800 test data points, and balanced labels. In Figure 5, we compare the empirical performance of our algorithm (A-I) with that of (A-II)-(A-IV). The left pane shows that (A-I) performs well, while the right pane illustrates that our classifier is significantly more robust to adversarial perturbations in data, compared to the strategic classification-based logistic regression developed recently in the literature [Dong et al., 2018].

[3] This dataset can be found at `https://www.kaggle.com/c/GiveMeSomeCredit`

[4] This dataset can be found at `https://archive.ics.uci.edu/ml/datasets/bank+marketing`

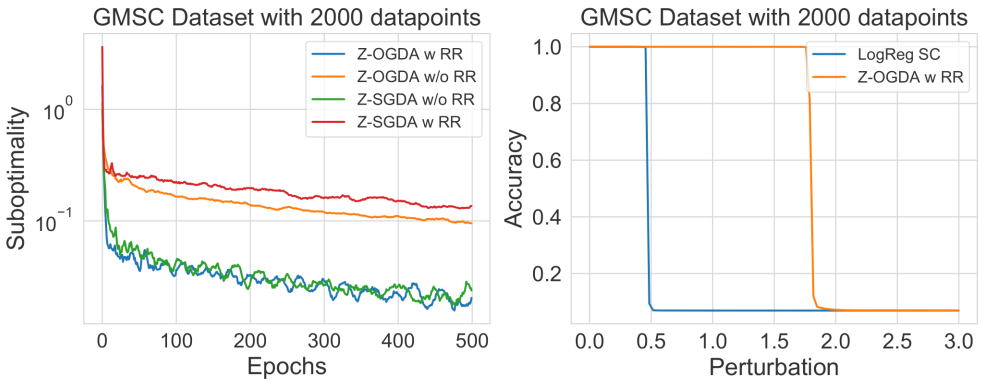


Figure 4: Experimental results for a balanced GiveMeSomeCredit dataset with $n = 2000$. (Left pane) Suboptimality iterates generated by the four algorithms (A-I), (A-II), (A-III), (A-IV), respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right pane) Comparison between decay in accuracy of strategic classification with logistic regression (originally trained with $\zeta = 0.05$) and Alg. (A-I) with changes in perturbation.

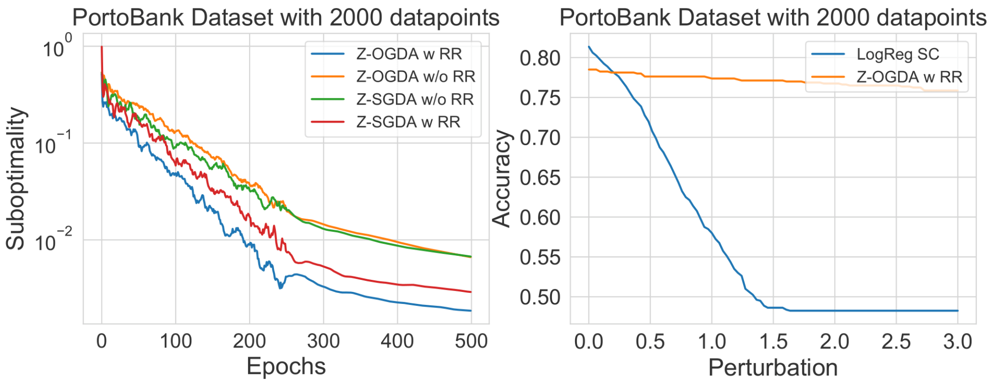


Figure 5: Experimental results for a balanced PortoBank dataset with $n = 2000$. (Left pane) Suboptimality iterates generated by the four algorithms (A-I), (A-II), (A-III), (A-IV), respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right pane) Comparison between decay in accuracy of strategic classification with logistic regression (originally trained with $\zeta = 0.05$) and Alg. (A-I) with change in perturbation.

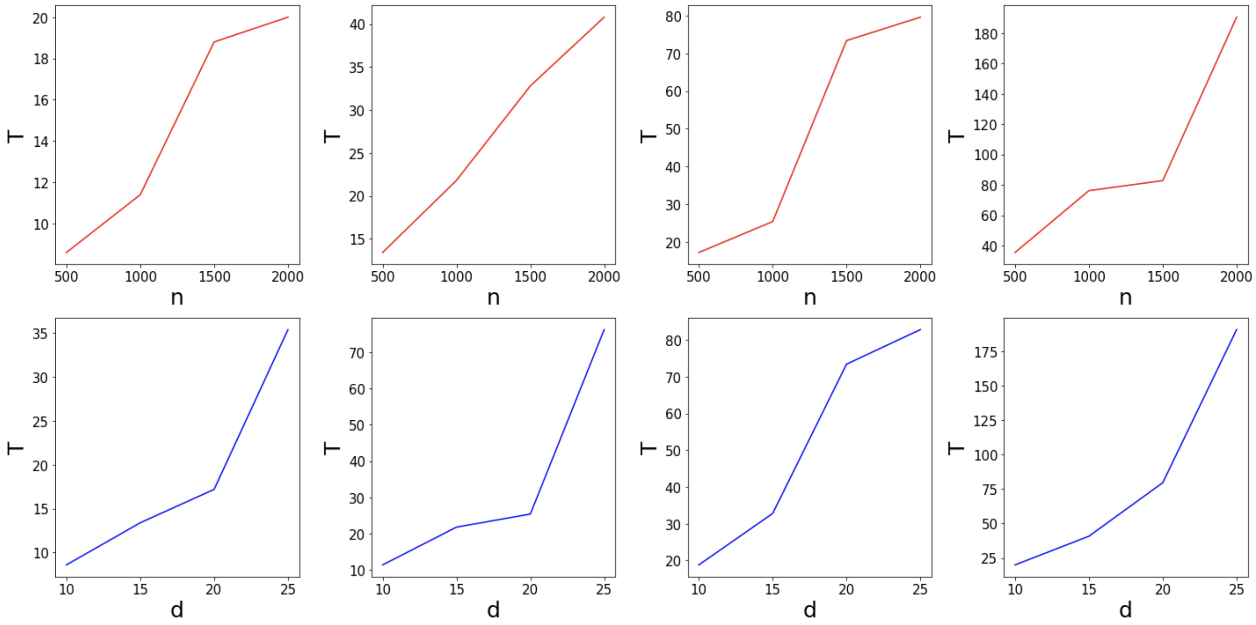


Figure 6: Experimental results presenting the number of samples required to reach $\epsilon-$suboptimality, with $\epsilon = 0.1$, for our algorithm (A-I) on synthetic dataset with varying values of $n \in \{500, 1000, 1500, 2000\}$ and $d \in \{10, 15, 20, 25\}$.

C.2.4 Effect of n, d on sample complexity

In this part, we demonstrate the empirical results that corroborates the theoretical dependence of sample complexity on n, d. For this purpose, we use synthetic dataset which is generated as per the method described in Section 6.1. Here we work in the setting where $n \in \{500, 1000, 1500, 2000\}$ and $d \in \{10, 15, 20, 25\}$. We fix the suboptimality to $\epsilon = 0.1$ and compute the number of samples required in each of the settings of n and d so that the iterates reach the $\epsilon-$suboptimality. We present the results in Figure 6.

C.3 Logistic regression as a Generalized linear model

The goal in logistic regression is to maximize the log-likelihood of the conditional probability of y (the *label*) given x (the *feature*). In this model, it is assumed that:

$$P(Y = 1|x, \theta) = \frac{1}{1 + \exp(-\langle x, \theta \rangle)}$$

This implies that:

$$P(Y = -1|x, \theta) = \frac{\exp(-\langle x, \theta \rangle)}{1 + \exp(-\langle x, \theta \rangle)}$$

Given a data point (x, y) the logistic loss is log-likelihood of observing y given x. For any θ and $y \in \{-1, 1\}$:

$$P(Y = y|x; \theta) = (P(Y = 1|x, \theta))^{\frac{1+y}{2}} \, (P(Y = -1|x, \theta))^{\frac{1-y}{2}}$$

Now, the log-likelihood is given by:

$$\begin{aligned}
L(x, y; \theta) &= \log(P(Y = y|x; \theta)) \\
&= \frac{1+y}{2} \log\left(\frac{1}{1+\exp(-\langle x, \theta\rangle)}\right) + \frac{1-y}{2} \log\left(\frac{\exp(-\langle x, \theta\rangle)}{1+\exp(-\langle x, \theta\rangle)}\right) \\
&= -\frac{1-y}{2} \langle x, \theta\rangle + \left(\frac{1+y}{2} + \frac{1-y}{2}\right) \log\left(\frac{1}{1+\exp(-\langle x, \theta\rangle)}\right) \\
&= -\frac{1-y}{2} \langle x, \theta\rangle - \log(1+\exp(-\langle x, \theta\rangle)) \\
&= \frac{y}{2} \langle x, \theta\rangle - \frac{1}{2} \langle x, \theta\rangle + \langle x, \theta\rangle - \log(1+\exp(\langle x, \theta\rangle)) \\
&= \frac{y}{2} \langle x, \theta\rangle + \frac{1}{2} \langle x, \theta\rangle - \log(1+\exp(\langle x, \theta\rangle))
\end{aligned}$$

The goal is to maximize the log likelihood, which is equivalent to minimizing the negative log likelihood. Thus the logistic regression minimizes the following loss:

$$\tilde{L}(x, y; \theta) = -L(x, y; \theta) = -\frac{y}{2} \langle x, \theta\rangle + \phi(\langle x, \theta\rangle)$$

where $\phi(\beta) = \log(1 + \exp(\beta)) - \frac{\beta}{2}$. If $y = 1$, then the above loss becomes:

$$\log(1 + \exp(\langle x, \theta\rangle)) - \langle x, \theta\rangle = \log(1 + \exp(-\langle x, \theta\rangle))$$

Otherwise, if $y = -1$, then the above loss becomes $\log(1 + \exp(\langle x, \theta\rangle))$. Thus, the above loss is equivalent to $\log(1 + \exp(-y \langle x, \theta\rangle))$.