Adversarial Classification via Distributional Robustness with Wasserstein Ambiguity

Nam Ho-Nguyen[1] and Stephen J. Wright[2]

[1]The University of Sydney
[2]University of Wisconsin–Madison

April 16, 2021

Abstract

We study a model for adversarial classification based on distributionally robust chance constraints. We show that under Wasserstein ambiguity, the model aims to minimize the conditional value-at-risk of the distance to misclassification, and we explore links to adversarial classification models proposed earlier and to maximum-margin classifiers. We also provide a reformulation of the distributionally robust model for linear classification, and show it is equivalent to minimizing a regularized ramp loss objective. Numerical experiments show that, despite the nonconvexity of this formulation, standard descent methods appear to converge to the global minimizer for this problem. Inspired by this observation, we show that, for a certain separable distribution, the only stationary point of the regularized ramp loss minimization problem is the global minimizer.

Keywords: Adversarial Classification, Distributional Robustness, Wasserstein Ambiguity, Ramp Loss, Non-convex

1 Introduction

Optimization models have been used for prediction and pattern recognition in data analysis as early as the work of Mangasarian [34] in the 1960s. Recent developments have seen models grow in size and complexity, with success on a variety of tasks, which has spurred many practical and theoretical advances in data science. However, it has been observed that models that achieve remarkable prediction accuracy on unseen data can lack robustness to small perturbations of the data [25, 45]. For example, the correct classification of a data point for a trained model can often be switched to incorrect by adding a small perturbation, carefully chosen. This fact is particularly problematic for image classification tasks, where the perturbation that yields misclassification can be imperceptible to the human eye[1].

This observation has led to the emergence of *adversarial machine learning*, a field that examines robustness properties of models to (potentially adversarial) data perturbations. Two streams of work in this area are particularly notable. The first is *adversarial attack* [12, 13, 36], where the aim is to "fool" a trained model by constructing adversarial perturbations. The second is *adversarial defense*, which focuses on model training methods that produce classifiers that are robust to perturbations [7, 15, 20, 33, 37, 46, 47, 48]. Most models for adversarial defense are based on robust optimization, where the training error is minimized subject to arbitrary perturbations of

[1] See, for example, `https://adversarial-ml-tutorial.org/introduction/`.

the data in a ball defined by some distance function (for example, a norm in feature space). As such, these algorithms are reminiscent of iterative algorithms from robust optimization [4, 27, 38]. Theoretical works on adversarial defense also focus on the robust optimization model, discussing several important topics such as hardness and fundamental limits [11, 22, 23, 24], learnability and risk bounds [54, 55], as well as margin guarantees and implicit bias for specific algorithms [14, 31].

In optimization under uncertainty and data-driven decision-making, the concept of *distributional robustness* offers an intriguing alternative to stochastic optimization and robust optimization [8, 17, 35, 50]. Instead of considering perturbations of the data (as in robust optimization), this approach considers perturbations in the *space of distributions* from which the data is drawn, according to some distance measure in distribution space (for example, ϕ-divergence or Wasserstein distance). This technique enjoys strong statistical guarantees and its numerical performance often outperforms models based on stochastic or robust optimization. In particular, for perturbations based on Wasserstein distances, the new distributions need not have the same support as the original empirical distribution.

In this paper, we explore adversarial defense by using ideas from distributional robustness and Wasserstein ambiguity sets. We focus on the fundamental classification problem in machine learning, and its formulation as an optimization problem in which we seek to minimize the probability of misclassification. We study a distributionally robust version of this problem and explore connections between maximum-margin classifiers and conditional value-at-risk objectives. We then focus on the linear classification problem. While convex linear classification formulations are well-known [5], the model we study is based on a "zero-one" loss function $r \mapsto \mathbf{1}(r \leq 0)$, which is discontinuous and thus nonconvex. However, we show that in the case of binary linear classification, the reformulation of the distributionally robust model gives rise to the "ramp loss" function L_R defined in (31), and we propose efficient first-order algorithms for minimization. While the ramp loss is nonconvex, the nonconvexity is apparently "benign"; the global minimizer appears to be found for sufficiently dense distributions. Indeed, we prove that in a special case, the global minimizer is the only stationary point. Numerical experiments confirm this observation.

1.1 Problem Description

Suppose that data ξ is drawn from some distribution P over a set S. In the learning task, we need to find a decision variable w, a classifier, from a space $\mathcal{W}$. For each $(w, \xi) \in \mathcal{W} \times S$, we evaluate the result of choosing classifier w for outcome ξ via a "safety function" $z : \mathcal{W} \times S \to \mathbb{R} \cup \{+\infty\}$. We say that w "correctly classifies" the point ξ when $z(w, \xi) > 0$, and w "misclassifies" ξ when $z(w, \xi) \leq 0$. Thus, we would like to choose w so as to minimize the probability of misclassification, that is,

$$\inf_{w \in \mathcal{W}} \mathbb{P}_{\xi \sim P} \left[z(w, \xi) \leq 0 \right]. \tag{1}$$

This fundamental problem is a generalization of the binary classification problem, which is obtained when $S = X \times \{\pm 1\}$, where $\xi = (x, y)$ is a feature-label pair, $\mathcal{W}$ describes the space of classifiers under consideration (e.g., linear classifiers or a reproducing kernel Hilbert space), and $z(w, (x, y)) = yw(x)$; so w correctly classifies (x, y) if and only if $\text{sign}(w(x)) = y$.

In the context of adversarial classification, we are interested in finding decisions $w \in \mathcal{W}$ which are *robust to (potentially adversarial) perturbations of the data* ξ. In other words, if our chosen w correctly classifies ξ (that is, $z(w, \xi) > 0$), then any small perturbation $\xi + \Delta$ should also be correctly classified, that is, $z(w, \xi + \Delta) > 0$ for "sufficiently small" Δ. To measure the size of perturbations, we use a distance function $c : S \times S \to [0, +\infty]$ that is nonnegative and lower semicontinuous with $c(\xi, \xi') = 0$ if and only if $\xi = \xi'$. (For binary classification $\xi = (x, y)$ mentioned above, the distance function can be $c((x,y),(x'y')) = \|x - x'\| + \mathbb{I}_{y=y'}(y,y')$ where $\mathbb{I}_A$ is the convex indicator of the set A.) For a classifier $w \in \mathcal{W}$, we define the *margin*, or *distance to misclassification*, of a point $\xi \in S$ as

$$d(w,\xi) := \inf_{\xi' \in S} \left\{ c(\xi,\xi') : z(w,\xi') \le 0 \right\}. \tag{2}$$

(Note that $d(w,\xi) = 0 \Leftrightarrow z(x,\xi) \le 0$.)

Two optimization models commonly studied in previous works on adversarial classification are the following:

$$\inf_{w \in \mathcal{W}} \mathbb{P}_{\xi \sim P} \left[d(w,\xi) \le \epsilon \right], \tag{3a}$$

$$\sup_{w \in \mathcal{W}} \mathbb{E}_{\xi \sim P} \left[d(w,\xi) \right]. \tag{3b}$$

The first model (3a), which is the most popular such model, aims to minimize the probability that the distance of a data point ξ to a bad result will be smaller than a certain threshold $\epsilon \ge 0$. Note that (1) is a special case of (3a) in which $\epsilon = 0$. The second model (3b) maximizes the expected margin. This model removes the need to choose a parameter ϵ, but Fawzi et al. [23, Lemma 1] has shown that this measure is inversely related to the probability of misclassification $\mathbb{P}_{\xi \sim P}[z(w,\xi) \le 0]$, that is, a lower probability of misclassification (good) leads to a lower expected distance (bad), and vice versa. Thus, this model is not used often.

With *distributional robustness*, rather than guarding against perturbations in the data points ξ, we aim to guard against perturbations of the distribution P of the data. In this paper, we study the following distributionally robust optimization (DRO) formulation (stated in two equivalent forms that will be used interchangeably throughout):

$$\inf_{w \in \mathcal{W}} \sup_{Q : d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q} \left[z(w,\xi) \le 0 \right] \quad \Leftrightarrow \quad \inf_{w \in \mathcal{W}} \sup_{Q : d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q} \left[d(w,\xi) = 0 \right], \tag{4}$$

that is, we aim to minimize the *worst-case* misclassification probability over a ball of distributions $\{Q : d_W(Q,P) \le \epsilon\}$. The ball is defined via the *Wasserstein distance* between two distributions, which is defined via the function c as follows:

$$d_W(P,Q) := \inf_{\Pi} \left\{ \mathbb{E}_{(\xi,\xi') \sim \Pi} \left[c(\xi,\xi') \right] : \Pi \text{ has marginals } P, Q \right\}. \tag{5}$$

In practice, the true distribution P is not known to us; we typically have only a finite sample $\xi_i \sim P$, $i \in [n]$ of training data, drawn from P, from which we can define the empirical distribution $P_n := \frac{1}{n} \sum_{i \in [n]} \delta_{\xi_i}$. We use P_n as the center of the ball of distributions, that is, we solve the formulation (4) in which P_n replaces P.

1.2 Contributions and outline

Other distance measures on distributions could be used in place of d_W in (4). In this paper, we explain why the choice of Wasserstein distance is natural in the context of adversarial defense. We do this by exploring the links between (4) and (3a) as well as maximum-margin classifiers (see Definition 2.1). Specifically, in Section 2.1, we show that for sufficiently small ϵ, (4) yields the maximum-margin classifier. In Section 2.2, we extend the link between the conditional value-at-risk of the distance function (2) and a chance-constrained version of (4) (observed by Xie [50]) to the probability minimization problem (4). This link yields an interpretation of optimal solutions of (4) for large ϵ, as optimizers of the conditional value-at-risk of the distance function (2).

In Section 3, we give a reformulation of (4) for linear classifiers, obtaining a regularized risk minimization problem with a "ramp loss" objective. This formulation highlights the link between distributional robustness and robustness to outliers, a criterion which has motivated the use of ramp loss in the past. We suggest a class of smooth approximations for the ramp loss, allowing problems with this objective to be solved (approximately) with standard continuous optimization algorithms.

In Section 4, we perform some numerical tests on a simple class of distributions. We observe that the regularized smoothed ramp loss minimization problem arising from (4), while nonconvex, is "benign" on this class, in the sense that the global minimum appears to be identified easily by smooth nonconvex optimization methods, when the number of training poitns n is sufficiently large. Motivated by this observation, we prove in Section 5 that in a certain restriction of the setting of Section 4, the ramp-loss problem indeed has only a single stationary point, which is therefore the global minimizer.

1.3 Related work

There are a number of works that explore distributional robustness for machine learning model training. These papers consider a distributionally robust version of empirical risk minimization, which seeks to minimize the worst-case risk over some ambiguity set of distributions around the empirical distribution. Lee and Raginsky [30] consider a distributionally robust ERM problem, exploring such theoretical questions as learnability of the minimax risk and its relationship to well-known function-class complexity measures. Their work targets smooth loss functions, thus does not apply to (4). Works that consider a Wasserstein ambiguity, similar to (4), include Chen and Paschalidis [16], Shafieezadeh-Abadeh et al. [41, 42], Sinha et al. [44]; whereas Hu et al. [28] uses a distance measure based on ϕ-divergences. For Wasserstein ambiguity, Sinha et al. [44] provide an approximation scheme for distributionally robust ERM by using the duality result of Lemma 2.3, showing convergence of this scheme when the loss is smooth and the distance c used to define the Wasserstein distance in (5) is strongly convex. When the loss function is of a "nice" form (e.g., logistic or hinge loss for classification, ℓ_1-loss for regression), Chen and Paschalidis [16], Kuhn et al. [29], Shafieezadeh-Abadeh et al. [41, 42] show that the incorporation of Wasserstein distributional robustness yields a *regularized* empirical risk minimization problem. This observation is quite similar to our results in Section 3, with a few key differences outlined in Remark 3.1. Also, discontinuous losses, including the "0-1" loss explored in our paper, are not considered by Chen and Paschalidis [16], Shafieezadeh-Abadeh et al. [41, 42], Sinha et al. [44]. Furthermore, none of these works provide an interpretation the optimal classifier like the one we provide in Section 2.

In this sense, the goals of Section 2 are similar to those of Hu et al. [28], who work with ϕ-divergence ambiguity sets. Their paper shows that the formulation that incorporates ϕ-divergence ambiguity does not result in classifiers different from those obtained by simply minimizing the empirical distribution. They suggest a modification of the ambiguity set and show experimental improvements over the basic ϕ-divergence ambiguity set. The main difference between our work and theirs is that we consider a different (Wasserstein-based) ambiguity set, which results in an entirely different analysis and computations. Furthermore, using ϕ-divergence ambiguity does not seem to have a strong theoretical connection with the traditional adversarial training model (3a), whereas we show that the Wasserstein ambiguity (4) has close links to (3a).

We mention some relevant works from the robust optimization-based models for adversarial training. Charles et al. [14] and Li et al. [31] both provide margin guarantees for gradient descent on an adversarial logistic regression model. We also give margin guarantees for the distributionally robust model (4) in Section 2, but ours are algorithm-independent, providing insight into use of the Wasserstein ambiguity set for adversarial defense. Bertsimas and Copenhaver [6] and Xu et al. [51, 52] have observed that for "nice" loss functions, (non-distributionally) robust models for ERM also reformulate to a regularized ERM problem.

Finally, we mention that our results concerning uniqueness of the stationary point in Section 5 are inspired by, and are similar in spirit to, local minima results for low-rank matrix factorization (see, for example, Chi et al. [18]).

2 Margin Guarantees and Conditional Value-at-Risk

In this section, we highlight the relationship between the main problem (4) and a generalization of maximum-margin classifiers, as well as the conditional value-at-risk of the margin function $d(w,\xi)$.

2.1 Margin guarantees for finite support distributions

We start by exploring the relationship between solutions to (4) and maximum margin classifiers. We recall the definition (2) of *margin* $d(w,\xi)$ for any $w \in \mathcal{W}$ and data point $\xi \in S$. We say that a classifier w has a *margin of at least* γ if $d(w,\xi) \geq \gamma$ for all $\xi \in S$. When $\gamma > 0$, this implies that a perturbation of size at most γ (as measured by the distance function c in (2)) for any data point ξ will still be correctly classified by w. In the context of guarding against adversarial perturbations of the data, it is clearly of interest to find a classifier w with maximum margin, that is, the one that has the largest possible γ. On the other hand, some datasets S cannot be perfectly separated, that is, for any classifier $w \in \mathcal{W}$, there will exist some $\xi \in S$ such that $d(w,\xi) = 0$. To enable discussion of maximum margins in both separable and non-separable settings, we propose a generalized margin concept in Definition 2.1 as the value of a bilevel optimization problem. We then show that solving the DRO formulation (4) is exactly equivalent to finding a generalized maximum margin classifier for small enough ambiguity radius ϵ. This highlights the fact that the Wasserstein ambiguity set is quite natural for modeling adversarial classification. We work with the following assumption on P.

Assumption 2.1. The distribution P has finite support, that is, $P = \sum_{i\in[n]} p_i \delta_{\xi_i}$, where each $p_i > 0$ and $\sum_{i\in[n]} p_i = 1$.

We make this assumption because for most *continuous* distributions, even our generalization of the margin will always be 0, so that a discussion of margin for such distributions is not meaningful. Since any training or test set we encounter in practice is finite, the finite-support case is worth our focus.

Under Assumption 2.1, we define the notion of *generalized margin* of P. For $w \in \mathcal{W}$ and $\rho \in [0,1]$, we define

$$\begin{aligned}
I(w) &:= \{i \in [n] : d(w,\xi_i) = 0\} \\
&\qquad \text{(points misclassified by } w\text{)} \\
\mathcal{I}(\rho) &:= \left\{ I \subseteq [n] : \sum_{i\in I} p_i \leq \rho \right\} \\
&\qquad \text{(subsets of } [n] \text{ with cumulative probability at most } \rho\text{)} \\
\eta(w) &:= \min_{i\in[n]\setminus I(w)} d(w,\xi_i) \\
&\qquad \text{(margin of } w \text{ with misclassified points excluded)} \\
\gamma(\rho) &:= \sup_{w\in\mathcal{W}} \{\eta(w) : I(w) \in \mathcal{I}(\rho)\}
\end{aligned}$$

(max. margin with at most fraction $\leq \rho$ of points misclassified).

The usual concept of margin is $\gamma(0)$. Given these quantities, we define the generalized maximum margin of P with respect to the classifiers $\mathcal{W}$ as the value of the following bilevel optimization problem.

Definition 2.1. Given P and $\mathcal{W}$, the *generalized maximum margin* is defined to be

$$\gamma^* := \sup_{w \in \mathcal{W}} \left\{ \eta(w) : w \in \arg\min_{w' \in \mathcal{W}} \mathbb{P}_{\xi \sim P}[d(w', \xi) = 0] \right\}. \tag{6}$$

Note that Definition 2.1 implicitly assumes that the arg min over $w' \in \mathcal{W}$ is achieved in (6). We show that under Assumption 2.1, this is indeed the case, and furthermore that $\gamma^* > 0$.

Proposition 2.1. *Suppose Assumption 2.1 holds. Define*

$$\rho^* := \inf \{\rho \in [0,1] : \gamma(\rho) > 0\}. \tag{7}$$

Then $\rho^* = \inf_{w' \in \mathcal{W}} \mathbb{P}_{\xi \sim P}[d(w', \xi) = 0]$, there exists $w \in \mathcal{W}$ such that

$$\mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] = \rho^*, \text{ and } \gamma^* = \gamma(\rho^*) > 0.$$

Proof. We first prove that under Assumption 2.1, the function $\rho \mapsto \gamma(\rho)$ is a right-continuous non-decreasing step function. The fact that $\gamma(\rho)$ is non-decreasing follows since $\mathcal{I}(\rho) \subseteq \mathcal{I}(\rho')$ for $\rho \leq \rho'$. To show that it is a right-continuous step function, consider the finite set of all possible probability sums $\mathcal{P} = \left\{\sum_{i \in I} p_i : I \subseteq [n]\right\} \subset [0,1]$. Let us order $\mathcal{P}$ as $\mathcal{P} = \left\{\rho^1, \ldots, \rho^K\right\}$ where $\rho^1 < \cdots < \rho^K$. There is no configuration $I \subseteq [n]$ such that $\rho^k < \sum_{i \in I} p_i < \rho^{k+1}$. Thus, $\mathcal{I}(\rho) = \mathcal{I}(\rho^k)$ and hence $\gamma(\rho) = \gamma(\rho^k)$ for $\rho \in [\rho_k, \rho_{k+1})$, proving the claim.

To prove the proposition, first note that there exists no classifier $w \in \mathcal{W}$ such that $\mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] = \underline{\rho} < \rho^*$, otherwise, we have $\gamma(\underline{\rho}) \geq \eta(w) > 0$, contradicting the definition of ρ^*. This shows that $\rho^* \leq \inf_{w' \in \mathcal{W}} \mathbb{P}_{\xi \sim P}[d(w', \xi) = 0]$. By the description of ρ^* as the infimal ρ such that $\gamma(\rho) > 0$ and by the right-continuity of $\gamma(\cdot)$ and the fact that $\gamma(\cdot)$ is a step function, we must have $\gamma(\rho^*) > 0$. Since $\gamma(\rho^*) > 0$, there must exist some $w \in \mathcal{W}$ such that $I(w) \in \mathcal{I}(\rho^*)$, that is, $\sum_{i \in I(w)} p_i = \mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] \leq \rho^*$. Since we cannot have $\mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] < \rho^*$ we conclude that $\mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] = \rho^*$. Therefore $\inf_{w \in \mathcal{W}} \mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] = \rho^*$. Furthermore, by definition, we have $\gamma^* = \gamma(\rho^*) > 0$. □

The following result gives a precise characterization of the worst-case misclassification probability of a classifier w for a radius ϵ that is smaller than the probability-weighted margin of w. It also gives a lower bound on worst-case error probability when ϵ is larger than this quantity.

Proposition 2.2. *Under Assumption 2.1, for $w \in \mathcal{W}$ such that*

$$\epsilon \leq \min_{i \in [n] \setminus I(w)} d(w, \xi_i) p_i,$$

we have

$$\sup_{Q : d_W(P,Q) \leq \epsilon} \mathbb{P}_{\xi \sim Q}[d(w, \xi) = 0] = \sum_{i \in I(w)} p_i + \frac{\epsilon}{\eta(w)}.$$

For $w \in \mathcal{W}$ such that $\epsilon > \min_{i \in [n] \setminus I(w)} d(w, \xi_i) p_i$, we have

$$\sup_{Q : d_W(P,Q) \leq \epsilon} \mathbb{P}_{\xi \sim Q}[d(w, \xi) = 0] > \sum_{i \in I(w)} p_i + \min_{i \in [n] \setminus I(w)} p_i.$$

To prove Proposition 2.2, we will use the following key key duality result for the worst-case error probability. Note that Lemma 2.3 does not need Assumption 2.1.

Lemma 2.3 (Blanchet and Murthy [8, Theorem 1, Eq. 15])**.** *For any $w \in \mathcal{W}$, we have*

$$\sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[z(w,\xi) \le 0] = \inf_{t>0} \left\{ \epsilon t + \mathbb{E}_{\xi \sim P} \left[\max\left\{0, 1 - t d(w,\xi)\right\}\right]\right\}. \tag{8a}$$

Proof of Proposition 2.2. First, by using (8a) in Lemma 2.3, using Assumption 2.1 and linear programming duality, we have

$$\begin{aligned} \sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[d(w,\xi) = 0] &= \inf_{t>0} \left\{ \epsilon t + \sum_{i \in [n]} p_i \max\{0, 1 - t d(w,\xi_i)\} \right\} \\ &= \max_{v} \left\{ \sum_{i \in [n]} v_i : \begin{array}{c} 0 \le v_i \le p_i, \;\; i \in [n] \\ \sum_{i \in [n]} d(w,\xi_i) v_i \le \epsilon \end{array} \right\}. \end{aligned}$$

The right-hand side is an instance of a fractional knapsack problem, which is solved by the following greedy algorithm:

> In increasing order of $d(w,\xi_i)$, increase v_i up to p_i or until the budget constraint $\sum_{i \in [n]} d(w,\xi_i) v_i \le \epsilon$ is tight, whichever occurs first.

Note that when $i \in I(w)$ we have $d(w,\xi_i) = 0$, so we can set $v_i = p_i$ for such values without making a contribution to the knapsack constraint. Hence, the value of the dual program is at least $\sum_{i \in I(w)} p_i$.

When $w \in \mathcal{W}$ is such that $\epsilon \le d(w,\xi_i) p_i$ for all $i \in [n] \setminus I(w)$, we will not be able to increase any v_i up to p_i for those $i \in [n] \setminus I(w)$ in the dual program. According to the greedy algorithm, we choose the smallest $d(w,\xi_i)$ amongst $i \in [n] \setminus I(w)$ — whose value corresponds to $\eta(w)$ — and increase this v_i up to $\epsilon / d(w,\xi_i) = \epsilon / \eta(w) \le p_i$. Therefore, we have

$$\sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[d(w,\xi) = 0] = \sum_{i \in I(w)} p_i + \frac{\epsilon}{\eta(w)}.$$

When $w \in \mathcal{W}$ is such that $\epsilon > d(w,\xi_i) p_i$ for some $i \in [n] \setminus I(w)$, the greedy algorithm for the dual program allows us to increase v_i up to p_i for at least one $i \in [n] \setminus I(w)$. Thus, by similar reasoning to the above, we have that a lower bound on the optimal objective is given by

$$\sum_{i \in I(w)} p_i + \min_{i \in [n] \setminus I(w)} p_i,$$

verifying the second claim. □

The main result in this section, which is a consequence of the first part of this proposition, is that as long as the radius $\epsilon > 0$ is sufficiently small, solving the DRO formulation (4) is equivalent to solving the bilevel optimization problem (6) for the generalized margin, that is, finding the w that, among those that misclassifies the smallest fraction of points $\mathbb{P}_{\xi \sim P}[d(w,\xi) = 0] = \rho^*$, achieves the largest margin $\eta(w) = \gamma^*$ on the correctly classified points. The required threshold for radius ϵ is $\epsilon = (\bar{\rho} - \rho^*)\gamma^*$, where $\bar{\rho}$ is the smallest probability that is strictly larger than ρ^*, that is,

$$\mathcal{P} := \left\{ \sum_{i \in I} p_i : I \notin \mathcal{I}(\rho^*) \right\} = \left\{ \sum_{i \in I} p_i : \sum_{i \in I} p_i > \rho^* \right\}, \quad \bar{\rho} := \min\left\{\rho : \rho \in \mathcal{P}\right\}. \tag{9}$$

We show too that classifiers that satisfy $\sum_{i \in I(w)} p_i = \rho^*$ but whose margin may be slightly suboptimal (greater that $\gamma^* - \delta$ but possibly less than γ^*) are also nearly optimal for (4).

Theorem 2.4. *Let Assumption 2.1 be satisfied. Suppose that $0 < \epsilon < (\bar{\rho} - \rho^*)\gamma^*$. Then, referring to the DRO problem (4), we have*

$$\min_{w \in \mathcal{W}} \sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q} \left[d(w, \xi) = 0 \right] = \rho^* + \frac{\epsilon}{\gamma^*}.$$

Furthermore, for any δ with $0 < \delta < \gamma^* - \epsilon/(\bar{\rho} - \rho^*)$, we have

$$\begin{aligned} &\{ w \in \mathcal{W} : I(w) \in \mathcal{I}(\rho^*), \ \eta(w) \ge \gamma^* - \delta \} \\ &= \left\{ w \in \mathcal{W} : I(w) \in \mathcal{I}(\rho^*), \sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q} \left[d(w, \xi) = 0 \right] \le \rho^* + \frac{\epsilon}{\gamma^* - \delta} \right\}. \end{aligned}$$

In particular, if there exists some $w \in \mathcal{W}$ such that $\mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] = \rho^*$, $\eta(w) = \gamma^*$, then w solves (4), and vice versa.

Proof. Since $\sup_{w \in \mathcal{W}: I(w) \in \mathcal{I}(\rho^*)} \eta(w) = \gamma(\rho^*) = \gamma^* > \epsilon/(\bar{\rho} - \rho^*)$, there exists some $w \in \mathcal{W}$ such that $\sum_{i \in I(w)} p_i = \rho^*$ and $\eta(w) > \epsilon/(\bar{\rho} - \rho^*)$, that is, $\epsilon < \eta(w)(\bar{\rho} - \rho^*)$. Now, since $\bar{\rho} - \rho^* \le p_i$ for all $i \in [n] \setminus I(w)$ (by definition of $\mathcal{P}$ and $\bar{\rho}$ in (9)), and since $\eta(w) \le d(w, \xi_i)$ for all $i \in [n] \setminus I(w)$, we have that $\epsilon < d(w, \xi_i) p_i$ for all $i \in [n] \setminus I(w)$. Therefore, by Proposition 2.2, we have for this w that

$$\sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[d(w, \xi) = 0] = \sum_{i \in I(w)} p_i + \frac{\epsilon}{\eta(w)} = \rho^* + \frac{\epsilon}{\eta(w)} < \bar{\rho}.$$

This implies that any $w \in \mathcal{W}$ such that $\sum_{i \in I(w)} p_i \ge \bar{\rho}$ is suboptimal for (4). (This is because even when we set $Q = P$ in (4), such a value of w has a worse objective than the w for which $\sum_{i \in I(w)} p_i = \rho^*$.) Furthermore, from Proposition 2.2 and the definition of $\bar{\rho}$, any $w \in \mathcal{W}$ such that $\sum_{i \in I(w)} p_i = \rho^*$ and $\epsilon \ge \min_{i \in [n] \setminus I(w)} d(w, \xi_i) p_i$ has

$$\sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[d(w, \xi) = 0] \ge \sum_{i \in I(w)} p_i + \min_{i \in [n] \setminus I(w)} p_i = \rho^* + \min_{i \in [n] \setminus I(w)} p_i \ge \bar{\rho},$$

hence is also suboptimal.

This means that all optimal and near-optimal solutions $w \in \mathcal{W}$ to (4) with $0 < \epsilon < (\bar{\rho} - \rho^*)\gamma^*$ are in the set

$$\left\{ w \in \mathcal{W} : \sum_{i \in I(w)} p_i = \rho^*, \ \epsilon < \min_{i \in [n] \setminus I(w)} d(w, \xi_i) p_i \right\},$$

and, by Proposition 2.2, the objective values corresponding to each such w are

$$\sup_{Q: d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[d(w, \xi) = 0] = \rho^* + \frac{\epsilon}{\eta(w)}.$$

By definition of $\gamma(\rho^*) = \gamma^*$, the infimal value for this objective is $\rho^* + \epsilon/\gamma^*$, and it is achieved as $\eta(w) \to \gamma^*$. The first claim is proved.

For the second claim, consider any $\delta \in (0, \gamma^* - \epsilon/(\bar{\rho} - \rho^*))$. We have for any w with $\sum_{i \in I(w)} p_i = \rho^*$ that

$$\eta(w) \ge \gamma^* - \delta \iff \rho^* + \frac{\epsilon}{\eta(w)} \le \rho^* + \frac{\epsilon}{\gamma^* - \delta}.$$

Furthermore, for such w and δ, we have $\epsilon < (\gamma^* - \delta)(\bar{\rho} - \rho^*) = (\gamma(\rho^*) - \delta)(\bar{\rho} - \rho^*) \leq \eta(w)(\bar{\rho} - \rho^*)$ so, by noting as in the first part of the proof that $\bar{\rho} - \rho^* \leq p_i$ and $\eta(w) \leq d(w, \xi_i)$ for all $i \in [n] \setminus I(w)$, we have $\epsilon < d(w, \xi_i)p_i$ for all $i \in [n] \setminus I(w)$. By applying Proposition 2.2 again, we obtain

$$\sup_{Q: d_W(P,Q) \leq \epsilon} \mathbb{P}_{\xi \sim Q}\left[z(w, \xi) \leq 0\right] = \rho^* + \frac{\epsilon}{\eta(w)} \leq \rho^* + \frac{\epsilon}{\gamma^* - \delta},$$

as required.

The final claim follows because, using the second claim, we have

$$\begin{aligned}
&\{w \in \mathcal{W} : \mathbb{P}_{\xi \sim P}[d(w, \xi) = 0] = \rho^*, \eta(w) = \gamma^*\} \\
&= \bigcap_{\delta > 0} \{w \in \mathcal{W} : I(w) \in \mathcal{I}(\rho^*),\ \eta(w) \geq \gamma^* - \delta\} \\
&= \bigcap_{\delta > 0} \left\{ w \in \mathcal{W} : I(w) \in \mathcal{I}(\rho^*),\ \sup_{Q: d_W(P,Q) \leq \epsilon} \mathbb{P}_{\xi \sim Q}\left[d(w, \xi) = 0\right] \leq \rho^* + \frac{\epsilon}{\gamma^* - \delta} \right\} \\
&= \left\{ w \in \mathcal{W} : \sup_{Q: d_W(P,Q) \leq \epsilon} \mathbb{P}_{\xi \sim Q}\left[d(w, \xi) = 0\right] = \rho^* + \frac{\epsilon}{\gamma^*} \right\},
\end{aligned}$$

as desired.

□

Theorem 2.4 shows that, for small Wasserstein ball radius ϵ, the solution of (4) matches the maximum-margin solution of the classification problem, in a well defined sense. How does the solution of (4) compare with the minimizer of the more widely used model (3a)? It is not hard to see that when the parameter ϵ in (3a) is chosen so that $\epsilon < \gamma^*$, the solution of (3a) will be a point w with margin $\eta(w) \geq \epsilon$. (Such a point will achieve an objective of zero in (3a).) However, in contrast to Theorem 2.4, this point may not attain the maximum possible margin γ^*. The margin that we obtain very much depends on the algorithm used to solve (3a). For fully separable data, for which $\rho^* = 0$ and $\gamma^* = \gamma(0) > 0$, Gunasekar et al. [26] show that gradient descent applied to a special case of this problem achieves a margin of γ with iteration count exponential in $(\gamma^* - \gamma)^{-1}$. Charles et al. [14] and [31] show that when adversarial training methods are applied to this same problem, where the balls around the adversarial samples are chosen with size γ, then gradient descent achieves a separation of γ in an iteration count *polynomial* in $(\gamma^* - \epsilon)^{-1}$. However, for (3a) to be a useful formulation, ϵ should be taken as close to γ^* as possible, to strengthen the margin guarantee. By contrast, Theorem 2.4 shows that, in the more general setting of non-separable data, the maximum-margin solution is attained from (4) when the parameter ϵ is taken to be *any* value below the threshold $\gamma^*(\bar{\rho} - \rho^*)$.

2.2 Conditional value-at-risk characterization

Section 2.1 gives insights into the types of solutions that the distributionally robust model (4) recovers when the Wasserstein radius ϵ is below a certain threshold. When ϵ is above this threshold however, (4) may no longer yield a maximum-margin solution. In this section, we show in Theorem 2.6 that, in general, (4) is intimately related to optimizing the conditional value-at-risk of the distance random variable $d(w, \xi)$. Thus, when ϵ is above the threshold of Theorem 2.4, (4) still has the effect of pushing data points ξ away from the error set $\{\xi \in S : z(w, \xi) \leq 0\}$ as much as possible, thereby encouraging robustness to perturbations. We note that unlike Section 2.1, we make no finite support assumptions on the distribution P, that is, Assumption 2.1 need not hold for our results below.

In stochastic optimization, when outcomes of decisions are random, different risk measures may be used to aggregate these random outcomes into a single measure of desirability (see, for example, [3, 40]). The most familiar risk measure is expectation. However, this measure has the drawback of being indifferent between a profit of 1 and a loss of -1 with equal probability, and a profit of 10 and a loss of -10 with equal probability. In contrast, other risk measures can adjust to different degrees of risk aversion to random outcomes, that is, they can penalize bad outcomes more heavily than good ones. The conditional value-at-risk (CVaR) is a commonly used measure that captures risk aversion and has several appealing properties. Roughly speaking, it is the conditional expectation for the ρ-quantile of most risky values, for some user-specified $\rho \in (0,1)$ which controls the degree of risk aversion. Formally, for a non-negative random variable $\nu(\xi)$ where low values are considered risky (that is, "bad"), CVaR is defined as follows:

$$\mathrm{CVaR}_\rho(\nu(\xi); P) := \sup_{t>0} \left\{ t + \frac{1}{\rho} \mathbb{E}_{\xi\sim P} \left[\min\left\{0, \nu(\xi) - t\right\}\right] \right\}. \tag{10}$$

Xie [50, Corollary 1] gives a characterization of the chance constraint

$$\max_{Q: d_W(P,Q)\le\epsilon} \mathbb{P}_{\xi\sim Q}[z(w,\xi) \le 0] \le \rho$$

in terms of the CVaR of $d(w,\xi)$ when $P = P_n$, a discrete distribution. We provide a slight generalization to arbitrary P.

Lemma 2.5. *Fix $\rho \in (0,1)$ and $\epsilon > 0$. Then, for all $w \in \mathcal{W}$, we have*

$$\sup_{Q: d_W(P,Q)\le\epsilon} \mathbb{P}_{\xi\sim Q}[z(w,\xi) \le 0] \le \rho \iff \rho\, \mathrm{CVaR}_\rho(d(w,\xi); P) \ge \epsilon. \tag{11}$$

Proof. We prove first the reverse implication in (11). Suppose that (following (10)) we have

$$\rho\, \mathrm{CVaR}_\rho(d(w,\xi); P) = \sup_{t>0} \left\{ \rho t + \mathbb{E}_{\xi\sim P} \left[\min\left\{0, d(w,\xi) - t\right\}\right]\right\} \ge \epsilon,$$

then for all $0 < \epsilon' < \epsilon$, there exists some $t > 0$ such that

$$\rho t + \mathbb{E}_{\xi\sim P} \left[\min\left\{0, d(w,\xi) - t\right\}\right] > \epsilon'.$$

Dividing by t, we obtain $\rho + \mathbb{E}_{\xi\sim P}\left[\min\left\{0, d(w,\xi)/t - 1\right\}\right] > \epsilon'/t$, so by rearranging and substituting $t' = 1/t$, we have

$$\begin{aligned} \rho &> \frac{\epsilon'}{t} + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - \frac{1}{t} d(w,\xi)\right\}\right] \\ &\ge \inf_{t'>0} \left\{ \epsilon' t' + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - t' d(w,\xi)\right\}\right]\right\}. \end{aligned}$$

Note that the function

$$\epsilon' \mapsto \inf_{t'>0} \left\{ \epsilon' t' + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - t' d(w,\xi)\right\}\right]\right\} \in [0,\rho]$$

is concave and bounded, hence continuous. This fact together with the previous inequality implies that

$$\inf_{t'>0} \left\{ \epsilon t' + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - t' d(w,\xi)\right\}\right]\right\} \le \rho,$$

which, when combined with (8a), proves that the reverse implication holds in (11).

We now prove the forward implication. Suppose that the left-hand condition in (11) is satisfied for some $\rho \in (0,1)$, and for contradiction that there exists some $\epsilon' \in (0,\epsilon)$ such that

$$\rho \operatorname{CVaR}_\rho(d(w,\xi);P) = \sup_{t>0} \left\{\rho t + \mathbb{E}_{\xi\sim P}\left[\min\left\{0, d(w,\xi)-t\right\}\right]\right\} \le \epsilon' < \epsilon.$$

Then for all $t > 0$, we have

$$\begin{aligned} &\rho t + \mathbb{E}_{\xi\sim P}\left[\min\left\{0, d(w,\xi)-t\right\}\right] \le \epsilon' \\ \implies\ & \rho \le \frac{\epsilon'}{t} + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-\frac{1}{t}d(w,\xi)\right\}\right] \\ \implies\ & \rho \le \inf_{t'>0}\left\{\epsilon' t' + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-t'd(w,\xi)\right\}\right]\right\}. \end{aligned}$$

Since $\epsilon' < \epsilon$, and using the left-hand condition in (11) together with (8a), we have

$$\begin{aligned} &\rho \le \inf_{t>0}\left\{\epsilon' t + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-td(w,\xi)\right\}\right]\right\} && (12)\\ &\quad \le \inf_{t>0}\left\{\epsilon t + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-td(w,\xi)\right\}\right]\right\} \le \rho, && (13)\\ \implies\ & \rho = \inf_{t>0}\left\{\epsilon t + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-td(w,\xi)\right\}\right]\right\}. && (14) \end{aligned}$$

Since $\epsilon' < \epsilon$, there cannot exist any $t > 0$ such that

$$\rho = \epsilon t + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-td(w,\xi)\right\}\right].$$

Let ρ_k and t_k be sequences such that $1 > \rho_k > \rho$, $\rho_k \to \rho$, $t_k > 0$, and

$$\rho_k \ge \epsilon t_k + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-t_k d(w,\xi)\right\}\right] > \rho.$$

Since $\epsilon > 0$, there cannot be any subsequence of t_k that diverges to ∞, since in that case $\epsilon t_k + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-t_k d(w,\xi)\right\}\right] \ge \epsilon t_k$ could not be bounded by $\rho_k < 1$. Thus $\{t_k\}$ is bounded, and there exists a convergent subsequence, so we assume without loss of generality that $t_k \to \tau$. By the dominated convergence theorem, $\mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-t_k d(w,\xi)\right\}\right] \to \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-\tau d(w,\xi)\right\}\right]$, and $\epsilon t_k \to \epsilon\tau$. But then since

$$\rho < \epsilon t_k + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-t_k d(w,\xi)\right\}\right] \le \rho_k \to \rho,$$

we have by the squeeze theorem that

$$\epsilon\tau + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-\tau d(w,\xi)\right\}\right] = \rho.$$

But then, by the fact noted after (14), we must have $\tau = 0$ so $\rho = 1$ (from (14)), which contradicts our assumption that $\rho \in (0,1)$. □

In the case of classification, the minimizers of (4) correspond exactly to the maximizers of $\operatorname{CVaR}_\rho(d(w,\xi);P)$, where ρ is the optimal worst-case error probability, as we show now.

Theorem 2.6. *Fix some $\rho \in [0,1]$ and define ϵ (using (10)) as follows:*

$$\epsilon := \rho \sup_{w\in\mathcal{W}} \operatorname{CVaR}_\rho(d(w,\xi);P) = \sup_{t>0}\left\{\rho t + \mathbb{E}_{\xi\sim P}\left[\min\left\{0, d(w,\xi)-t\right\}\right]\right\}. \tag{15}$$

If $0 < \epsilon < \infty$*, then*

$$\rho = \inf_{w \in \mathcal{W}} \sup_{Q : d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[z(w,\xi) \le 0].$$

Furthermore, the optimal values of w *coincide, that is,*

$$\arg\min_{w \in \mathcal{W}} \sup_{Q : d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[z(w,\xi) \le 0] = \arg\max_{w \in \mathcal{W}} \mathrm{CVaR}_\rho(d(w,\xi); P).$$

Proof. For any $w \in \mathcal{W}$ and $t > 0$, we have from (15) and (10) that

$$\epsilon \ge \sup_{t'>0} \left\{ \rho t' + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t'\right\}\right] \right\} \ge \rho t + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t\right\}\right].$$

Dividing by t and rearranging, we obtain

$$\frac{\epsilon}{t} + \mathbb{E}_{\xi \sim P}\left[\max\left\{0, 1 - \frac{1}{t} d(w,\xi)\right\}\right] \ge \rho.$$

Taking the infimum over $t > 0$, using (8a) (noting that $1/t > 0$), then taking the infimum over $w \in \mathcal{W}$, we obtain

$$\rho \le \inf_{w \in \mathcal{W}} \sup_{Q : d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[z(w,\xi) \le 0]. \tag{16}$$

In the remainder of the proof, we show that equality is obtained in this bound, when $0 < \epsilon < \infty$.

Trivially, the inequality in (16) can be replaced by an equality when $\rho = 1$. We thus consider the case of $\rho < 1$, and suppose for contradiction that there exists some $\rho' \in (\rho, 1]$ such that for all $w \in \mathcal{W}$, we have

$$\rho < \rho' < \sup_{Q : d_W(P,Q) \le \epsilon} \mathbb{P}_{\xi \sim Q}[z(w,\xi) \le 0].$$

It follows from Lemma 2.5 that for all $w \in \mathcal{W}$, we have

$$\sup_{t>0} \left\{ \rho' t + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t\right\}\right] \right\} < \epsilon. \tag{17}$$

By taking the supremum over $w \in \mathcal{W}$ in this bound, and using $\rho' > \rho$ and the definition of ϵ in (15), we have that

$$\begin{aligned} \epsilon &\ge \sup_{w \in \mathcal{W}, t>0} \left\{ \rho' t + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t\right\}\right] \right\} \\ &\ge \sup_{w \in \mathcal{W}, t>0} \left\{ \rho t + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t\right\}\right] \right\} = \epsilon, \end{aligned}$$

so that

$$\begin{aligned} \epsilon &= \sup_{w \in \mathcal{W}, t>0} \left\{ \rho' t + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t\right\}\right] \right\} \\ &= \sup_{w \in \mathcal{W}, t>0} \left\{ \rho t + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w,\xi) - t\right\}\right] \right\}. \end{aligned} \tag{18}$$

From $\rho < \rho'$, (17), and (18), we can define sequences ϵ_k, $t_k > 0$, and $w_k \in \mathcal{W}$ such that $\epsilon_k \nearrow \epsilon$ and

$$\epsilon_k < \rho t_k + \mathbb{E}_{\xi \sim P}\left[\min\left\{0, d(w_k,\xi) - t_k\right\}\right] < \epsilon.$$

By rearranging these inequalities, we obtain

$$\frac{\epsilon_k}{t_k} + \mathbb{E}_{\xi \sim P}\left[\max\left\{0, 1 - \frac{1}{t_k} d(w_k,\xi)\right\}\right]$$

$$\leq \rho < \frac{\epsilon}{t_k} + \mathbb{E}_{\xi \sim P}\left[\max\left\{0, 1 - \frac{1}{t_k} d(w_k, \xi)\right\}\right].$$

Since $\epsilon_k \to \epsilon$, we have either that t_k is bounded away from 0, in which case $\epsilon/t_k + \mathbb{E}_{\xi\sim P}[\max\{0, 1 - d(w_k,\xi)/t_k\}] \to \rho$; or there exists a subsequence on which $t_k \to 0$. In the former case, we have for k sufficiently large that

$$\begin{aligned}
&\frac{\epsilon}{t_k} + \mathbb{E}_{\xi\sim P}[\max\{0, 1 - d(w_k,\xi)/t_k\}] \leq \rho + \frac{\rho' - \rho}{2} < \rho' \\
&\implies \epsilon < \rho' t_k + \mathbb{E}_{\xi\sim P}[\min\{0, d(w_k,\xi) - t_k\}] \\
&\implies \epsilon < \sup_{w\in\mathcal{W}} \sup_{t>0} \left\{\rho' t + \mathbb{E}_{\xi\sim P}[\min\{0, d(w,\xi) - t\}]\right\},
\end{aligned}$$

which contradicts (17). We consider now the other case, in which there is a subsequence for which $t_k \to 0$, and assume without loss of generality that the full sequence has $t_k \to 0$. Since $\mathbb{E}_{\xi\sim P}[\min\{0, d(w_k,\xi) - t_k\}] \leq 0$ for any k, it follows that

$$\begin{aligned}
0 &\geq \limsup_{k\to\infty} \left\{\rho' t_k + \mathbb{E}_{\xi\sim P}[\min\{0, d(w_k,\xi) - t_k\}]\right\} \\
&\geq \limsup_{k\to\infty} \left\{\rho t_k + \mathbb{E}_{\xi\sim P}[\min\{0, d(w_k,\xi) - t_k\}]\right\} \\
&\geq \lim_{k\to\infty} \epsilon_k = \epsilon,
\end{aligned}$$

so that $\epsilon \leq 0$. This contradicts the assumption that $\epsilon > 0$, so we must have

$$\rho = \inf_{w\in\mathcal{W}} \sup_{Q: d_W(P,Q)\leq\epsilon} \mathbb{P}_{\xi\sim Q}[z(w,\xi) \leq 0].$$

This completes our proof of the first claim of the theorem.

Let $w \in \mathcal{W}$ be a maximizer of the CVaR, so that $\epsilon = \rho\,\mathrm{CVaR}_\rho(d(w,\xi); P)$. Then by Lemma 2.5, we have

$$\sup_{Q: d_W(P,Q)\leq\epsilon} \mathbb{P}_{\xi\sim Q}[z(w,\xi) \leq 0] \leq \rho,$$

so the same value of w is also a minimizer of the worst-case error probability. A similar argument shows that minimizers of the worst-case error probability are also maximizers of the CVaR. □

3 Reformulation and Algorithms for Linear Classifiers

In this section, we formulate (4) for a common choice of distance function c and safety function z, and discuss algorithms for solving this formulation. We make use of the following assumption.

Assumption 3.1. We have $\mathcal{W} = \mathbb{R}^d \times \mathbb{R}$ and $S = \mathbb{R}^d \times \{\pm 1\}$. Write $\bar{w} = (w_0, b_0) \in \mathbb{R}^d \times \mathbb{R}$ and $\xi = (x, y) \in \mathbb{R}^d \times \{\pm 1\}$. Define $c(\xi, \xi') := \|x - x'\| + \mathbb{I}_{y=y'}(y, y')$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$ and $\mathbb{I}_A(\cdot)$ is the convex indicator function where $\mathbb{I}_A(y, y') = 0$ if $(y, y') \in A$ and ∞ otherwise. Furthermore, $z(\bar{w}, \xi) := y(\langle w_0, \xi\rangle + b_0)$ where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product on $\mathbb{R}^d$.

From Lemma 2.3, the DRO problem (4) is equivalent to

$$\inf_{\bar{w}=(w_0,b_0)\in\mathbb{R}^d\times\mathbb{R},\, t>0} \left\{\epsilon t + \mathbb{E}_{\xi\sim P}\left[\max\{0, 1 - t d(\bar{w}, \xi)\}\right]\right\}. \tag{19}$$

Letting $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$ from Assumption 3.1, the distance to misclassification $d(\bar{w},\xi)$ is as follows

$$d(\bar{w},\xi) = d((w_0,b_0),(x,y)) = \begin{cases} \frac{\max\{0,y(\langle w_0,x\rangle+b_0)\}}{\|w_0\|_*}, & w_0 \neq 0 \\ \infty, & w_0 = 0,\ yb_0 > 0 \\ 0, & w_0 = 0,\ yb_0 \leq 0. \end{cases} \tag{20}$$

When $w_0 \neq 0$, we can define the following nonlinear transformation:

$$w \leftarrow \frac{tw_0}{\|w_0\|_*}, \quad b \leftarrow \frac{tb_0}{\|w_0\|_*}, \tag{21}$$

noting that $t = \|w\|_*$, and substitute (20) into (19) to obtain

$$\inf_{w\in\mathbb{R}^d, b\in\mathbb{R}} \left\{\epsilon\|w\|_* + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-\max\left\{0, y(\langle w,x\rangle + b)\right\}\right\}\right]\right\}. \tag{22}$$

In fact, the next result shows that this formulation is equivalent to (19) even when $w_0 = 0$. (Here, we use the term "δ-optimal solution" to refer to a point whose objective value is within δ of the optimal objective value for that problem.)

Theorem 3.1. *Under Assumption 3.1, (22) is equivalent to (19). Moreover, any δ-optimal solution (w,b) for (22) can be converted into a δ-optimal solution t and $\bar{w} = (w_0,b_0)$ for (19) as follows:*

$$t = \|w\|_*, \quad (w_0,b_0) := \begin{cases} \left(\frac{w}{\|w\|_*}, \frac{b}{\|w\|_*}\right) & w \neq 0 \\ (0,b), & w = 0. \end{cases} \tag{23}$$

Proof. The first part of the proof shows that the optimal value of (22) is less than or equal to that of (19), while the second part proves the converse.

To prove that the optimal value of (22) is less than or equal to that of (19), it suffices to show that given any $\bar{w} = (w_0,b_0)$, we can construct a sequence $\{(w^k,b^k)\}_{k\in\mathbb{N}}$ such that

$$\begin{aligned} &\epsilon\|w^k\|_* + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-\max\left\{0, y(\langle w^k,x\rangle + b^k)\right\}\right\}\right] \\ &\quad \to \inf_{t\geq 0}\left\{\epsilon t + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-td(\bar{w},\xi)\right\}\right]\right\}. \end{aligned} \tag{24}$$

Consider first the case of $w_0 \neq 0$, and let $t_k > 0$ be a sequence such that

$$\begin{aligned} &\lim_{k\to\infty}\left\{\epsilon t_k + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-t_k d(\bar{w},\xi)\right\}\right]\right\} \\ &= \inf_{t>0}\left\{\epsilon t + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1-td(\bar{w},\xi)\right\}\right]\right\}. \end{aligned} \tag{25}$$

Following (21), we define $w^k := t_k w_0/\|w_0\|_*$ and $b^k := t_k b_0/\|w_0\|_*$. We then have from (20) that

$$\begin{aligned} \max\{0, y(\langle w^k,x\rangle + b^k)\} &= \max\left\{0, t_k\frac{y(\langle w_0,x\rangle + b_0)}{\|w_0\|_*}\right\} \\ &= t_k\frac{\max\{0, y(\langle w_0,x\rangle + b_0)}{\|w_0\|_*} = t_k d(\bar{w},\xi). \end{aligned}$$

Thus, the left-hand sides of (25) and (24) are equivalent, so (24) holds.

Next, we consider the case of $\bar{w} = (w_0, b_0)$ with $w_0 = 0$. Note that $d(\bar{w}, \xi) = 0$ when $yb_0 \leq 0$ and $d(\bar{w}, \xi) = \infty$ when $yb_0 > 0$, we have $\max\{0, 1 - td(\bar{w}, \xi)\} = \mathbf{1}(yb \leq 0)$ for all $t > 0$, where $\mathbf{1}(\cdot)$ has the value 1 when its argument is true and 0 otherwise. Thus, we have

$$\begin{aligned}
&\inf_{t>0} \{\epsilon t + \mathbb{E}_{\xi\sim P}[\max\{0, 1 - td(\bar{w}, \xi)\}]\} \\
&= \mathbb{P}_{\xi\sim P}[yb_0 \leq 0] = \begin{cases} \mathbb{P}_{\xi\sim P}[y \leq 0], & b_0 > 0 \\ 1, & b_0 = 0 \\ \mathbb{P}_{\xi\sim P}[y \geq 0], & b_0 < 0. \end{cases}
\end{aligned} \tag{26}$$

Now choose $w^k = 0$ and $b^k = kb_0$ for $k = 1, 2, \ldots$. We then have

$$\begin{aligned}
&\max\left\{0, 1 - \max\left\{0, y(\langle w^k, x\rangle + b^k)\right\}\right\} \\
&= \max\{0, 1 - \max\{0, kyb_0\}\} \\
&= \max\{0, 1 - \max\{0, kyb_0\}\}\,\mathbf{1}(b_0 > 0) + \mathbf{1}(b_0 = 0) \\
&\quad + \max\{0, 1 - \max\{0, kyb_0\}\}\,\mathbf{1}(b_0 < 0) \\
&= (\max\{0, 1 - kyb_0\}\mathbf{1}(y > 0) + \mathbf{1}(y \leq 0))\,\mathbf{1}(b_0 > 0) + \mathbf{1}(b_0 = 0) \\
&\quad + (\mathbf{1}(y \geq 0) + \max\{0, 1 - kyb_0\}\mathbf{1}(y < 0))\,\mathbf{1}(b_0 < 0).
\end{aligned} \tag{27}$$

Now notice that for the first and last terms in this last expression, we have by taking limits as $k \to \infty$ that

$$\begin{aligned}
(\max\{0, 1 - yb_0k\}\mathbf{1}(y > 0) + \mathbf{1}(y \leq 0))\,\mathbf{1}(b_0 > 0) &\to \mathbf{1}(y \leq 0)\mathbf{1}(b_0 > 0), \\
(\mathbf{1}(y \geq 0) + \max\{0, 1 - yb_0k\}\mathbf{1}(y < 0))\,\mathbf{1}(b_0 < 0) &\to \mathbf{1}(y \geq 0)\mathbf{1}(b_0 < 0),
\end{aligned}$$

both pointwise, and everything is bounded by 1. Therefore, by the dominated convergence theorem, we have from (27) that

$$\mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - \max\left\{0, y(\langle w^k, x\rangle + b^k)\right\}\right\}\right] \to \begin{cases} \mathbb{P}_{\xi\sim P}[y \leq 0], & b_0 > 0 \\ 1, & b_0 = 0 \\ \mathbb{P}_{\xi\sim P}[y \geq 0], & b_0 < 0. \end{cases} \tag{28}$$

By comparing (26) with (28), we see that (24) holds for the case of $w_0 = 0$ too. This completes our proof that the optimal value of (22) is less than or equal to that of (19).

We now prove the converse, that the optimal value of (19) is less than or equal to that of (22). Given w and b, we show that there exists $\bar{w} = (w_0, b_0)$ such that

$$\begin{aligned}
&\epsilon\|w\|_* + \mathbb{E}_{\xi\sim P}[\max\{0, 1 - \max\{0, y(\langle w, x\rangle + b)\}\}] \\
&\geq \inf_{t>0} \{\epsilon t + \mathbb{E}_{\xi\sim P}[\max\{0, 1 - td(\bar{w}, \xi)\}]\}.
\end{aligned} \tag{29}$$

When $w \neq 0$, we take $t = \|w\|_*$, $w_0 = w/\|w\|_* = w/t$, and $b_0 = b/\|w\|_* = b/t$, and use (20) to obtain (29).

Specifically, we have

$$\begin{aligned}
&\epsilon\|w\|_* - \mathbb{E}_{\xi\sim P}[\max\{0, 1 - \max\{0, y(\langle w, x\rangle + b)\}\}] \\
&= \epsilon t - \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - \max\left\{0, \frac{ty(\langle w_0, x\rangle + b_0)}{\|w_0\|_*}\right\}\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \epsilon t - \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - t\frac{\max\{0, y(\langle w_0, x\rangle + b_0)\}}{\|w_0\|_*}\right\}\right] \\
&= \epsilon t - \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - td(\bar{w}, \xi)\right\}\right] \\
&\geq \inf_{t>0}\left\{\epsilon t - \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - td(\bar{w}, \xi)\right\}\right]\right\},
\end{aligned}
$$

as claimed.

For the case of $w = 0$, we set $b_0 = b$ and obtain

$$
\begin{aligned}
&\epsilon\|w\|_* + \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - \max\left\{0, y(\langle w, x\rangle + b)\right\}\right\}\right] \\
&= \mathbb{E}_{\xi\sim P}\left[\max\left\{0, 1 - \max\left\{0, yb\right\}\right\}\right] \\
&\geq \mathbb{P}_{\xi\sim P}\left[yb \leq 0\right] = \mathbb{P}_{\xi\sim P}\left[yb_0 \leq 0\right].
\end{aligned}
$$

By comparing with (26), we see that (29) holds in this case too. Hence, the objective value of (19) is less than or equal to that of (22).

For the final claim, we note that the optimal values of the problems (19) and (22) are equal and, from the second part of the proof above, the transformation (23) gives a solution t and $\bar{w} = (w_0, b_0)$ whose objective in (19) is at most that of (w, b) in (22). Thus, whenever (w, b) is δ-optimal for (22), then the given values of t and $\bar{w}$ are δ-optimal for (19). □

The formulation (22) can be written as the regularized risk minimization problem

$$
\inf_{w,b}\left\{\epsilon\|w\|_* + \mathbb{E}_{\xi\sim P}\left[L_R(y(\langle w, x\rangle + b))\right]\right\}, \tag{30}
$$

where L_R is the ramp loss function defined by

$$
L_R(r) := \max\left\{0, 1 - r\right\} - \max\{0, -r\} = \begin{cases} 1, & r \leq 0 \\ 1 - r, & 0 < r < 1 \\ 0, & r \geq 1. \end{cases} \tag{31}
$$

Here, the risk of a solution (w, b) is defined to be the expected ramp loss $\mathbb{E}_{\xi\sim P}\left[L_R(y(\langle w, x\rangle + b))\right]$, and the regularization term $\|w\|_*$ is defined via the norm that is dual to the one introduced in Assumption 3.1.

Remark 3.1. The formulation (22) is reminiscent of Kuhn et al. [29, Proposition 2] (see also references therein), where other distributionally robust risk minimization results were explored, except the risk was defined via the expectation of a *continuous and convex* loss function, and the reformulation was shown to be the regularized risk defined on *the same* loss function. In contrast, the risk in (4) is defined as the expectation of the *discontinuous and non-convex* 0-1 loss function $\mathbf{1}(y(\langle w, x\rangle + b) \leq 0)$, and the resulting reformulation uses the ramp loss L_R, a continuous but still nonconvex approximation of the 0-1 loss. ■

Remark 3.2. The ramp loss L_R has been studied in the context of classification by Shen et al. [43], Wu and Liu [49], and Collobert et al. [21] to find classifiers that are robust to outliers. The reformulation (30) suggests that the ramp loss together with a regularization term may have the additional benefit of also encouraging robustness to adversarial perturbations in the data. In previous work, there has been several variants of ramp loss with different slopes and break points. The formulation (32) suggests a principled form for ramp loss in classification problems. ■

In practice, the distribution P in (30) is taken to be the empirical distribution P_n on given data points $\{\xi_i\}_{i\in[n]}$, so (30) becomes

$$\inf_{w,b} \epsilon\|w\|_* + \frac{1}{n}\sum_{i\in[n]} L_R(y_i(\langle w, x_i\rangle + b)). \tag{32}$$

This problem can be formulated as a mixed-integer program (MIP) and solved to global optimality using off-the-shelf software; see [2, 10]. Despite significant advances in the computational state of the art, the scalability of MIP-based approaches with training set size m remains limited. Thus, we consider here an alternative approach based on smooth approximation of L_R and continuous optimization algorithms.

Henceforth, we consider $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$ to be the Euclidean norm. For a given ϵ in (32), there exists $\bar{\epsilon} \geq 0$ such that a strong local minimizer $(w(\epsilon), b(\epsilon))$ of (32) with $w(\epsilon) \neq 0$ is also a strong local minimizer of the following problem:

$$\min_{w,b} \frac{1}{2}\bar{\epsilon}\|w\|^2 + \frac{1}{n}\sum_{i\in[n]} L_R(y_i(\langle w, x_i\rangle + b)), \tag{33}$$

where we define $\bar{\epsilon} = \epsilon/\|w(\epsilon)\|$. In the following result, we use the notation

$$g(w,b) := \frac{1}{n}\sum_{i\in[n]} L_R(y_i(\langle w, x_i\rangle + b)),$$

for the summation term in (32) and (33).

Theorem 3.2. *Suppose that for some $\epsilon > 0$, there exists a local minimizer $(w(\epsilon), b(\epsilon))$ of (32) with $w(\epsilon) \neq 0$ and a constant $\tau > 0$ such that for all $(v, \beta) \in \mathbb{R}^d \times \mathbb{R}$ sufficiently small, we have*

$$\epsilon\|w(\epsilon)\| + g(w(\epsilon), b(\epsilon)) + \tau\|v\|^2 \leq \epsilon\|w(\epsilon) + v\| + g(w(\epsilon) + v, b(\epsilon) + \beta). \tag{34}$$

Then for $\bar{\epsilon} = \epsilon/\|w(\epsilon)\|$, $w(\epsilon)$ is also a strong local minimizer of (33), in the sense that

$$\frac{1}{2}\bar{\epsilon}\|w(\epsilon)\|^2 + g(w(\epsilon), b(\epsilon)) + \frac{\tau}{2}\|v\|^2 \leq \frac{1}{2}\bar{\epsilon}\|w(\epsilon) + v\|^2 + g(w(\epsilon) + v, b(\epsilon) + \beta),$$

for all (v, β) sufficiently small.

Proof. For simplicity of notation, we denote $(w, b) = (w(\epsilon), b(\epsilon))$ throughout the proof.

From a Taylor-series approximation of the term $\|w + v\|$, we have

$$\begin{aligned}
&\epsilon\|w\| + g(w,b) + \tau\|v\|^2 \\
&\leq \epsilon\|w + v\| + g(w + v, b + \beta) \\
&= \left[\epsilon\|w\| + \frac{\epsilon}{\|w\|}w^T v + \frac{1}{2}\frac{\epsilon}{\|w\|}v^T\left(I - \frac{ww^T}{w^T w}\right)v\right] \\
&\quad + O(\|v\|^3) + g(w + v, b + \beta) \\
&\leq \left[\epsilon\|w\| + \frac{\epsilon}{\|w\|}w^T v + \frac{1}{2}\frac{\epsilon}{\|w\|}v^T v\right] + O(\|v\|^3) + g(w + v, b + \beta) \\
&= \frac{1}{2}\epsilon\|w\| + \frac{1}{2}\frac{\epsilon}{\|w\|}(w + v)^T(w + v) + O(\|v\|^3) + g(w + v, b + \beta).
\end{aligned}$$

By rearranging this expression, and taking v small enough that the $O(\|v\|^3)$ term is dominated by $(\tau/2)\|v\|^2$, we have

$$\frac{1}{2}\epsilon\|w\| + g(w,b) + \frac{\tau}{2}\|v\|^2 \leq \frac{1}{2}\frac{\epsilon}{\|w\|}(w+v)^T(w+v) + g(w+v, b+\beta).$$

By substituting $\bar{\epsilon} = \epsilon/\|w\|$, we obtain the result. □

We note that the condition (34) is satisfied when the local minimizer satisfies a second-order sufficient condition.

To construct a smooth approximation for $L_R(r) = \max\{0, 1-r\} - \max\{0, -r\}$, we follow Beck and Teboulle [1] and approximate the two max-terms with the softmax operation: For small $\sigma > 0$,

$$\max\{a,b\} \approx \sigma \log\left(\exp\left(\frac{a}{\sigma}\right) + \exp\left(\frac{b}{\sigma}\right)\right).$$

Thus, we can approximate $L_R(r)$ by the smooth function $\psi_\sigma(r)$, parametrized by $\sigma > 0$ and defined as follows:

$$\begin{aligned}\psi_\sigma(r) &:= \sigma\log\left(1+\exp\left(\frac{1-r}{\sigma}\right)\right) - \sigma\log\left(1+\exp\left(-\frac{r}{\sigma}\right)\right)\\ &= \sigma\log\left(\frac{\exp(1/\sigma)+\exp(r/\sigma)}{1+\exp(r/\sigma)}\right).\end{aligned} \tag{35}$$

For any $r \in \mathbb{R}$, we have that $\lim_{\sigma\downarrow 0}\psi_\sigma(r) = L_R(r)$, so the approximation (35) becomes increasingly accurate as $\sigma \downarrow 0$.

By substituting the approximation ψ_σ in (35) into (33), we obtain

$$\min_{w,b}\left\{F_{\bar{\epsilon},\sigma}(w) := \frac{1}{2}\bar{\epsilon}\|w\|^2 + \frac{1}{n}\sum_{i\in[n]}\psi_\sigma(y_i(\langle w, x_i\rangle + b))\right\}. \tag{36}$$

This is a smooth nonlinear optimization problem that is nonconvex because $\psi''_\sigma(r) < 0$ for $r < 1/2$ and $\psi''_\sigma(r) > 0$ for $r > 1/2$. It can be minimized by any standard method for smooth nonconvex optimization. Stochastic gradient approaches with minibatching are best suited to cases in which n is very large. For problems of modest size, methods based on full gradient evaluations are appropriate, such as nonlinear conjugate gradient methods (see [39, Chapter 5] or L-BFGS [32]. Subsampled Newton methods (see for example [9, 53]), in which the gradient is approximated by averaging over a subset of the n terms in the summation in (36) and the Hessian is approximated over a typically smaller subset, may also be appropriate. It is well known that these methods are highly unlikely to converge to saddle points, but they may well converge to local minima of the nonconvex function that are not global minima. We show in the next section that, empirically, the global minimum is often found, even for problems involving highly nonseparable data. In fact, as proved in Section 5, under certain (strong) assumptions on the data, spurious local solutions do not exist.

4 Numerical Experiments

We report on computational tests on the linear classification problem described above, for separable and nonseparable data sets. We observe that on separable data, despite the nonconvexity of the

n	#sols	$\sin\theta(w,\bar{w})$
100	8	.155, .177, .294, .208, .224, .207, .218, .151
300	6	.184, .172, .159, .167, .180, .173
1000	2	.0735, .0821
3000	3	.0335, .0405, .0390
10000	1	.0397
30000	1	.0127

Table 1: Separable random data for n training points in $d = 10$ dimensions. Single data set for each n, 20 random starting points, $\bar{\epsilon} = .1$, $\sigma = .02$.

problem, the global minimizer of the smoothed formulation (36) is apparently found reliably by standard procedures for smooth nonlinear optimization, for sufficiently large n. Moreover, the hyperplane obtained from the hinge-loss formulation is remarkably robust to label corruption: A solution close to the separating hyperplane is frequently identified even when a large fraction of the labels from the separable data set are flipped randomly to incorrect values, and even when many adversarial data points are added.

Our results are intended to be "proof of concept" in that they both motivate and support our analysis in Section 5 that the minimizer of the regularized risk minimization problem (30) is the only point satisfying even first-order conditions. In fact, our computational tests go beyond the case considered in the next section by (a) increasing the dimension beyond $d = 2$, and (b) considering nonseparable data sets. Further analysis is required to prove that the global minimizer can be found reliably in these contexts too.

We report on computations with the formulation (36) with $\sigma = .02$ and various values of $\bar{\epsilon}$ and n. (The results are not sensitive to the choice of σ, except that smaller values yield functions that are less smooth and thus require more iterations to minimize.) We use dimension $d = 10$ throughout. The data points x_i were generated uniformly at random in the box $[-10, 10]^d$, and we set $y_i = \text{sign}((x_i)_1)$ for all i.

We tried various smooth unconstrained optimization solvers for the resulting smooth nonconvex optimization problem — the PR+ version of nonlinear conjugate gradient [39, Chapter 5], the L-BFGS method [32], and Newton's method with diagonal damping — all in conjunction with a line-search procedure that ensures weak Wolfe conditions. These methods behaved in a roughly similar manner and all were effective in finding a minimizer. Our tables report results obtained only with nonlinear conjugate gradient.

For all tests, we solved (36) for the hyperplane (w, b), starting from a random point on the unit ball in $\mathbb{R}^{d+1}$. A statistic of particular interest is the closeness of the solution obtained from the local minimization procedure to the "canonical" separating hyperplane $\bar{w} = (1, 0, 0 \ldots, 0)^T$ and intercept $b = 0$. We report the sine of the angle between $\bar{w}$ and each calculated w; smaller values correspond to more closely oriented hyperplanes.

Table 1 shows results for the separable data set described above, for various training set sizes n, with $\bar{\epsilon} = 0.1$. For each value of n, we generate a single random data set and solve the ramp-loss problem from 20 random starting points. We tabulate the number of local minima found during these 20 trials, showing for each local minimizer the angle with the canonical supporting hyperplane $\bar{w}$. We note that $\bar{w}$ becomes closer to optimal as the size of the training set is increased, because the the training points become more closely packed around the separating hyperplane for larger n. Note that as n grows larger, the empirical problem (36) appears to have a single global minimizer;

$\bar{\epsilon}$	$\|w\|$	imputed ϵ	$\sin\theta(w, \bar{w})$
.001	3.744, 3.768	.003744, .003768	.0119, .0127
.01	1.744	.01744	.0194
.1	.7878	.07878	.0235
1	.3670	.3670	.0472
10	.1703	1.703	.0522

Table 2: Separable random data for 10000 training points in $d = 10$ dimensions and varying values of $\bar{\epsilon}$. Single data set for each n, 20 random starting points, $\sigma = .02$. The same minimizer (presumably the global minimizer) was reached from all starting points, except for $\bar{\epsilon} = .001$, where two minimizers were found.

%flipped	av. # solutions	av. $\sin\theta(w, \bar{w})$	av. $\sin\theta(w_{\text{hinge}}, \bar{w})$
10	1	.0347	.0693
20	1.5	.0476	.0987
30	1.4	.0589	.1220
40	3.9	.1164	.1917

Table 3: Nonseparable random data obtained for $n = 10000$, $\bar{\epsilon} = 0.1$, $\sigma = .02$ in $d = 10$ dimensions, modifying the original separable data set by flipping a specified percentage of labels randomly in the separable data set. Results averaged over 10 data sets in each row, with 20 random starting points for the hinge loss.

in this respect its behavior appears to match that of the continuous problem that is analyzed in the next section. We note that for all these problems, the summation term in (36) involving smoothed hinge losses is nonzero, because there is no clear margin between the two classes. Regardless of the orientation and scaling of (w, b), some points will always be close to the hyperplane and thus incur nonzero values of ψ_σ.

Table 2 has a similar setup to Table 1, except that here we fix the training set size n at 10000 and survey the effects of different values of $\bar{\epsilon}$. Two (similar) local minimizers were found for $\bar{\epsilon} = .001$, while for larger values of $\bar{\epsilon}$ a single (presumably global) minimizer was found. We tabulate the norm of the solution w and the value of ϵ that is “imputed” by the relationship $\epsilon = \bar{\epsilon}\|w\|$ that we mention in the previous section when describing the relationship between the formulations (32) and (33). Note that there is a slight drift of the orientation of the optimal w away from the canonical separating hyperplane $\bar{w}$ as $\bar{\epsilon}$ increases. We believe that this can be explained in terms of the original DRO motivation. As ϵ increases, we are maximizing the loss over an increasingly large ball of distributions in (4), centered on the empirical distribution P_n. Thus, for any fixed w, more points are adversarially perturbed to the decision boundary. Correspondingly, the proportion of points that are unperturbed, and hence contribute to correct classification with respect to P_n, decreases, hence the objective is more likely to shift away from the canonical hyperplane.

In Table 3, we show results obtained with nonseparable data sets that are obtained by constructing the separable data set as explained above, then flipping the labels on a randomly chosen fraction of training points. In this table, we fix $n = 10000$ and $\bar{\epsilon} = 0.1$, generate 10 data sets for each choice of number of labels flipped (10%, 20%, 30%, and 40%), and run the ramp loss minimization from 20 starting points for each data set. We tabulate the average number of local minima detected over the 10 runs. (For example, with 20% of labels flipped, eight of the data sets yielded just one minimizer from the 20 starting points, one data set yielded two minimizers, and one data

ϕ (%adv)	$\sin\theta(w,\bar{w})$	intercepts	misclass	$\sin\theta(w_{\text{hinge}},\bar{w})$	intercept	misclass
10	.0509	.0922	1111	.0364	.2682	1560
20	.0466	.0426	2072	.0318	.359	2903
30	.0401	-.0028	3067	.3425	1.00	3537

Table 4: Each row corresponds to a random data set with $n = 10000$ in which a certain percentage ϕ of the points (chosen randomly) are modified by setting $x_{i,1} = -10$ and $y_i = 1$ to produce adversarial examples. (Same minimizer is identified in 20 trials from different starting points for each instance of smoothed ramp loss.)

set yielded five local minimizers. The overall average was 1.5 local minimizers.) We also compute solutions on the same data sets solved with (36) in which ψ_σ is replaced with a version of the hinge-loss function $h(r) = \max(1-r, 0)$, smoothed in the same way as ψ_σ to allow solvers for smooth unconstrained optimization to be applied. Since this problem is convex, there is a unique optimal objective value, found from any starting point. Comparing the solutions for ramp loss and hinge loss, we note that the orientation of the hyperplane is distorted a little further from its canonical value by the hinge-loss solution than the ramp-loss solution, suggesting that the "outliers" created by label flipping are disrupting the hinge-loss solution slightly more than the ramp-loss solution.

Finally, we tested on examples in which the random data set described above is modified with numerous adversarial examples. Starting with the usual random data set of size $n = 10000$, we modify a randomly chosen fraction ϕ of the points by resetting their first component to -10 and their label to 1. (The remaining $(1-\phi)n$ points from the original data set have approximately equal representation from each of the two classes.) All ϕn of the adversarial points are mislabelled and far on the wrong side of the "canonical" decision boundary defined by $w = \bar{w}$ and $b = 0$. When the separating hyperplane is not too far from this canonical value, the smoothed ramp-loss function assigns a loss of approximately 1 to all these points. The hinge-loss function, however, incurs a much higher loss for these points, so the hyperplane that minimizes the hinge loss is likely to be significantly distorted, with the intercept shifted significantly, resulting in misclassification of many (non-adversarial) points that were close to the canonical decision boundary.

Results are shown in Table 4. We generated a single data set for each value $\phi = .1, .2, .3$ of the fraction of adversarial points. We started the smoothed ramp-loss objective from 20 points and found that it converged to a single point in all three instances. For the ramp-loss instances, the hyperplane obtained is close to the canonical hyperplane, and few points are misclassified apart from the adversarial points (which are "correctly" misclassified). For the hinge-loss objective, the intercept b is shifted more and more as ϕ increases, and for the case of $\phi = .3$, the orientation of the hyperplane is also shifted significantly. These changes results in many points being misclassified in addition to the adversarial points.

5 Benign Nonconvexity of Ramp Loss on Linearly Separable Data

We consider (33), setting $b = 0$ for simplicity to obtain

$$\min_{w} \left\{ F_\epsilon(w) := \frac{1}{2}\epsilon\|w\|_2^2 + \mathbb{E}_{(x,y)\sim P}\left[L_R(y\langle w, x\rangle)\right] \right\}. \tag{37}$$

We also consider the smoothed approximation to this problem in which L_R is replaced by ψ_σ defined in (35). In this section, we explore the question: is the nonconvex problem (37) *benign*, in the sense that, for reasonable data sets, descent algorithms for smooth nonlinear optimization will find the global minimum? In the formulation (37), we make use of the true distribution P rather than its empirical approximation P_n, because results obtained for P will carry through to P_n for large n, with high probability. Exploring this question for general data distributions is difficult, so we consider the following simplified problem, similar to those studied in Section 4, but with $d = 2$.

Assumption 5.1. The data (x, y) comprising P has x uniformly distributed on $[-1, 1]^2$, and $y = \text{sign}(\langle w^*, x\rangle)$ where $w^* = (1, 0)$.

Under this assumption, we show that F_ϵ defined in (37) has a single local minimizer of the form $w(\epsilon) = (w_1(\epsilon), 0)$ for some $w_1(\epsilon) > 0$. Since the function is also bounded below (by zero) and coercive, this local minimizer is the global minimizer. We conjecture that the smoothed version of (37) in which L_R is replaced by ψ_σ (35) has similar properties, so that any gradient-based descent method applied to the latter problem will converge to the global minimizer, which will be close to $w(\epsilon)$.

We assume $d = 2$ in Assumption 5.1 for tractability of our analysis. While we believe the results below are also true for general d, we do not have rigorous proofs for that case.

Since $d = 2$ we write $w = (w_1, w_2)$ and $x = (x_1, x_2)$. It follows from Assumption 5.1 that $x_1, x_2 \sim U(-1, 1)$ independently. Since $y = \text{sign}(x_1)$, we can write $x_1 = yr$ where $r \sim U(0, 1)$ and y is a Rademacher random variable, that is, $\mathbb{P}[y = 1] = \mathbb{P}[y = -1] = 1/2$ independent of r, x_2. Since $x_2 \sim U(-1, 1)$ is symmetric, we have $x_2 \overset{d}{=} yx_2$. Therefore, $y\langle w, x\rangle = yw_1x_1 + yw_2x_2 \overset{d}{=} w_1r + w_2x_2$. Henceforth, without loss of generality, we can ignore y and replace x_1 with r, so that

$$F_\epsilon(w) = \frac{1}{2}\epsilon\|w\|_2^2 + \mathbb{E}_{(r\sim U(0,1),\, x_2\sim U(-1,1))}\left[L_R(w_1r + w_2x_2)\right]. \tag{38}$$

In the remainder of the paper, we use $\mathbb{E}$ to denote $\mathbb{E}_{(r\sim U(0,1),\, x_2\sim U(-1,1))}$, unless explicitly indicated otherwise.

We now investigate differentiability properties of the objective F_ϵ.

Lemma 5.1. *When $w \neq (0, 0)$, the function $F_\epsilon(w)$ is differentiable in w with gradient*

$$\nabla F_\epsilon(w) = \epsilon w - \mathbb{E}\left[\mathbf{1}\left(0 < w_1r + w_2x_2 < 1\right)\begin{pmatrix} r \\ x_2 \end{pmatrix}\right].$$

At $w = (0, 0)$, the directional derivatives of F_ϵ in the directions $(1, 0)$ and $(-1, 0)$ are $-1/2$ and 0, respectively.

Proof. We appeal to Clarke [19, Theorem 2.7.2] which shows how to compute the generalized gradient of a function defined via expectations. We note that for every r, x_2, $w \mapsto L_R(w_1r + w_2x_2)$ is a regular function since it is a difference of two convex functions, and is differentiable everywhere except when $w_1r + w_2x_2 \in \{0, 1\}$, with gradient $\mathbf{1}(0 < w_1r + w_2x_2 < 1)(r, x_2)$. When $w \neq (0, 0)$, the set of $(r, x_2) \sim P$ such that $w_1r + w_2x_2 \in \{0, 1\}$ is a measure-zero set under the distribution on r, x_2, so Clarke [19, Theorem 2.7.2] states that the generalized gradient of $\mathbb{E}[L_R(w_1r + w_2x_2)]$ is the singleton set $\{-(\mathbb{E}\left[\mathbf{1}\left(0 < w_1r + w_2x_2 < 1\right)r\right], \mathbb{E}\left[\mathbf{1}\left(0 < w_1r + w_2x_2 < 1\right)x_2\right])\}$. As it is a singleton, this coincides with the gradient at w. This proves the first claim.

For the final claims, note first that the gradient of the regularization term $\frac{1}{2}\epsilon\|w\|^2$ is zero at $w = (0, 0)$. Thus we need consider only the L_R term in applying the definition of directional derivative to (37). For the direction $(1, 0)$, we have

$$F'_\epsilon((0, 0); (1, 0)) = \lim_{\alpha\downarrow 0}\frac{1}{\alpha}\left[\mathbb{E}_{(x,y)\sim P}[L_R(y(\alpha x_1))] - \mathbb{E}_{(x,y)\sim P}[L_R(0)]\right]$$

$$
\begin{aligned}
&= \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[\mathbb{E}_{(r,x_2)}[L_R(\alpha r)] - 1 \right] \\
&= \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[\mathbb{E}_{(r,x_2)}[(1 - \alpha r)] - 1 \right] \\
&= -\mathbb{E}_{(r,x_2)}[r] = -\frac{1}{2},
\end{aligned}
$$

as claimed. For the direction $(-1, 0)$, we have

$$
F_\epsilon((0,0) + \alpha(-1,0)) = \mathbb{E}\left[L_R(-\alpha r)\right] = \mathbb{E}\left[L_R(0)\right] = -1, \quad \text{for } \alpha > 0,
$$

proving the claim. □

Lemma 5.1 shows that $w = (0,0)$ is not a local minimum of F_ϵ, hence any reasonable descent algorithm will not converge to it. We now investigate stationary points $\nabla F_\epsilon(w) = 0$ for $w \neq (0,0)$, for which the stationarity condition can be written as follows:

$$
\epsilon w_1 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] \tag{39a}
$$

$$
\epsilon w_2 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right]. \tag{39b}
$$

The following results use these expressions to eliminate most w from being stationary points.

Proposition 5.2. *Under Assumption 5.1, the conditions (39) cannot be satisfied by any (w_1, w_2) with $w_1 < 0$, or by any w of the form $(0, w_2)$ with $w_2 \neq 0$.*

Proof. Suppose that $w_1 < 0$. Then since $\mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] \geq 0$, there is inconsistency between the two sides in the equation (39a), proving the first claim.

Considering now w with $w_1 = 0$ and $w_2 \neq 0$, we have

$$
\mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] = \mathbb{P}[0 < w_2 x_2 < 1]\mathbb{E}[r] > 0,
$$

so (39a) cannot be satisfied in this case either, proving the second claim. □

By inspection of (39), we see that ϵw_1 is obtained as the expectation of the nonnegative random variable $r \sim U(0,1)$ over a certain domain, whereas ϵw_2 is the expectation of a random variable $x_2 \sim U(-1,1)$ symmetrically distributed about 0 on the *same domain.* Hence, we would expect at least that $|w_2| \ll w_1 > 0$. In fact, we show in the next result, whose highly technical proof is deferred to Appendix A, that no w with $w_2 \neq 0$ can be stationary for F_ϵ.

Proposition 5.3. *Under Assumption 5.1, if $w = (w_1, w_2)$ has $w_2 \neq 0$, then it cannot satisfy (39).*

Propositions 5.2 and 5.3 show that the only possible stationary points for F_ϵ in (37) have either $w = (w_1, 0)$ for some $w_1 > 0$, or $w = (0,0)$. We have shown already that there is a descent direction from $(0,0)$, so the only possible local minima are the points with $w = (w_1, 0)$ and $w_1 > 0$. For a given value of ϵ, we obtain these stationary points by solving (39).

When $w_1 > 0$, $w_2 = 0$, (39b) is clearly satisfied, since

$$
0 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r < 1\right) x_2\right] = \mathbb{P}[0 < w_1 r < 1]\mathbb{E}[x_2] = 0.
$$

For (39a), we have

$$
\epsilon w_1 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r < 1\right) r\right] = \int_0^{\min\{1, 1/w_1\}} r dr,
$$

For $w_1 \geq 1$, we have

$$\epsilon w_1 = \int_0^{\min\{1,1/w_1\}} r dr = \frac{1}{2w_1^2} \implies w_1 = \sqrt[3]{\frac{1}{2\epsilon}}, \tag{40}$$

which is consistent with $w_1 \geq 1$ only when $\epsilon \leq 1/2$. For $w_1 < 1$, we have

$$\epsilon w_1 = \int_0^{\min\{1,1/w_1\}} r dr = \frac{1}{2} \implies w_1 = \frac{1}{2\epsilon}, \tag{41}$$

which is consistent with $w_1 < 1$ only when $\epsilon > 1/2$. Note that when $w_1 > 0$ and $w_2 = 0$, we have $y(w_1 x_1 + w_2 x_2) = w_1 r \geq 0$, so

$$\mathbb{E}[L_R(y(w_1 x_1 + w_2 x_2))] = \int_{r=0}^{r=\min\{1,1/w_1\}} (1 - w_1 r) dr = \begin{cases} 1 - \frac{w_1}{2}, & w_1 < 1 \\ \frac{1}{2w_1}, & w_1 \geq 1. \end{cases}$$

Therefore, letting $w(\epsilon)$ be the global minimizer of F_ϵ defined by (40) and (41), and by substituting this value into (37), we obtain

$$w(\epsilon) = (w_1(\epsilon), w_2(\epsilon)) = \begin{cases} \left(\frac{1}{\sqrt[3]{2\epsilon}}, 0\right) & \epsilon \leq 1/2 \\ \left(\frac{1}{2\epsilon}, 0\right) & \epsilon > 1/2, \end{cases} \qquad F_\epsilon(w(\epsilon)) = \begin{cases} 3\sqrt[3]{\frac{\epsilon}{32}} & \epsilon \leq 1/2 \\ 1 - \frac{1}{8\epsilon} & \epsilon > 1/2. \end{cases}$$

For the smoothed counterpart of (37), which is

$$\min_w \left\{ F_{\epsilon,\sigma}(w) := \frac{1}{2}\epsilon \|w\|_2^2 + \mathbb{E}_{(x,y)\sim P}\left[\psi_\sigma(y\langle w, x\rangle)\right] \right\}, \tag{42}$$

we can show that the origin $w = (0,0)$ is not a stationary point. Since $\psi'_\sigma(0) = \frac{1-\exp(1/\sigma)}{2(\exp(1/\sigma)+1)}$, we have

$$\begin{aligned} \frac{\partial}{\partial w_1} \mathbb{E}_{(x,y)\sim P}\left[\psi_\sigma(y\langle w, x\rangle)\right]\Big|_{w=(0,0)} &= \frac{1-\exp(1/\sigma)}{2(\exp(1/\sigma)+1)} \mathbb{E}_{(x,y)\sim P}[\, y x_1] \\ &= \frac{1-\exp(1/\sigma)}{2(\exp(1/\sigma)+1)} \mathbb{E}_{r\sim U(0,1)}[\, r] \\ &= \frac{1-\exp(1/\sigma)}{4(\exp(1/\sigma)+1)} < 0 \end{aligned}$$

for any $\sigma > 0$. In fact, the limit of this quantity as $\sigma \downarrow 0$ is $-1/4$, which is the average of the directional derivatives of F_ϵ at $w = (0,0)$ in the directions $(-1, 0$ and $(1, 0)$ (see Lemma 5.1), as expected.

Label Flipping. Our experiments in Section 4 showed that solutions of the problems analyzed in this section showed remarkable resilience to "flipping" of the labels y on a number of samples. To give some informal insight into this phenomenon, we write the smoothed objective (42) and its gradient with distribution P replaced by a discrete distribution on $(x_i, y_i) \in \mathbb{R}^2 \times \{\pm 1\}$, $i \in [n]$, as follows:

$$F_{\epsilon,\sigma}(w) = \frac{1}{2}\epsilon \|w\|_2^2 + \frac{1}{n}\sum_{i\in[n]} \psi_\sigma(y_i\langle w, x_i\rangle), \tag{43a}$$

$$\nabla F_{\epsilon,\sigma}(w) = \epsilon w + \frac{1}{n}\sum_{i\in[n]} \psi'_\sigma(y_i\langle w, x_i\rangle) y_i x_i. \tag{43b}$$

Note that $\nabla F_{\epsilon,\sigma}(w) = 0$ when

$$w = -\frac{1}{n\epsilon} \sum_{i \in [n]} \psi'_\sigma(y_i \langle w, x_i \rangle) y_i x_i. \tag{44}$$

We note that $\psi'_\sigma(y_i \langle w, x_i \rangle) < 0$ for all i (since in fact $\psi'_\sigma(t) < 0$ for all t). Thus since $y_i = \mathrm{sign}(x_{i,1})$, the first components of all terms in the summation in (44) are negative. For the second components, the negative coefficients $\psi'_\sigma(y_i \langle w, x_i \rangle)$ are the same as for the first components, but about half of the terms $y_i x_{i,2}$ are negative and half positive. Thus, even by this informal analysis, we can see that w satisfying (44) will have $w_1 \gg |w_2|$.

Considering now the case in which some fraction $\phi \in [0, 1/2)$ of labels y_i, chosen randomly, are flipped. We can again compare first and second components of the right-hand side of (44) to note that in the first component, a fraction of $(1 - \phi)$ of terms are negative with the remainder positive, while in the second component we still have that approximately half the terms are negative and half positive. The negative coefficients $\psi'_\sigma(y_i \langle w, x_i \rangle)$ are of course the same for both components, so the imbalance in the first component suggests again that $w_1 \geq |w_2|$ when ϕ is not too close to 1/2. This informal analysis extends immediately to the case of $d > 2$, explaining to a large extent the results noted in Table 3.

Acknowledgements

This work was supported by NSF Awards 1628384, 1634597, 1740707, and 2023239; Subcontract 8F-30039 from Argonne National Laboratory; and Award N660011824020 from the DARPA Lagrange Program. The research was conducted in part while the first author was a postdoctoral research associate at the University of Wisconsin–Madison.

References

[1] A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012. doi: 10.1137/100818327.

[2] P. Belotti, P. Bonami, M. Fischetti, A. Lodi, M. Monaci, A. Nogales-Gómez, and D. Salvagnin. On handling indicator constraints in mixed integer programming. *Computational Optimization and Applications*, 65(3):545–566, 2016. doi: 10.1007/s10589-016-9847-8.

[3] A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007. doi: 10.1111/j.1467-9965.2007.00311.x.

[4] A. Ben-Tal, E. Hazan, T. Koren, and S. Mannor. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, 2015. doi: 10.1287/opre.2015.1374.

[5] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, 1992. doi: 10.1080/10556789208805504.

[6] D. Bertsimas and M. S. Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931 – 942, 2018. ISSN 0377-2217. doi: 10.1016/j.ejor.2017.03.051.

[7] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019. doi: 10.1287/ijoo.2018.0001.

[8] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936.

[9] R. Bollapragada, R. H. Byrd, and J. Nocedal. Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019.

[10] J. P. Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2):467–479, 2011. doi: 10.1287/opre.1100.0854.

[11] S. Bubeck, Y. T. Lee, E. Price, and I. Razenshteyn. Adversarial examples from computational constraints. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 831–840, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/bubeck19a.html.

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

[13] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 3–14, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352024. doi: 10.1145/3128572.3140444. URL https://doi.org/10.1145/3128572.3140444.

[14] Z. Charles, S. Rajput, S. Wright, and D. Papailiopoulos. Convergence and Margin of Adversarial Training on Separable Data. Technical report, May 2019. URL https://arxiv.org/abs/1905.09209.

[15] J. Chen, X. Wu, V. Rastogi, Y. Liang, and S. Jha. Robust attribution regularization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14300–14310. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9577-robust-attribution-regularization.pdf.

[16] R. Chen and I. C. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018. URL http://jmlr.org/papers/v19/17-295.html.

[17] Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over Wasserstein balls. *arXiv e-prints*, art. arXiv:1809.00210, Sep 2018.

[18] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[19] F. H. Clarke. *Optimization and Nonsmooth Analysis.* Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611971309.

[20] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning*

Research, pages 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/cohen19c.html.

[21] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 201–208, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143870.

[22] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NeurIPS'18, pages 1186–1195, USA, 2018. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=3326943.3327052.

[23] A. Fawzi, O. Fawzi, and P. Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, Mar 2018. ISSN 1573-0565. doi: 10.1007/s10994-017-5663-3.

[24] J. Gilmer, N. Ford, N. Carlini, and E. Cubuk. Adversarial examples are a natural consequence of test error in noise. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2280–2289, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/gilmer19a.html.

[25] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

[26] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9461–9471. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8156-implicit-bias-of-gradient-descent-on-linear-convolutional-networks.pdf.

[27] N. Ho-Nguyen and F. Kılınç-Karzan. Online first-order framework for robust convex optimization. *Operations Research*, 66(6):1670–1692, 2018. doi: 10.1287/opre.2018.1764.

[28] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/hu18a.html.

[29] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, chapter 6, pages 130–166. 2019. doi: 10.1287/educ.2019.0198.

[30] J. Lee and M. Raginsky. Minimax statistical learning with Wasserstein distances. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2687–2696. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7534-minimax-statistical-learning-with-wasserstein-distances.pdf.

[31] Y. Li, E. X.Fang, H. Xu, and T. Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkgTTh4FDH.

[32] D. C. Liu and J. Nocedal. On the limited-memory BFGS method for large scale optimization. 45:503–528, 1989.

[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[34] O. L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13(3):444–452, 1965. doi: 10.1287/opre.13.3.444.

[35] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sep 2018. ISSN 1436-4646. doi: 10.1007/s10107-017-1172-1.

[36] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[37] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard. Robustness via curvature regularization, and vice versa. Technical report, Nov 2018. URL https://arxiv.org/abs/1811.09716.

[38] A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*, 24(3):381–406, 2009. doi: 10.1080/10556780802712889.

[39] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.

[40] R. T. Rockafellar. *Coherent Approaches to Risk in Optimization Under Uncertainty*, chapter 3, pages 38–61. 2007. doi: 10.1287/educ.1073.0032.

[41] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1576–1584, Cambridge, MA, USA, 2015. MIT Press.

[42] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019. URL http://jmlr.org/papers/v20/17-633.html.

[43] X. Shen, G. C. Tseng, X. Zhang, and W. H. Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003. doi: 10.1198/016214503000000639.

[44] A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

[45] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

[46] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

[47] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6586–6595, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/wang19i.html.

[48] E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5286–5295, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[49] Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007. doi: 10.1198/016214507000000617.

[50] W. Xie. On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming*, 2019. doi: 10.1007/s10107-019-01445-5. Article in advance.

[51] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(51):1485–1510, 2009. URL http://jmlr.org/papers/v10/xu09b.html.

[52] H. Xu, C. Caramanis, and S. Mannor. Robust optimization in machine learning. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*. The MIT Press, 2011. ISBN 026201646X.

[53] P. Xu, J. Yang, F. Roosta, C. Ré, and M. W. Mahoney. Sub-sampled Newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems*, pages 3000–3008, 2016.

[54] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7085–7094, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/yin19b.html.

[55] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zhang19p.html.

A Proof of Proposition 5.3

We first give a slightly simplified expression for (39). Since $x_2 \sim U(-1,1)$ is a symmetric random variable, we know that $x_2 \overset{d}{=} -x_2 \overset{d}{=} \operatorname{sign}(w_2)x_2$. Therefore we can simplify the gradient condition (39) as follows:

$$\begin{aligned}
\epsilon w_1 &= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] \\
&= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + |w_2| \operatorname{sign}(w_2) x_2 < 1\right) r\right] \\
&= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + |w_2| x_2 < 1\right) r\right] \\
\epsilon w_2 &= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \\
&= \operatorname{sign}(w_2)\mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + |w_2| \operatorname{sign}(w_2) x_2 < 1\right) \operatorname{sign}(w_2) x_2\right] \\
&= \operatorname{sign}(w_2)\mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + |w_2| x_2 < 1\right) x_2\right].
\end{aligned}$$

It follows that (39) is equivalent to

$$\begin{aligned}
\epsilon w_1 &= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + |w_2| x_2 < 1\right) r\right] \\
\epsilon |w_2| &= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + |w_2| x_2 < 1\right) x_2\right].
\end{aligned}$$

Note that if (w_1, w_2) satisfies these conditions, so will $(w_1, -w_2)$. Thus, without loss of generality, in searching for stationary points for F_ϵ, we can assume that (39) holds for $w_2 \geq 0$, that is,

$$\epsilon w_1 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] \tag{45a}$$

$$\epsilon w_2 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right]. \tag{45b}$$

In the proof below, we have made extensive use of Wolfram Alpha[2] to compute and simplify various integral, polynomial, and rational expressions.

Proof of Proposition 5.3. We showed in Proposition 5.2 that if $w_2 \neq 0$, then $w_1 \leq 0$ cannot yield a stationary point. Thus, for this proof, we need show only that there is no (w_1, w_2) satisfying (45) with $w \in \mathbb{R}^2_{++} = \{(w_1, w_2) : w_1 > 0, w_2 > 0\}$. Our proof works with by partitioning $\mathbb{R}^2_{++}$ into five subsets, finding in each case a contradiction between (45) and the conditions that define the subset. (Each of the five cases is numbered below.)

We make extensive use of the function g defined by

$$g(r) := \int_{x_2=-w_1 r/w_2}^{(1-w_1 r)/w_2} x_2 d\mathbb{P}(x_2) = \frac{1}{2}\int_{x_2=\max\{-1, -w_1 r/w_2\}}^{\min\{1,(1-w_1 r)/w_2\}} x_2 dx_2. \tag{46}$$

We have from (45b) that

$$\begin{aligned}
\epsilon w_2 &= \mathbb{E}_{r\sim U(0,1),\, x_2 \sim U(-1,1)}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \\
&= \int_{r=0}^{r=1/w_1} g(r) d\mathbb{P}(r) + \int_{r=1/w_1}^{\infty} g(r) d\mathbb{P}(r).
\end{aligned} \tag{47}$$

When $r \geq 1/w_1$, the upper bound $(1-w_1 r)/w_2$ of the integrand defining $g(r)$ is nonpositive, so $g(r) \leq 0$, and we have

$$\epsilon w_2 = \mathbb{E}_{r\sim U(0,1),\, x_2 \sim U(-1,1)}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \leq \int_{r=0}^{r=1/w_1} g(r) d\mathbb{P}(r). \tag{48}$$

[2] wolframalpha.com

We claim that g is anti-symmetric about the point $1/(2w_1)$, that is, if $r - 1/(2w_1) = 1/(2w_1) - r'$, then $g(r) = -g(r')$. To see this, note that $r = 1/w_1 - r'$, so that

$$g(r) = g(1/w_1 - r') = \int_{x_2=-(1-w_1r')/w_2}^{w_1r'/w_2} x_2 d\mathbb{P}(x_2) = -\int_{x_2=-w_1r'/w_2}^{(1-w_1r')/w_2} x_2 d\mathbb{P}(x_2) = -g(r'),$$

where the second equality follows from symmetry of x_2.

1. Consider $w_1 \geq 1$, $w_2 > 0$. Since $r \sim U(0,1)$, we have from the antisymmetry property, (48), and (45b) that

$$\begin{aligned} &\int_{r=0}^{r=1/w_1} g(r) d\mathbb{P}(r) = 0 \\ &\implies \epsilon w_2 = \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)} \left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \leq 0, \end{aligned}$$

so $w_2 \leq 0$, which contradicts our assumption that $w_2 > 0$. Thus, there can be no stationary point with $w_1 \geq 1$ and $w_2 > 0$.

We now verify cases when $0 < w_1 < 1$, for which we have $1/w_1 > 1$. Since $r \sim U(0,1)$, we have $\int_{r=1/w_1}^{\infty} g(r) d\mathbb{P}(r) = 0$. Thus from (47), we have

$$\begin{aligned} \epsilon w_2 &= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)} \left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \\ &= \int_{r=0}^{1/w_1} g(r) d\mathbb{P} r = \int_{r=0}^{1} g(r) dr \geq 0. \end{aligned} \tag{49}$$

2. Consider now $0 < w_1 < 1$ and $w_2 \geq 1$. We have for $r \sim U(0,1)$ that $-1 < -w_1 r/w_2 \leq (1 - w_1 r)/w_2 \leq 1$, and since $x_2 \sim U(-1,1)$, we have

$$\begin{aligned} g(r) = \int_{x_2=-w_1r/w_2}^{(1-w_1r)/w_2} x_2 d\mathbb{P}(x_2) &= \frac{1}{2}\int_{x_2=-w_1r/w_2}^{(1-w_1r)/w_2} x_2 dx_2 \\ &= \frac{1}{4w_2^2}\left((1-w_1r)^2 - (w_1r)^2\right) = \frac{1-2w_1r}{4w_2^2}, \end{aligned}$$

so that from (47) we have in this case that

$$\begin{aligned} \epsilon w_2 &= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)} [\mathbf{1}(0 < w_1 r + w_2 x_2 < 1) x_2] \\ &= \int_{r=0}^{1} g(r) dr = \int_{r=0}^{1} \frac{1-2w_1r}{4w_2^2} dr \\ &= \frac{1-w_1}{4w_2^2}. \end{aligned}$$

We also have from (45a) that

$$\begin{aligned} \epsilon w_1 = \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] &= \mathbb{E}\left[\mathbf{1}\left(-w_1 r/w_2 < x_2 < (1-w_1r)/w_2\right) r\right] \\ &= \frac{1}{2w_2}\mathbb{E}\left[r\right] = \frac{1}{4w_2}. \end{aligned}$$

By combining these last two expressions, we have

$$\frac{w_2}{w_1} = \frac{\epsilon w_2}{\epsilon w_1} = \frac{(1-w_1)/(4w_2^2)}{1/(4w_2)} = \frac{1-w_1}{w_2} < 1,$$

so that $w_2 < w_1$, which contradicts the conditions that define this subset.

3. Consider now $0 < w_1 < 1$, $w_2 > 0$ and $w_1 + w_2 < 1$. We have $w_1 r + w_2 x_2 < 1$ for all $r \in [0,1]$ and $x_2 \in [-1,1]$, so

$$\begin{aligned}
\epsilon w_2 &= \mathbb{E}_{r\sim U(0,1),\, x_2 \sim U(-1,1)} \left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \\
&= \mathbb{E}_{r\sim U(0,1),\, x_2 \sim U(-1,1)} \left[\mathbf{1}\left(0 < w_1 r + w_2 x_2\right) x_2\right] \\
&= \mathbb{E}_{r\sim U(0,1),\, x_2 \sim U(-1,1)} \left[\mathbf{1}\left(-w_1 r/w_2 < x_2\right) x_2\right] \\
&= \int_{r=0}^{\min\{w_2/w_1,1\}} \left[\frac{1}{2}\int_{x_2=-w_1 r/w_2}^{1} x_2 dx_2\right] d\mathbb{P}(r) \\
&\quad + \int_{r=\min\{w_2/w_1,1\}}^{1} \left[\frac{1}{2}\int_{x_2=-1}^{1} x_2 dx_2\right] d\mathbb{P}(r) \\
&= \int_{r=0}^{\min\{w_2/w_1,1\}} \left[\frac{1}{2}\int_{x_2=-w_1 r/w_2}^{1} x_2 dx_2\right] d\mathbb{P}(r) \\
&= \int_{r=0}^{\min\{w_2/w_1,1\}} \left(\frac{1}{4} - \frac{r^2 w_1^2}{4 w_2^2}\right) dr \\
&= \begin{cases} \frac{w_2}{6w_1}, & w_2 \le w_1 \\ \frac{1}{4} - \frac{w_1^2}{12 w_2^2}, & w_2 > w_1. \end{cases}
\end{aligned}$$

Also, we have

$$\begin{aligned}
\epsilon w_1 &= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] \\
&= \mathbb{E}\left[\mathbf{1}\left(0 < w_1 r + w_2 x_2\right) r\right] \\
&= \mathbb{E}\left[\mathbf{1}\left(-w_1 r/w_2 < x_2 < 1\right) r\right] \\
&= \frac{1}{2}\int_{r=0}^{\min\{w_2/w_1,1\}} (1 + w_1 r/w_2) r dr + \int_{r=\min\{w_2/w_1,1\}}^{1} r dr \\
&= \begin{cases} \frac{5w_2^2}{12w_1^2} + \frac{1}{2} - \frac{w_2^2}{2w_1^2} & w_2 \le w_1 \\ \frac{w_1}{6w_2} + \frac{1}{4} & w_2 > w_1 \end{cases} \\
&= \begin{cases} \frac{1}{2} - \frac{w_2^2}{12w_1^2} & w_2 \le w_1 \\ \frac{w_1}{6w_2} + \frac{1}{4} & w_2 > w_1 \end{cases}
\end{aligned}$$

We consider two subcases.

(a) When $w_2 \le w_1$, the first order conditions are

$$\epsilon w_1 = \frac{1}{2} - \frac{w_2^2}{12 w_1^2}, \quad \epsilon w_2 = \frac{w_2}{6 w_1}.$$

Since $w_2 > 0$, we solve the second equation to obtain $w_1 = 1/(6\epsilon)$. By substituting into the first equation, we obtain

$$\epsilon w_1 = \frac{1}{6} = \frac{1}{2} - \frac{w_2^2}{12 w_1^2} \implies \frac{w_2^2}{12 w_1^2} = \frac{1}{3} \implies \frac{w_2^2}{w_1^2} = 4,$$

but the final inequality is incompatible with $0 < w_2 \le w_1$.

(b) When $w_2 > w_1$, the first order conditions are

$$\epsilon w_1 = \frac{1}{4} + \frac{w_1}{6w_2}, \quad \epsilon w_2 = \frac{1}{4} - \frac{w_1^2}{12w_2^2}.$$

But since $0 < w_1 < w_2$, these expressions imply that $\epsilon w_1 > 1/4 > \epsilon w_2$, a contradiction.

We conclude that there is no solution of (45) with $0 < w_1 < 1$, $w_2 > 0$ and $w_1 + w_2 < 1$.

4. Now suppose that $0 < w_1 \leq w_2 < 1$, $1 < w_1 + w_2$. It follows that $(1 - w_2)/w_1 < 1$, $(1 - w_1)/w_2 < 1$, and $w_2 > 1/2$. Note that

   - $r \in [0, 1] \implies -w_1 r/w_2 \geq -1$
   - $r < (1 - w_2)/w_1 \implies (1 - w_1 r)/w_2 > 1$
   - $r \geq (1 - w_2)/w_1 \implies (1 - w_1 r)/w_2 \leq 1$.

   For w_2, we have

$$\begin{aligned}
\epsilon w_2 &= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)} \left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) x_2\right] \\
&= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)} \left[\mathbf{1}\left(-w_1 r/w_2 < x_2 < (1 - w_1 r)/w_2\right) x_2\right] \\
&= \int_{r=0}^{(1-w_2)/w_1} \left[\frac{1}{2}\int_{x_2=-w_1 r/w_2}^{1} x_2 dx_2\right] dr \\
&\quad + \int_{r=(1-w_2)/w_1}^{1} \left[\frac{1}{2}\int_{x_2=-w_1 r/w_2}^{(1-w_1 r/w_2)} x_2 dx_2\right] dr \\
&= \frac{3w_2 - 2w_2^3 - 1}{12w_1 w_2^2} + \frac{1 - w_1^2 - 3w_2 + w_1 w_2 + 2w_2^2}{4w_1 w_2} \\
&= \frac{-3w_1^2 w_2 + 3w_1 w_2^2 + 4w_2^3 - 9w_2^2 + 6w_2 - 1}{12w_1 w_2^2} \\
&= \frac{-3w_1^2 w_2 + 3w_1 w_2^2 + (4w_2^3 - 9w_2^2 + 6w_2 - 1)}{12w_1 w_2^2} \\
&= -\frac{w_1}{4w_2} + \frac{1}{4} + \frac{(1 - w_2)^2(4w_2 - 1)}{12w_1 w_2^2}.
\end{aligned}$$

   Recall that $w_2 > 1/2$ in this subset, hence $(1 - w_2)^2(4w_2 - 1) > 0$. Therefore $-\frac{w_1}{4w_2} + \frac{1}{4} + \frac{(1-w_2)^2(4w_2-1)}{12w_1 w_2^2}$ is convex as a function of w_1 on the interval $1 - w_2 < w_1 \leq w_2$, so the maximum occurs at $w_1 = 1 - w_2$ or $w_1 = w_2$. In the former case of $w_1 = 1 - w_2$, we have

$$\epsilon w_2 = -\frac{w_1}{4w_2} + \frac{1}{4} + \frac{(1 - w_2)^2(4w_2 - 1)}{12w_1 w_2^2} = \frac{2w_2 + 2w_2^2 - 1}{12w_2^2} < 1/4 \quad \forall w_2 \in (1/2, 1).$$

   In the latter case of $w_1 = w_2$, we have

$$\epsilon w_2 = -\frac{w_1}{4w_2} + \frac{1}{4} + \frac{(1 - w_2)^2(4w_2 - 1)}{12w_1 w_2^2} = \frac{(1 - w_2)^2(4w_2 - 1)}{12w_2^3} \leq 1/6 \quad \forall w_2 \in (1/2, 1).$$

   Thus $\epsilon w_2 < 1/4$ for the chosen range of values of w_1 and w_2.

For w_1, we have

$$\begin{aligned}
\epsilon w_1 &= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)}\ \left[\mathbf{1}\left(0 < w_1 r + w_2 x_2 < 1\right) r\right] \\
&= \mathbb{E}_{r\sim U(0,1)}\ \left[\mathbb{P}_{x_2\sim U(-1,1)}\left[-w_1 r/w_2 < x_2 < (1-w_1 r)/w_2 \mid r\right]\cdot r\right] \\
&= \frac{1}{2}\mathbb{E}_{r\sim U(0,1)}\ \left[\left(\min\left\{1, (1-w_1 r)/w_2\right\} - \max\left\{-1, -w_1 r/w_2\right\}\right) r\right] \\
&= \frac{1}{2}\mathbb{E}_{r\sim U(0,1)}\ \left[\min\left\{1, (1-w_1 r)/w_2\right\} r + w_1 r^2/w_2\right] \\
&= \frac{1}{2}\left(\mathbb{E}_{r\sim U(0,1)}\ \left[\min\left\{1, (1-w_1 r)/w_2\right\} r\right] + \frac{w_1}{3w_2}\right) \\
&= \int_{r=0}^{(1-w_2)/w_1} \frac{r}{2} dr + \int_{r=(1-w_2)/w_1}^{1} \frac{1-w_1 r}{2w_2} r dr + \frac{w_1}{6w_2} \\
&= \frac{(1-w_2)^2}{4w_1^2} - \frac{2w_1^3 - 3w_1^2 + 2w_2^3 - 3w_2^2 + 1}{12w_1^2 w_2} + \frac{w_1}{6w_2} \\
&= \frac{1}{4w_2} - \frac{(1-w_2)^3}{12w_1^2 w_2} \\
&> \frac{1}{4w_2} - \frac{1-w_2}{12w_2} = \frac{1}{6w_2} + \frac{1}{12} > \frac{1}{4},
\end{aligned}$$

where the third, fourth, and fifth equalities follow from our assumptions on w_1, w_2 (outlined in the bullet points above), the first inequality follows from $(1-w_2)^2/w_1^2 < 1$ and the second inequality follows from $1/2 < w_2 < 1$.

We have thus shown that $\epsilon w_1 > 1/4 > \epsilon w_2$, which contradicts one of the inequalities that defined this subset, namely, $w_2 \geq w_1$.

5. Finally, we consider $0 < w_2 < w_1 < 1$, $1 < w_1 + w_2$. In evaluating the expressions in (45) here, we integrate first over $r \in [0,1]$ and then over $x_2 \in [-1,1]$. We adjust the limits for r when x_2 lies in certain subintervals of its range, as follows:

   - $x_2 \in [-1, 0] \implies r \in [-w_2 x_2/w_1, 1]$;
   - $x_2 \in [0, (1-w_1)/w_2] \implies r \in [0,1]$;
   - $x_2 \in [(1-w_1)/w_2, 1] \implies r \in [0, (1-w_2 x_2)/w_1]$.

   For w_1, and using (45) again, we obtain

$$\begin{aligned}
\epsilon w_1 &= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)}[\mathbf{1}(0 < w_1 r + w_2 x_2 < 1) r] \\
&= \int_{x_2=-1}^{0} \frac{1}{2}\left(\int_{r=-w_2 x_2/w_1}^{1} r dr\right) dx_2 + \int_{x_2=0}^{(1-w_1)/w_2} \frac{1}{2}\left(\int_{r=0}^{1} r dr\right) dx_2 \\
&\quad + \int_{x_2=(1-w_1)/w_2}^{1} \frac{1}{2}\left(\int_{r=0}^{(1-w_2 x_2)/w_1} r dr\right) dx_2 \\
&= \left(\frac{1}{4} - \frac{w_2^2}{12w_1^2}\right) - \frac{w_1 - 1}{4w_2} + \frac{w_1^3 + (w_2-1)^3}{12w_1^2 w_2} \\
&= \frac{w_1^2(3w_2 - 2w_1 + 3) + 3w_2(1-w_2) - 1}{12w_1^2 w_2}. \qquad (50)
\end{aligned}$$

For w_2, we obtain

$$\begin{aligned}
\epsilon w_2 &= \mathbb{E}_{r\sim U(0,1),\, x_2\sim U(-1,1)}[\mathbf{1}(0 < w_1 r + w_2 x_2 < 1)x_2] \\
&= \int_{x_2=-1}^{0} \frac{1}{2}\left(\int_{r=-w_2x_2/w_1}^{1} dr\right) x_2 dx_2 + \int_{x_2=0}^{(1-w_1)/w_2} \frac{1}{2}\left(\int_{r=0}^{1} dr\right) x_2 dx_2 \\
&\quad + \int_{x_2=(1-w_1)/w_2}^{1} \frac{1}{2}\left(\int_{r=0}^{(1-w_2x_2)/w_1} dr\right) x_2 dx_2 \\
&= \int_{x_2=-1}^{0} \frac{(w_1+w_2x_2)x_2}{2w_1} dx_2 + \int_{x_2=0}^{(1-w_1)/w_2} \frac{x_2}{2} dx_2 \qquad (51) \\
&\quad + \int_{x_2=(1-w_1)/w_2}^{1} \frac{(1-w_2x_2)x_2}{2w_1} dx_2 \\
&= \left(\frac{w_2}{6w_1} - \frac{1}{4}\right) + \frac{(w_1-1)^2}{4w_2^2} - \frac{2w_1^3 - 3w_1^2 + 2w_2^3 - 3w_2^2 + 1}{12w_1w_2^2} \\
&= \frac{3(1-w_1)w_2^2 - (1-w_1)^3}{12w_1w_2^2}. \qquad (52)
\end{aligned}$$

Multiplying (50) by $12w_1^2w_2$ and (52) by $12w_1^3$ makes the left-hand side of both equations $12\epsilon w_1^2 w_2$. When we multiply the right-hand sides by these same factors and equate them, we obtain

$$w_1^2(3w_2 - 2w_1 + 3) + 3w_2(1-w_2) - 1 = 3(1-w_1)w_1^2 - (1-w_1)^3\frac{w_1^2}{w_2^2},$$

which after rearrangement becomes

$$-w_1^2(1-w_1)^3/w_2^2 = -3w_2^2 + 3(w_1^2+1)w_2 + (w_1^3 - 1). \qquad (53)$$

Observe that the left hand side of (53) is strictly negative for all w_1, w_2 in the range we are considering here. Writing the right-hand side as $p(w_1, w_2)$, we aim to show that

$$0 < \min_{w_1\in(1/2,1)} \min_{w_2\in[1-w_1,w_1]} p(w_1,w_2). \qquad (54)$$

We first solve the inner minimization problem (over w_2). This is a bound-constrained strictly concave quadratic in w_2, so its minimum must occur at one of the endpoints. Hence, (54) is equivalent to

$$0 < \min_{w_1\in(1/2,1)} p(w_1, 1-w_1), \quad 0 < \min_{w_1\in(1/2,1)} p(w_1,w_1). \qquad (55)$$

For the first inequality in (55), we obtain

$$\begin{aligned}
p(w_1, 1-w_1) &= -3(1-w_1)^2 + 3(w_1^2+1)(1-w_1) + (w_1^3-1) \\
&= -2w_1^3 + 3w_1 - 1 \\
&= (w_1 - 1)(-2w_1^2 - 2w_1 + 1).
\end{aligned}$$

Since $-2w_1^2 - 2w_1 + 1$ has roots at $-1/2 \pm \sqrt{3}/2$, it is strictly negative for $w_1 \in (1/2, 1)$. Meanwhile, $w_1 - 1 < 0$ for $w_1 \in (1/2,1)$, so we have $p(w_1, 1-w_1) > 0$ for $w_1 \in (1/2,1)$, as required. For the second inequality in (55), we have

$$p(w_1,w_1) = -3w_1^2 + 3(w_1^2+1)w_1 + (w_1^3-1) = 4w_1^3 - 3w_1^2 + 3w_1 - 1.$$

The gradient of this expression is $12w_1^2 - 6w_1 + 3$, whose minimum occurs at $w_1 = 1/4$ with value $9/4$, hence the gradient is strictly positive for all w_1. Thus we have $p(w_1, w_1) > p(1/2, 1/2) = 1/4$ for all $w_1 \in (1/2, 1)$.

We conclude that (53) cannot be satisfied for any w_1, w_2 in the subset being considered in this case.

We have shown that all 5 cases lead to a contradiction, so we conclude that no points (w_1, w_2) with $w_2 > 0$ satisfying (45) exist. This completes the proof. □